Engineering and science in our world

Sandip Tiwari

Sandip Tiwari
May 2023

SANDIP TIWARI

ENGINEERING & SCIENCE IN OUR WORLD:

THIS I BELIEVE.

Sandip Tiwari is the Charles N. Mellowes Professor in Engineering Emeritus at Cornell University and a Distinguished Visiting Professor at Indian Institute of Technology at Kanpur. Having grown in a household graced by friends who were teachers, literary figures, Gandhian social workers, and freedom fighters, and having an innate love for the activities of the mind and the hands, his personal tastes and passion gravitated to science and engineering with a skeptical streak. He has spent about half of his working life conducting research in industry (*IBM*) and the other in academe, where he has spent periods at many of the world's premiere institutions in *USA*, Europe and India, led national academia-based major efforts, and served on the advisory boards of national efforts and of industry. His research and technological contributions have been recognized with numerous awards. In recent times, particularly satisfying to him is the Electroscience Book Series from Oxford University Press and sparking in young minds the balancing of rigor and intuition in figuring things out.

Semiconductors-, computation-, and more broadly information-centered pursuit is now pervasive in our society. It powers and drives the society through the commerce and the social schema in the ˵medium is the message.˝ Labor almost appears as a secondary afterthought and a Marxist anachronism. This enterprise as an engine has hybridized with economics, defense, class, culture, power and religion.

These essays, technical and societal, are views that crystallized during the past decade as I took a step back and pursued ideas and interests as an individual rather than for the affinity group as we all are wont to do carried on in the flow of life. They are written with those technically inclined in mind. The students of Indian Institute of Technology (*IIT*) were the audience when I articulated these ideas in the Kanpur lectures of 2023. While at times they may become sufficiently mathematical that a general reader may not absorb the analytics of the argument it is my hope that the flow of ideas being conveyed will still be digestible.

Starting with a discussion of education, particularly higher education, the sequence explores some of the major questions and challenges of the current state—technical and worldly—with my guesses for the future. The initial three are a technical perspective viewed through my experiences. They discuss, by viewing complexity in electronics as a mix of determinism and uncertainty, the questions arising in the giant integration scales now possible, the indeterministic view of the world of open boundaries, and of computation for the complex and incomplete problems that can now be tackled through machine learning and neural networks. The latter is an entirely new way of tackling complex problems that I did not have in my formative years.

The last two essays follow through from a very personal frame of reference, my mind's eye and I, cultural and humanist lessons of the circumscribed science and engineering pursuits in a real world with its open boundaries. It is a quest of learning and of change in a dynamic world. I end with thoughts probing the future in which semiconductors—an early research love—as an essential physical layer for nearly every economic pursuit and a critical national pursuit for my birth country.

To the young faculty and creators,

who took a chance at building a new institution

that fostered the intellectual blossoming of young Indians

of a newly independent nation,

and for instilling in them that it is an individual's responsibility

to figure out and act on what kind of a person they wished to be.

PRINTED WORDS HAVE NO LIFE,

BUT YOU CAN GIVE LIFE TO THE PRINTED WORD IF YOU ARE SERIOUS.

Jiddu Krishnamurti

Acknowledgments

Mari and I spent the spring of 2023 at IIT Kanpur. For me it was a pilgrimage, my *calf-grund*, and a chance to relive a formative period, perhaps a *hiraeth*, but this time watching a vibrant new world of young self-confident humans, bicycles everywhere on a more diverse campus, and the nature's cycle.

The past of all cultures has distilled essentials of human condition to pithy phrases. *Calf-grund* is an old Scots expression describing a territory on which one was calved and whose imprint remains, no matter where one goes later. *Hiraeth* is a Welsh word for nostalgia and homesickness for a place that no longer exists. My maternal grandmother would say to me in Bundelkhandi ˵*Gam khao,*˶ literally ˵*Eat sorrow,*˶ and meaning calm down and take it slowly. I am still working on this instruction from the summers of my childhood.

The waves of new winter flowers, the bougainvilleas cascading in shades of pinks, reds, oranges, yellows and whites, the flowering trees, specially the Gulmohars and Alamtas, have been a feast. So also the cacophony of numerous species of birds that love this oasis in the Gangetic planes. I made new friends: Uday jee and Pawan jee, guards at the Director's house, who also took care of family rabbits, earning away from their family in their villages, Khemji, from Nepal terai and now settled with family in Nankari, Kushwaha bhai, who explained to me how to get Cinerarias to germinate, grow and bloom, and Sanjaya jee, who may be a cleaner despite having a bachelor's degree in economics but committed to getting the best education for his children. The teaching let me put together the first draft and practice the remaining volume of the Oxford Electroscience series placing information at the center of statistical, quantum and computational mechanics. Also lighting a few fires in the young minds and warding away higher-order illiteracies. The students across the spectrum of disciplines and years-in-school made the writing of the Python journal files and organizing and sequencing the information-centric arguments a pleasure. Also by probing through the questions in the evening talks.

The writing here come from the five talks of the Kanpur lectures series. This series let me place recent research work within the context of life learning and share broader thoughts in the last two by expanding to the social context, and the societal-technical entanglement of development through science and technology.

The talks can be accessed at www.iitk.ac.in/scdt/Sandip_Tiwari_Kanpur_Lectures.html web location.

The essay on interdisciplinarity and cultures of science and humanities is in honor of Prof. K. R. Sarma, who was present, and who was my advisor in 1975–76 in the life-instructive design project.

After a personal view on education as a preamble, the initial essays are technical. The first sets the vast integration of semiconductors in thermodynamic context. The next two expand to the intersection of complexity and machine learning. We now have a very powerful new tool that is opening a vast new science and engineering territory to exploration by the young and keen mind.

One's writing is reflective of emotional, moral and analytic gravity. All that one imagines, thinks, infers, and writes must change with time as one learns. This is what makes life worth living where immutable answers need not exist. We are only participants in state changes under cause and chance. Since I am venturing personal opinions at times, some discussions here will provoke nodding and some will cause indignation. Understanding comes from hearing and amicably resolving these different points of view.

In 1971 I had arrived to pursue physics in its inaugural integrated program but also had an engineering vein. *IIT* and the rest of life has allowed me to inhabit both these worlds. Lest one thinks there are hard boundaries it was corrected in the very first year when we had to write an essay in our humanities class on some philosophical thinking theme of our choice. I have always been interested in Buddha's teachings, in non violence and satyagraha as working principles of life, and in suffering and its resolution. I quickly hit my articulation limits as I put pen to paper. So I wrote to my father. When the submitted essay was returned, Ms. Chitra Ray, our tutor, had circled a part in red with a note, paraphrased here after fifty+ years, ``This is from the heart. Where is the rest from?'' Bless her. The circled part was my father. This cathartic event opened a door for me. Thinking is not definitive and complete—in as much as it can be—until one has argued it out on paper for archiving.

Who would have thought that atheists can have a pilgrimage just as satisfying as the Amarnath yatra or Sabarimala trek or Mecca or the Camino de Santiago? Being at *IIT* was transformational, then and now. The 1971–76 period was of nonlinear learning and of lessons such as of the incompleteness of ontological and epistemic boxes. The 2023 period helped me pull some of this unfinished business into an information-centered perspective. It was rejuvenating.

This was all made possible by the Institute hosting us, the love and the effort that Prof. Sundar Iyer, Mr. Dharmendra Swain and Prof. R. Vijaya put into all that makes life work out. The memory of 2023 will always bring a smile of sharing this afterimage, of the new friends made, of the changes one can see in a dynamic India, and of being still-a-student after having survived so far all the complexity that existence throws at one in life.

Sandip Tiwari, *Kanpur, May, 2023*

Contents

1
Preamble: On education

We are born accidentally. We do not choose the family, the environment, whether rich or poor or powerful or not. We just materialize, are sheltered in early years, are sent off to schools, meet and are influenced by individuals and groups, and we are expected to channel a course through the world. Nothing is quite free will here, a lot of it is a product of the environment, but we can and do clear a path for ourselves building up from the circumstances with personal propensity. Education builds the path and provides the means to personal fulfillment. It begins at home, is extended and expanded through schooling and settles itself as an evolving track in the vicissitudes of the real world. The mechanisms of education have changed over history. This is particularly true since the dawn of modern science. The traditional liberal arts have been supplanted, sometimes gradually and sometimes rapidly, by modern sciences and those of the engineering, medicine, management and other professions that support living. It is therefore not surprising that much tension exists in what education is about, what education should be, and how the society should pursue this aspiration for every human and to what extent. This preamble is of thoughts from a sentimental and sometimes accidental wanderer in this spectrum. Some of these are bequeathed by history and some are from observations of a life spent in science and engineering in an east-west emerging-developed world.

In times past—hundreds of years and beyond—education happened at home and by living in the world that surrounded us. Institutionalized education of the masses—early schools through colleges or universities or institutes of various nomenclatures and pedigrees—is largely a modern phenomena, and a change that has been an incredible enhancement for human development and in the life of a human being.

This transformation also causes a clash of the old and the new that reverberates constantly, within us, within families, within institutions, and within all of the different configurations in the society.

Science and engineering in the world

1.1 Education: Childhood, studenthood and life

What is education? What are its goals? What is the best way of attempting these objectives via schooling and what should the schooling emphasize to achieve the goals at the various stages given what we understand of child and adult development? How far must one go in schooling? How does one separate indoctrination from free thought? How is one free given all the pressures of needs and powers? And so on and on. These and others are diverse questions rooted in the past and of the needs of the future. On one side is a conveyed tradition and on the other is the great unknown that the education to the young ones is for.

No wonder that there is much conflict and debate and social and systemic wars no matter where one resides in the world and the terroir of that environment. The struggle is integral to being human.

Historically, education was *artes liberales.* In the ashrams, dominantly for the children of the aristocracy, it was for developing an understanding of traditions, developing capabilities of reasoning for making small and large decisions, and also to build capacity for war. In the Greek tradition it was the learning for a free person. The original Greek form of education concentrating on theology, law, medicine and philosophy that folded in natural sciences is not that different from the Indian tradition. It too was for the ruling class—the ``free˝—and not for population at large—slaves or slave-like folks—an indentured laborer in the British euphamism, for example—regardless of what we call them in different parts of the ancient world.

In the middle of this millennium, the classical education began being complemented by *scientiae lucrativae,* an education for material gain.

In the West, in the institutions of learning, the earliest of the changes in the 16th century were in the faculty of arts turning into the faculty of philosophy. The new sciences—chemistry (originally alchemy) and botany particularly as tied to medicine—brought in the empirical and rational way of exploration that did not depend entirely on bias-constrained mental arguments and discussions of old times. This is the start of the early conflicts between old and new sciences, the old having traditionally been folded into philosophy.

Liberal arts had its emphasis mainly on classics—*Ramayana* or *Mahabharata* or *Republic* or *The Iliad*—and classic traditions, where thoughts and discussions of human proclivities were employed for education.

Is this liberal tradition merely ornamental and are lucrative sciences mainly utilitarian? Is this just a duel between mental delights and material goods?

I do not think so.

Over time, it is the amalgamation of the liberal approach and the scientific approach that became the enlightenment: an interaction of ideas and social reality or superficially of reason and rationality.

With the introduction of scientific method of observation—empiricism—drawing on the work of Francis Bacon and John Locke, and others, the amalgamated enlightenment happened in the mid-1600s with Rene Descartes' musings on logic and method—*Cogito, ergo sum. I think, therefore I am.*—and in India, with Raja Rammohun Roy's Brahmo Samaj in early 1800s—*Truth and virtue do not necessarily belong to wealth and power and distinctions of big mansions.*.

As we learn more, we are educated more, and we evolve how we see the past and project to the future. This is enlightenment. As an example, Descartes' reasoning of rational methods now need a thorough questioning in light of what we now know about the human brain.

This evolution—hundreds of year in making— is an overthrow of the long period of purely mental exercises such as of Aristotle or of Plato and Ptolemy or of the Indian rishis, even if there is much in classics of Greece and of India or of China to be revered and admired.

This supplanting—perhaps today an overpowering and not just complementing—forced education, and places for education to change.

Without sciences and engineering and physiology, we cannot build a new world, but without literature—ancient and modern—and music, dance, and arts, is the new world worthwhile?

Liberal education with its goal of human completeness, the teacher backing the nature of the students, nurturing their hunger and their capacity, and reinforcing keeps the building up alive. This is how generations progress and tackle the permanent concerns of mankind. Specially so since tastes change, what angers one changes, and what one does is affected by ‶the medium is the message.″ Liberal education gives one the tools to deal with the problems of the world one has to inevitably deal with once we leave parental security. Science and engineering and others give us the vocational and even free-spirit means for making a living and exploring our living. They need to coexist for the society to function.

This brings us to the question of enlightenment. What is enlighenment? It is the shining of light on darkness, the replacing of opinions—superstitions—by scientific knowledge of nature. Start from a phenomenon that can be seen by all and end with some rational demonstrative conclusion that can also be seen by all.

I think, therefore I am is an incredibly Gödelian self-conflicted statement. Is it provable or not provable? The brain, our understanding of neurosciences, the processes of complexity have much to do with the logic and method argument. Antonio Damasio's *Descartes' error*, Penguin, ISBN 0-399-13894-3 (1994) is an incredible discussion—now dated because of the progress in neurosciences—of emotion and reason. Does a person who loses cognition cease to exist? A more recent writing is R. Sapolsky, *Behave. The biology of humans at our best and worst*, Penguin, ISBN 978-1-59420-507-1 (2017) discussing the reward system of frontal cortex determining how we respond to situations. Does free will exist? It cannot. It is not that long before we will be manipulating the brain through neural approaches. Planting memories and controlling behavior is a time-honored warfare and popular-fiction book technique. The myth building by which societies and often institutions operate does pretty well with this controlling.

Aristotle's failure is best embodied in the claim that heaver objects fall faster as being a pure mental exercise. It took the famous Tower of Pisa experiment to dislodge that heresy.

It is the breakdown of symmetries that underlies our existence and heterogeneity. Symmetry is a Platonic ideal, and geocentrism—earth as the center of universe—is ascribed to Ptolemy. Plato is to be befriended. Plato's *Republic* is education, then and now. Plato, when disagreeing with Aristotle on the nature of good, was still underscoring the society of the mind. The world is neither ideal nor is there anything unique about the planet or us. Ramsey's 1920 combinatorics theorem tells us that there exists a quantifiable minimum-sized collection for one to find any specific set of relationships. The self-centered notion of our uniqueness and not knowing of intelligent life beyond the earth is a problem of spacetime. It most likely exists, somewhere, since the universe is immense, and it did in the past too, but can it reach us under constraints of Einstein's general relativity and of thermodynamics of matter?

Enlightenment is a free pursuit unconstrained by punishments from the society to explore reason, cause and effect. This is intellectually honing. With the German philosophical traditions standing tall, the enlightenment broadened empirical and rational inquiry from that of Locke, Bacon, Descartes and others of that 17th century period, to all disciplines, old and new.

Sciences, with their worldly emphasis, kept expanding. The invention of engines converting energy in various forms through various carriers—human, steam, electric, combustion, nuclear, solar—of electricity as a convenient means to transportation of energy so that the society did not need to stay confined at energy sources and water bodies alone, the understanding of matter, of sickness and of nature have transformed the society and how we interact and what we pursue. This happened through the developments of science and its use that blossomed into the large area of engineering. Science remains more abstract and engineering more applied, and together they have transformed the society. This is what makes *science and engineering a societal force* on par with all the others that make being human worthwhile.

We arrived at this seeming bifurcation over the past five hundred years as Indian, Chinese, Greek, Latin or Middle Eastern heirlooms gave way to new knowledge that furthered the material and nonmaterial living.

This then brings up the debates and discussions of the age old question of what it means to be enlightened, rational, or human. We need a contemporary interpretation and on how one should pursue that course, and in turn with the use of the science and engineering, progress the society along a humanist path. At its foundation, this is the pursuit of truth, ephemeral as it may be, work towards human completeness, keep the natural world alive and progressing through local and global reach. On one side are the big questions, of reason versus revelation, of freedom versus necessity, of democracy versus plutocracy or aristocracy, of good versus evil, of body versus soul, of soul versus other, of together versus individual, of eternity versus present, or of being versus nothing and on the other side is the real world position and flow in adaptation and survival.

This is an idealistic view. This is not how society practices it. The second dominates since survival, wealth creation, and upliftment of poor is how the societal systems are organized, whether democracy, plutocracy, authoritarian, monarchy or communist.

In the ideal view the role of precollege schooling is the development of the child—standing up on all ten, learning the past and developing an understanding and consideration of others and the nature—and preparing for an adulthood and membership of the society through an all round development—of integrity, accountability, and leading a life in the light. The precollege education across the different directions of knowledge aim towards this so that the child can make a learned choice of how they wish to pursue their place in the world. This happens aligned with the maturation of the mind while the body is undergoing hormone and chemical changes that are the major stimuli of the teenage period. At the end of the high schooling, the young person could be going directly to work and learning on the job, or get a vocational education, or go on to higher education at the academe.

We sometimes refer to one part of this development—related to the person in the society—as values. Values are however only the products of people's minds and have relevance only to those minds. They change with time. Nietzsche, in late 19th century, for example, says that humans are losing capacity to value and therefore humanity. We have to be sympathetic to this view as the world wars happened within a few decades. He viewed self-satisfaction—the feeling of being adjusted and a comfort in having solved one's problems—as a sign of incapacity to look up to perfection and the overcoming of oneself. Nietzsche also saw problems with how the word is used. Authentic values create culture. Religions' teachings aim towards this. But, if they are not rational and in the nature of that community, then it is being imposed. So any opposite values must be subjugated. Rational persuasion cannot make anybody a believer since values and believing in them are acts of the will. It therefore turns into a problem of lack of will. A value as a value is life preserving and life enhancing. These are all conditionals, and if one depended entirely on schools—teachers and cohort—and family for development, it cannot really succeed too well.

This leads to culture in the discussion of development. Culture as art is a supreme expression of human's creativity and of the capacity to break out of nature's narrow bond. It builds dignity. The culture of a community is the fabric of relations in which the self exhibits a diverse and elaborate expression. It comes from self and is also the product. So it is a production and a product. In this sense, like value, culture too is relativist. This is why wars, great cruelty instead of compassion, is to Nietzsche a fundamental phenomena. We indulge in it all the time, world wars just continue after an intermission as right now (2023) in Europe and Middle East, and these cycles will keep repeating. War is the fundamental phenomenon upon which peace is sometimes forced. This is Nietzsche's criticism that this is civilized reanimalization.

A child needs going beyond the bounds placed in imagination, creativity and aspirations. Creativity implies separating oneself from others in some unique way. This is contrary to rationalism or egalitarianism. Childhood is the perfect time for such dreaming to be promoted for building uniqueness, which we later on attempt to corral. The soul's longing in the midst of constraints and conditionals needs encouragement from the very beginning. Books are central to this transformation. Books—children's books such as *Black Beauty* by Anna Sewell, or the *Peter Rabbit* stories of Beatrix Potter, or the *Panchtantra* tales, or *Idgah* or *Bade Bhai Saheb* of Premchand, or *Afeemchee* of Hazariprasad Dwivedi, so many others, are an outlet of encouragement to the young child's mind to let go off their local confines and let it wander. Such a child then knows that the reciprocity of rights is fundamental. There is no need of adherence to any religion or creed or culture or ethnicity.

These early years are enormously important in human development. Psychology studies tell us that the character comes from the early preschool years. Character along with knowledge make for education. Character is associated with vitality, courage, sensitivity, intelligence, curiosity, and other similar traits. Vitality is a physiological characteristic. If you watch people across the age spectrum you will notice that it is the one characteristic that dwindles with age. Courage is sometimes the absence of rational or irrational fear and sometimes it is in the control of these fears. Sensitivity is affection and sympathy towards others. Sensitiveness is a modulation on courage promoting affection and sympathy towards others. Intelligence arises in alert curiosity. It is the genuine love for knowledge.

There are plenty of studies of large cohorts that also tell us that childhood trauma, desperation and violence have a long arm stretching throughout life. Only a few are lucky to overcome the neural consequences. Wars, poverty, ostracism in all forms continue to then flow. This is a reason to best see a society through the lens of how it treats the child and the equality in education that it provides to every child through the age of 18. This is equal opportunity. The current Germanic societies do this exceptionally well.

We are all born with the potential for all these traits and it is the environment of the early formative years that establish them as lifelong habits. In the Chinese way of thinking—a Confucian–Taoist view—logic and morals and wisdom of life are all one. This is another way of looking at character. If the later life is in a supportive environment, the desirable traits will continue to solidify and grow. The traits also help knowledge acquisition part of education.

These early years of school education are also for taming raw passions. But it is not to suppress or eliminate them. If one did that, one is excising the energy that makes us us. It is a molding process. This process is largely informed by the liberal arts. Later in life—at the university and beyond—and this is the most difficult part, one needs a harmonization of the enthusiastic parts of the soul with that rational part that builds up later. As one ages, one feels more and more incomplete the larger the disconnection is in between the inner self—the internal aspirations not seen by the world—and the outer self—the one molded and pushed by the society—that one presents. A person can never be whole without this harmonization. Music and poetry as another reasoned form of music, are a delicate

In the later essays, I point to the correspondence between this view and the Kullback-Leibler divergence view in information theory.

balance of passion and reason, This is why love of music and arts is so important to harmony in life. Even when music and arts turns more edgy and tight, as also religion, which is both warlike and erotic, the scales are generally tipped towards passion.

The ambitions get molded by the models experienced in the books we read and we develop an inner feel for the yearnings of the heart. I recall my early school years with my father bringing home the colorful magazines published by the American Embassy, one of which had Martin Luther King and his *I have a dream* speech as the focus that is still stuck in my head. I also received the free Soviet magazines (*Sputnik*) with collective farms of golden fields spread out to infinity and giant harvesters doing the hard work. All glossy. The young mind is so impressionable. Both of these were propaganda. Both fallacious in that the dream is still unfulfilled and the farms didn't quite work out either. Yet, these are good examples of how real life is also part of the education process. I could dream. I was still a mere preschooler watching the Sputnik making its rounds in the night sky and me riding it around in the universe. Later on, with Mir Publishers and English Language Book Society supplying inexpensive books, *Science Today* and *Scientific American* as the magazines, and popular science books such as George Gamow's *One, two, three, infinity,* and so many others.

A university or the narrower college or institute education is the place for what we have traditionally called higher education, a place to go beyond the broader learning of the basics to enter the society. They exist to mold students to become discoverers and doers who go beyond being mechanical parts of a societal machinery.

Higher education exists to provide us with the means—tools, methods, ability to put such resources together, and build a repertoire to handle our way through the world. How can higher education do that well and how does a student make the most of it? This world of higher education institutions is both the world of liberal arts and of the more utilitarian professions.

In the liberal view, we cannot be satisfied by our culture if we are to be a full human. Plato, in the *Republic,* draws this picture of a cave with us as prisoners in it. *The culture is the cave.* We should use nature as the standard for judging our lives and of lives of other people. This is what places philosophy in its important position in the liberal tradition.

Science and engineering are a body of systematized and verifiable knowledge. They express and utilize relationships between definable phenomena. This is in contrast to matters of common knowledge or of opinion or of belief. In a loose way, science and engineering are attempting to resolve *if this then that* question. It is an attempt at completeness in as much is possible given what one knows and by in some way accounting for what is not known.

Humanities in general and philosophies in particular are exploring an open system of *what if and why so* questions. The former—science—is attempting to close a matter of import and the latter—philosophy—is attempting to open a matter of import by making sure all assumptions and possibilities are properly questioned.

Understanding and forecasting the world around us—an educational goal—is experiential. We observe, we guess, we make models, we forecast, but in the most complex of the problems, it is preposterous to attempt it since there is so much not known, and so much not knowable, that we might as well drink tea and chew paan. This is trained ignorance. The university education is enables us to be able to differentiate and attempt both the complete and the incomplete problems that can be tackled.

Is a university a collection of a series of disciplines, each of which has its own rigidly ordered form of study, that is, as separate schools, or is an education one of a compendium of various disciplines and of the connections between them? Is it a place for critical thought and of an understanding of principles so that one can morally and ethically pursue some career? Or is it a place for learning the wherewithals of the narrow confines of one direction of application pursuit? It is here that the dilemma of modern education is. Old liberal ideals, new scientific and technical needs, or something else?

I am a card-carrying phenomenologist. I see phenomenologism as a scientific process, where one must not get addicted to abstractions and generalizations, as philosophy so very well teaches us. I see, I interpret it through my experiences and through my viewing of the viewpoints of others, where I am more in agreement with Husserl, Heidegger, and the later existentialists, rather than Ockham or Kant, or Vannevar Bush or Pasteur. I celebrate our existence. We are bombarded with all that is wrong, of despair-des-jour, but in the midst of all this we are friendly, attempting equality, living naturally, and not entirely depending on history and culture. This is what makes the best for ourselves and what sometimes succeed with those around us and by extension to a bigger world around us.

Many of the world's brightest young arrive at the Indian Institutes of Technology in pursuit of education and thereon to finding their place in the world. Because these are elite institutions, despite high costs now, the student is materially free to do what they want. Some of them are also spiritually free. But that depends on the experiences of earlier years.

A university education—any education—is by itself incomplete. Getting educated in schools of learning is only a process in the hu-

One can find many liberal education-oriented books, from Bertrand Russell's *Education and the good life,* Liveright, (1926), Irwin Edman's *Philosopher's holiday*, Viking (1938), Hazard Adams, *The academic tribes,* Liveright, ISBN 0-87140-623-3 (1976) to Allan Bloom, *The closing of the American mind,* Simon & Schuster, ISBN 978-5-551-86868-2 (1987) that argue based on philosophical ideals of making a human as a citizen of the society drawing on modern psychological understanding of the human. Society is also survival. Philosophers tend to be idealists. Nothing at all wrong with it. The challenge of living and of education however is finding the middle path between idealism and making our way through the world. Unfortunately, engineers, particularly those who gravitate to management, less so the scientists, become too enamored with the powering of the way through the world—plowing—and lose the humanness part. I subscribe to the middle path. For science and engineering, there is nothing quite equivalent as a discussion of ideas. There are older books, K. Popper, *The logic of scientific discovery*, ISBN 0–415–27843–0, Routledge (1934), W. A. Beveridge, *The art of scientific investigation*, Library of Congress 57-14582, W. W. Norton (1957), C. P. Snow, *The two cultures*, Cambridge, ISBN 0 521 06520 (1959) and T. S. Kuhn, *The structure of scientific revolutions*, ISBN: 0-226-45807-5, U. of Chicago, (1962), that get much attention, These are books on the process as one sees it. Not books that probe the entirety of spectrum of questions of why, how, why not, et cetera. For that, there are two books that I admire. One is by the great W. Heisenberg, *Across the frontiers,* Harper and Row, ISBN 0-06-011824-5 (1974), who discusses the meaning and ways for a modern university, and V. Narayanamurti and J. T. Tsao, *The genesis of technoscientific revolutions,* Harvard, ISBN 9780674518 54 (2021) which dwells on research and its nurturing, a direction to which the university is an integral part. An integrative view of science is J. Bronowski, *The common sense of science,* Vintage, (1956) whose discussion of truth and value in science is very remarkable. The Chapter 5 has a discussion of some of the arguments of the books while discussing culture's give and take with science and engineering.

man journey of finding one's place of comfort and pursuit in the world, where the world, like us, is continuously changing.

Higher education can help and higher education can hinder. For example, there is nothing in higher education that necessarily helps a troubled person. It cannot help with how one may want to conduct life such as in family matters, with the opposite sex, or with corruption. It is when we handle these within us from deep down in our consciousness that we learn what life's struggles are all about.

A student who arrives at the university comes already with a set of beliefs that have been either drummed in or have been acquired. Perhaps the student also has a love for art and music, and if that is so, then the early development has been quite successful. Classical music, specially instrumental music, but also vocal such as of Kumar Gandharva or Paluskar or the great operas of Verdi, Rossini or Bizet. It provides peace and never has the hard bite or discords that words can have even if they dwell on difficult human matters. Everyone is capable of appreciating these and some are good at creating these, They don't hurt. They speak of aspirations and something deep.

Today, music is everywhere, much more accessible than in the past. This music of the modern world is also not limited. It knows no class nor nation, so it does open the world to us. Classical music does this much more, it appeals to refinement, and in this it is doing the same as the classical books do.

Every culture has writers who shape and guide limits while we are young. For Indians, it may be *Mahabharata* as a conflict of living, or others, or something else including modern writing—mine was Naipaul, but others may have chosen Nehru, or Gandhi, or Tagore, or Premchand or Sharatchandra, or others, for French this may be Descartes and Pascal, for Germans Goethe and Mann, for Italians Dante and Machiavelli. They tell their people what their choice is, and they give a very perceptive view to life's perennial problems that weave the fabric of the soul. These are giving us a choice between reason and revelation, science and piety, choices from which rest will follow.

Perhaps the Draupadi cheer haran incident is what is referenced the most as one of the life matters from *Mahabharata*. But it was Eklavya's gurudakshina that is seared in my brain. What a lesson to give to a young one growing up. Was it the birth pedigree that was left intentionally obfuscated or was it a suppression of competition to the royalty? It is a conflict that only cleared once one grew up.

In the Naipaul or Gandhi, or Descartes and Pascal, one is making a choice between scientific rationalism and transcendent faith. Indians are mostly Gandhian. Faith rules. The French are divided equally between logical rationality and faith. They reflect a personal view. Such books are the independent gateway to education.

If a student arrives today with Elon Musk or Mark Zuckerberg as a hero model, they are following Ayn Rand's John Galt. Neither Cartesian nor Pascalian, but somebody who hasn't really read or imbibed books and learning. Such a lack of early education would mean that the student tries to get enlightenment wherever it is easily available, is incapable of distinguishing trash from sublime, propaganda from insight, or *AI*-generated from what is truly human. Tocqueville, in discussing his trip to America before the dawn of the American 20th century, where he saw it as Cartesian, also remarked that democracy's greatest danger is enslavement to public opinion. If public is not free and enlightened, then all else collapses. This is the great danger that is being highlighted in today's world, both in the West and in the East.

An artist's unconscious is full of monsters and dreams. Not that much of a scientist or an engineer. This is a reflection of liberal and utilitarian conflict. The future cannot be really predicted definitively. It is full of monsters and dreams.

This John Galt syndrome is higher education's dilemma now that sciences and engineering have become so powerful and it stresses why the enlightenment of reason and rationality is so crucial for a society. While growing up we all have a longing for overcoming of necessities, tensions, conflict, a resting of the soul, and the constant travail. A real education must respond to the need that one sees. A teacher sees the needs in the eyes of the students and in what is happening in the world around them. The intellectual structure of the educational institution—modes of working in the various branches and the organization of the studies—have to reflect this need.

1.2 *University as an institution for higher education and learning*

THE UNIVERSITY WAS TO BE AN ISLAND of intellectual freedom without restrictions in the idealized view. But in the process of being allowed to exist, it is an active and positive participatory venture in the society, and has to absorb the back flow of society's problems.

In Germany, when Alexander von Humboldt introduced the education reforms in early 19th century, he was advocating the idealized way of learning. Now it is rarely followed, even in Germany or Germanic countries. The Johns Hopkins model of late 19th century in the United States was to have a research pursuit—medicine being the original objective—guide the university. This approach still had the element of exploration, which centers on philosophy, in it. Education evolved under societal pressures.

Let me therefore return to the technical vocational enterprise that is Indian Institute of Technology. It was *IIT* Delhi that I first saw for an extended period during my high schooling. It was a few miles away, I could slip into its library while still a high school student—

school was early so the afternoons were free—and it was a great discovery along with that of the library at National Council of Education, Research and Training. *IIT* with a collection of buildings devoted to a higher purpose, not necessity or utility, or shelter or manufacturing, or trade, but with something ephemeral as an end in itself. The buildings had some tie to each other even though this was a campus in a big city. I admired that something like this could be put together. This was a personal enlightenment.

Coming to *IIT* then was part of growing up, it spoke of questions that ought to be addressed, what was important was not judged by money, friendships, shared experiences, models of discussion, et cetera, something that must have followed with some inkling of enlightenment.

But surprise, not just *IIT*s, but universities in general, are rarely Humboldtian or Hopkinsian. Departments tend to have their own culture and most universities are assemblages of departments organized into schools. Trained specialists—computer engineers, communication engineers, electronic circuits engineers, chemical engineers, doctors, lawyers, teachers, et cetera—are urgently demanded by the society, and specialized education from the beginning to the limit is the easy answer. This is addled by the society also looking for management people for which engineering is the easiest path demanding as it is in analytic skills.

We increasingly accept the notion that scientific thinking, ways of acquiring new knowledge, insights, learning sources of errors, building a logical sequence of arguments, et cetera, that is, learning, can be found by doing a web-based search or by listening to a webcast.

This is reflected in the current *IIT*s.

The liberal-utilitarian conflict now has a new persona. One can see it in various ways that education is pursued in different countries and even within countries. *IIT*s are quite discipline centered. The medical schools in India—any place where one goes to get a medical degree right after high school—are discipline centered. Tagore's Santiniketan was of the second type, and that method remains so for students of Philosophy at most advanced institutions.

But now in modern times we also have institutions of the third type, even more narrowing. Information technology, and another turn of specialization, data sciences, are of this kind.

We are increasingly forgetting what different personalities are like or talk to the personalities out of which some of the gems that one learns come out or the occasional hearing of something that can never be captured by the web or the books. This has come about because of the enormous expansion of science and engineering, the wealth that comes with it, and is now such that the age-old tradi-

I came to *IIT* Kanpur in 1971, was born in 1955, the population of India had not yet exploded, and for those lucky—accidentally—as in my case with the environment, it was not much of a competition to get in then. From our government school—Kendriya Vidyalaya or Central School as they were called then—of our class of 28 students from modest backgrounds, 9 tried the Joint Entrance Examination. All got in, and some of us did really well. Only one of these students had used coaching school—they existed even then—and he is now in finance. What we all had was wonderful teachers, specially in mathematics and the sciences, curious and absorptive us, and specially one student who loved abstractness in a way I have never seen again in life. It rubbed off. So there were plenty of spark points towards education and to promote interest from family, family friends, schooling, et cetera, and that is what channels and aligns best with the objectives of education.

tional disciplines are held in lower esteem. In United States, the medical schools are an entirely different environment with a further multiplicative factor in style that comes with this trend. In some institutions, it is the management schools. This society- and wealth-based change also leads to their being little safeguard against prejudice and idealogical delusion.

The unfortunate implication of this evolution is that even though sciences brought liberation from the dogmatic ways starting in the middle of this millennium, with the overspecialization and the wealth focus, the science and engineering academic do not necessarily have a central societal role. Breadth of thought should be a central condition even in sciences and engineering, and somewhere in this path of specializing, it got lost in learning and then it lost the vital connection to being a human.

There are limits to how long in time length the higher education schooling should be or could be. Humans have a finite lifetime and the knowledge keeps expanding. Inevitably, an engineering education or a science education with all its demands cannot be the same or much of liberal education. They serve two different motives. Yet, one is not complete without an understanding of the other. A liberal student must understand implications of thermodynamics or the economic relationships of people and effort or of *AI* just as a science student must understand the complexity of all the agents at work for a society to function and progress.

A teacher's task is to pay attention to the students, know their hunger and what the can digest, and find ways to get the knowledge across. A real education responds to felt needs.

1.3 *A smörgåsbord of dilemmas*

What does education, specially higher education, have to do with clearing of this path of societal progress? Specially higher education if one were lucky or could choose to pursue it.

The *IIT* students are lucky to have the enormous opportunity for the learning that the becomes accessible to them. But it is not the same as the wherewithals for living in the world, that is, of life. Life is quite distinct from that while in academe. Education is not the same as learning. Nor is the academe the right place for education unless one is self propelled or falls in the spell of a unique professor, a signed-and-delivered place for education.

Ideally, a teacher should be occupied in the pursuit of a vision, in capturing and making permanent something that can only be seen dimly at the moment, something that the teacher has loved with so much ardor that the joys of this world pale by comparison. This is true both in liberal arts and in applied sciences and unifies them.

But, in reality in sciences and engineering, many of the faculty are specialists who care mostly about their fields. As with any human, they care about advancements on their own terms, some are concerned only with their narrow specialty, and often largely with the rewards of professional distinction and recognition. This often means that if a student comes to a faculty and asks, "I am human and I know what I believe in right now. Can you help me evolve to more completeness and develop my real potential?" Even I would be at some loss in responding. There is plenty of incompleteness both in the person, in me, in the dynamics of the society and of the future. But this is something liberal arts has asked in one form or other for ever, and they have an answer. We develop your capabilities so that when the time comes you will be able to tackle and go forth on terms that you will be comfortable with. If a science and engineering program is quite well boxed, it keeps professors busy and they don't have to think about these being and nothingness endeavors.

This problem is not necessarily restricted to technical or vocational learning either. Humanities—the philosophers, language, arts—and specially the social science folks, who think of themselves as scientific and free of earlier thoughts struggle too. This is because this is the struggle of being human. It is a journey, we can develop and respond using the tools we have acquired and the mechanisms of rationality that we hold.

Logic is the asepsis of thinking for a student. In humanities, particularly of philosophy, it is a tool. In sciences and engineering, the analytics—formulas and coding—are all logic, whether deterministic or non-deterministic as in probabilities. Student sticks to apparatus through this stratagem. In this, the absence of worldly experience, disappears behind a fog. This is where the former with its sense of proportion and the latter with a future that is poorly divined collide. We see this often with inventions. Nuclear bombs or climate change are easy to see, they are playing out in front of us everyday, but the alienation arising from video games and the search and group-based internet think are the societal consequences that are harder to perceive. They play out too slowly to register.

In 18th century, real science was practiced by social oddities because institutions like Royal Society were a means to social climbing. One can see that in publishing and priority duel between Wallace and Darwin when Wallace came out with his pamphlet on adaptation, or between Newton and Leibniz with Leibniz's calculus taking hold. In continental Europe, the model was of a rich benefactor—Kings, Counts, Bishops, others in aristocracy—supporting the oddity, Kepler, Euler, for example.

The 19th century was about the idea of causes. In order to act, it is not required to have metaphysical beliefs that are universal, but that causes and effects can be related and that the resulting rules and other rules just like them become universal. This too is fallacious.

At the bottom of all general beliefs of this nature is a conflict with the principles of science. Laplace and classical mechanics say if we know the present completely, one can determine the future. It is not a scientific statement at all. Nor is it a literary one. Because it is not a statement about reality, either now or in future. It simply doesn't make sense to assert what would happen if we knew the present completely. We do not know the present completely and we never can. This is the great scientific idea of 20th century. The principle of uncertainty places limits on what can be known in some very special terms. Science is describing reality as being limited by limits on observation. It is not asserting anything else beyond observation. In Philosophy, this is also Nietzsche's relativism. Laplace is being scholastic not scientific. No different than the belief of karma. Free will too is simply a misunderstanding of history. History is neither determined nor random. With time, we keep moving forward into newer and new areas whose general shape can be known but whose boundaries are uncertain in a calculable way.

Science also brings troubles to the mind in the way some learned changes bring out troubles. It is a division that arises in habits with which we grew and new habits of thought that science has brought out. The two sidedness is exemplified in what we are taught to value vis-a-vis the aspiration for worldly success. Many actions of our own conduct can ashame us, but which we feel compelled to in face of the force of the society around us. Atom bomb, some of the abuses of generative *AI*, and others are symbols of this conflict. How do we choose between what we have been taught is right and something else that is succeeding. This is an empirical test of science. The empirical habit is teaching us that the traditional beliefs will have to slowly evolve even within sciences. Accepted codes of good and right conduct change. Every age needs to rediscover its own conscience. Thrift, sobriety, frugality have evolved, so have independence. This same holds in arts too. Books and paintings, for example, that are held to be harmful to public mind by reasons of being devoid of morality can become acceptable at a later time. When I was young this collection included Nirad Chaudhury, Naipaul, the worst was deemed to be Sasthi Brata, yet today, they read just fine and worthy of debates and discussion.

Scientists and engineers can be easily co-opted. Their interests are not threatened by the larger forces—social and humanist—that tend to be at war in troubled times. The connection of science and engineering to humane learning is not familial but abstract. There may be checking of boxes, invocation of rights for all, but nothing of burning shared convictions and interests. That is a shame. We cannot live without each other.

In at least USA's educational institutions, the engineering students have at least one module in teaching in their first years with ethics as a central theme. A teacher comes, picks a "case" study like problem—the Bhopal gas tragedy was common at the turn of the century, still known to the students then—and it ends up with a discussion and a set of questions that are pertinent to ethics for any generic technical project. There is little discussion of what all went wrong which led to the incident starting from the creation of a dangerous factory in a living community or of the human aftermath in short- and long-term remediation and support. The boxes to be checked were simple questions of harm or not, et cetera, with a lot of wiggle room that had no clarity. This is a subject where expediency should not outweigh a deep discussion. Future cannot be foretold and that is often a big question with the most difficult of science projects, so many of whom now tend to have warfare and human connection. This same lack of seriousness also exists with funding agencies where one often is expected to write a pro forma societal benefit paragraph.

This separateness did not exist in times past. Kant, known as a philosopher was also a natural scientist and inspiration for Einstein in this. Goethe may be known for his writing, *Faust* to us, but was also a botanist. Descartes was a logician and a mathematician who largely worked lying in his bed. Pascal may be known to many of us for the Pascal triangle, but French still grow up describing each other as being Cartesian or Pascalian in their approach to the world. For much of history, arts and natural sciences were united in being guided by being, freedom and beauty.

Much of what happens in academe is the teaching of the tools of the trade. Students of any discipline come to respect the techniques of its craft.

The goal of the academe though is to develop the mind, to think independently as oneself and with others, that is, create a collective of mind.

In a university, one gets to hear and talk to people and hear something whose power just cannot come from a book. And if one is blessed, one gets a professor whose lectures are beholden as one watches and learns from the thinking as it happens in the presence of the class. The society of mind may appear sometimes by chance during a walk, or a lecture, or by some other accident, or may emerge in a casual conversation among friends at a mess table where a conversation among friends turns into an exchange of ideas—clear with deep emotions—that start a fire.

Yet, it is also possible to hear many intellectual words but no ideas with life in them, with a campus where ghosts of mind walk around. One hears a vocabulary of ideas, like the language of new mathematics of late 1960s with few ideas, with many of these ideas dead on delivery. This is substituting methods for substance like a colorful gift wrapping of the grain of intellectual training and imaginative life.

What one needs is the development of strong character, tenacity, single-mindedness, and working in isolation. These are traits that need to be learned while young.

Liveliness of mind and acuteness of feelings disappear into nothing if there does not exist a discipline acquired over a period of orientation in principles, in how to handle facts logically, how to distinguish facts from fancy, and in the variety of ways that facts can be discovered.

A mechanized university is one where routines of the classroom, formal habits, dealing with the same materials year in and out, makes for an uncritical mind. It can destroy spirits, turn lovers of beauty and ideas into pedants, even petrify lovers of wisdom into doctrinaire practitioners, passionate people into submissive, revolutionaries into reactionaries, or more sadly, liberals lose their passion. But if the university becomes quite mechanized with routines of the lectures at 10 minutes past the hour, the formal habits of a professor repeating a course, to the uncritical young ones, it provides one with the equipment so that experiences can become enriched with meaning and this may show up later in life.

However, the worst stultification comes from outside, and not necessarily the academy. I have seen students and young people newly hired who had been awakened in the university with ideas, imagination and passion, have their spirit killed by the world. Regimentation and stunting comes comes not from the gears of the academic life, but from the gears of living.

To work around these dilemmas one has to keep an eye on the object, see any thought and thinking at its own terms, stay focused by not raising foolish and irrelevant questions, and force the discussion of central and relevant ones. Aristotle did have many beautiful ideas to discuss that have stood the test of time—*What does it mean to be?* certainly—that predate Nietzsche or Hesse or Kant or Schopenhauer or Aurobindo or Raman maharishi. And then one must act accordingly. That is phenomenology. The truth of an idea is tested by its use. That is what science tells us too.

1.4 *A teacher view*

REGARDING THE TEACHER, as Plato writes in *Republic*, it is not what the teacher but what the world teaches that will count in the long run. Teachers are catalysts. What a student can learn from the teacher comes from the habits that appeared in the first years of life and of temperament that was established in the childhood. A teacher can initiate enthusiasm, show logic towards a path, and inculcate discipline for those sensitive enough to be amenable to it. A teacher can only communicate passion, show methods, and not much more.

The teacher is the transient. A lucky teacher is one if sometime later a student feels the teacher's voice inside him or her. It is like music that is the ultimate, not the musician. A truly great artist plays a classical piece associated with the composer, and puts nothing between the composer and the audience. It is what the teacher teaches that counts, not the teacher. In accomplishing this, the teacher is both the composer and the listener.

A teacher has to put on a show, but a content-full show. Students are then attentive and there is a good chance that there will be something in the show that will capture them and they will flow with it. A show includes the teacher thinking on his or her feet in front of the class, the gentle art of arguments and derivations and what they imply, and when making an error, walking back to find that mistake, be it in the derivations or the argument. Teachers, who are intellectually brilliant, but cannot put on a show, should certainly write books since books are the other way a group of students can be enlightened.

Richard Feynman was beautiful with flow of ideas, was certainly a showman but also a magician with ideas. He could, being Pascalian, connect ideas from different streams, arrive intuitively at answers, and then go back and do the logical derivation of it. Feynman-Kac formula connecting parabolic partial differential equations to stochastic processes is a beautiful example of this. It is physics and mathematics woven together. Even pure mathematicians—a very reticent brotherhood—accept Feynman to their fraternity. On the other hand Hans Bethe was a logician—a Cartesian, he could logically follow the most difficult of mathematical paths. His Nobel for physics was the calculation of the 9-step process of fusion in the sun that lights and warms our world, a calculation reportedly outlined during a train trip back to Ithaca from Washington. Both had a giant impact through generations of students in their own way even if they had very different persona.

Feynman's *Character of Physical Law,* a profound scientific but also philosophical subject that he gave as Messenger Lectures at my academic home, Cornell, is very worthwhile viewing. It shows how arguments can be woven and a show put on. The lectures and a little more can be found at https:www.feynmanlectures.-caltech.edumessenger.html.

For a teacher, fascination with one's students leads to an awareness of the various kinds of soul and their various capacities for truth and error as well as learning. There is not much that one does for a good student, except perhaps to encourage them on by form or example.

1.5 *A student view*

Now a bit of my own student experience and its reflection in life and why I see education as a process in many stages, but most determined by the character and discipline developed in early life.

As a student, my favorites have been teachers who show the act of thinking and figuring things out in real time in front of a class. It rubbed in, emphasized not the objective but the journey, and the joy of what Feynman calls figuring things out.

Scholastic education, which is the university's purpose, is beneficial to only a small proportion of the population. Idle rich, for example, do not derive any scholarly benefits. For them, the university is a place for building networks of relationships.

As a young one on the campus, one could clearly see the intellectual and brilliant as such by the second year when all the past pre-institute learning had been normalized. Later in life, most of these are people who have done some of the most interesting things and have lived a productive life. The childhood and schooling's character building and learning discipline showed. The erudite adults wrote just as well as second year students as they do now appropriate to their ages. Campus politicians are still politicians, even if not in politics. There are some changes too that are observable. Aesthetes can become disillusioned or harsh. Sentimentalists can become cynics. But here too, the change is an evolution, not a parity change. Surprises exist too, though fewer. A very friendly and well-spoken student who was not as inclined towards the classes but still interested became an academic. A very serious and committed academic student didn't make it and spent his life on the large-company water wheel. A very bright student who always got to the heart of the matter through his questioning explored many of the varied human activities and pursuits and then followed through into academia, and a curious, always smiling, welcoming and ever-happy student chose not to get advanced degrees but did remarkable creative engineering work in India and for India. Still a very happy and satisfied person. Only the very original students, students who always sprung surprises, students that are difficult to classify, of various types, went on to being very different. Publishers, social workers, revolutionaries, and many other types, all very non-*IIT* follow throughs came from this rank. At least, one, if not more, started successful small companies with intellectual ideas. This is all type casting. It is a distribution's behavior, but therein is the essence that life brings surprises and many changes of directions and behavior to a select few. Life is full of educational surprises.

Students do well in the company of other students with whom they can build the society of mind. Students who want to progress intellectually—their educational evolution—gravitate to places where they can pursue those interests practicing, watching and being guided by people who they think will be good examples. These are the pre-eminent faculty and this is what seeds the critical mass of a society of mind resonance of faculty, students and ideas. It is no surprise that institutions evolve as this abstract collective is dynamic since the society is dynamic and the great problems of interest evolve as we progress.

1.6 A summation and plethora of cautions to the student

Institutes like *IIT* are a blessing unlike any other for students. One can see prosperity ahead, a confidence that one will make a good living. The institute provides a self-contained exceptional environment of peers, teachers, and a large support structure. No matter what the past has been, one now has the freedom and opportunity for a nonlinear blossoming whether the interest is in sciences or in engineering, or in between as it was for me.

There is much written about the extinguishing of desire of learning by coaching used to gain entrance. This is too one sided a criticism. In sciences and engineering and even in writing and philosophy and all liberal arts, it is very important to have much of the basic tools and methods become a part of the repertoire. They are to become so natural that one doesn't waste time in that mechanics. In turn this gives one the chance to push the frontiers by working on advanced themes that are not already internalized. If one did not know multiplication tables and other simple ways of doing complex mental arithmetics, the order of magnitude estimations themselves will become time consuming shorting the more important education and learning: probing of the question of the validity or the implications of order of magnitude estimates. The real objective is to do a good reasonably accurate calculation. The order of magnitude is only a step in the process. Those by-heart learnings—coaching and the criticized rote—are mechanical tools to aid in the next step. It is this opening of a new frontier that one should recognize and look for when coming to the institute.

The institutes will have plenty of good tool teachers, but if one gets at least one teacher a semester who shows one the process of thinking even as they teach, teaches one the tools, but jumps and connects various streams of ideas and shows the intuition and logic underlying it, one will learn, and that is what the education is about in all the disciplines, whether engineering, or sciences or humanities. One has to take advantage of such rare chances and learn.

The precision of scientific thinking is that truth and falsity must be decided objectively, that even though subjective opinion and personal commitment are important, they are not sufficient to making the scientist right.

One needs to be exposed to lots of ideas. But many of these ideas, unless properly understood, are only intellectual words, a vocabulary of ideas, are spoken by rote with no life in them. This kind of method cannot be a substitute for substance.

To accomplish, one needs a very strong character, a tenacity of purpose, and singleness of aim to work in isolation, specially when young. Liveliness of mind and sharp feelings can disintegrate into nothingness if one does not have the orientation in principles, of

Malcolm Gladwell is referenced for popularizing that one needs to have spent at least 10,000 hours of practice—in music or programming and in others by ˋhasty˝ generalization—based on research by Anders Ericsson. That is far more than the amount of time a student has in undergraduate college. 10,000 hours is about 8 years with 4 hours of concentrated effort every day of the year. Unlike music playing, science and engineering can be done without an instrument too. There is plenty of thinking that goes into an experiment of any importance. Same with a large software effort or in tackling complex problems.

intellectually handling facts, of discovering the way in which facts themselves distinguish from fancies and the way facts are discovered.

The thirst of experience can be sated but not satisfied if one is not equipped to endow experience with meaning. This is the intellectual training of the mind. From this comes a society of mind that is within and without. It is populated by people one has never met, or that one may have by accident, or appears as a surprise in casual conversations with one's circle of friends by transforming into an exchange of clearly seen ideas and directly felt emotions, in turn causing and spreading a mutual fire.

Sometimes it is said that a good student does not need a teacher. A good library and a time to play will do. No poor or good student deserves a bad or bored teacher. But a student is blessed if later on in life, dreaming back, she or he remembers five to ten of the people for passion for ideas, clarity of them, and the love of communicating them, and exemplifying it in in their own intellectual discipline and candor by giving meaning to facts that one would likely not have found on one's own.

But there are cautions too.

There are fashions in teaching and fashions in subjects. One has to choose carefully. One method does not fit all students. Every student has their preferences that fits with their being. The same holds for the teacher. This is one reason that different teachers and teaching styles will appeal to different students and diversity is a big help. Having some good examples around that fit with one's style works well in keeping interests peaking even as one works methodically with others that are not as appealing. It is ideas from those that appealed and those that did not appeal that will come together when attacking some interesting problem later in life.

A hasty ambition should be avoided. It is not likely to lead anywhere successful. The subjects that are popular today because of their impact on society, or industrial need, or other, will probably have a life cycle smaller than a student's working life.

As a student at *IIT*, one will find much froth since *IIT* is a microcosm encapsulation of a world envisioned, not the local Indian environment one grew in. Once out there in the world, it is like a refugee camp where unfriendly and or self-centered people idle away.

So it is reasonable to question what good is liberal arts education?

Science changes values by injecting new ideas into a culture and then subjects it to the pressure of technical change. Gradually, the whole basis of the culture is imperceptibly remade. The sensibility of people changes. The Tiktok generation followed the Facebook generation that followed the Email generation that followed the people who wrote postcards. Learning and thinking styles changed through this technology change. This will continue. I can already imagine what the *AI* tools will do for those of us who spend considerable time at a keyboard. A good liberal arts education keeps a person grounded. Without imagining monsters and having dreams, science and engineering can rapidly become demonic.

Just as liberal arts were supplanted by lucrative science with its utilitarian character, the new world has added a layer of finance and wealth acquisition on top of it. Business degree is not the moral equivalent of medical degree. But it is enormously lucrative even if there is not much by way of scholarly achievement in it. This is tourism by such students in a science and engineering institution and it is the modern time's change to that evolution from liberales to lucrativae to another turn to enormous wealth.

I will call it *Novus opulanatiae*, one of creating a new aristocracy by monetizing results of learning exploiting gaps in society and governance.

We now have *Artes liberales.*, *Scientiae lucrativae*, and *Novus opulanatiae*.

It can be seen in *USA* as also in India. Just as the example of Kepler or Euler, the institutions now turn to these products of the institutions or other major donors to support them. The dilemma here is that again, as liberales got coopted by lucrativae, the institutions—many of the private institutions around the world—are at the risk of being coopted by the agenda of the new opulence.

One of the most insidious consequence of such changes of the *Novus opulanatiae* is that appearance of competence is more important than the evidence of it. Mind is dulled when engineering is co-opted as a means to business and wealth. Words like innovation and creativity to name two in vogue lose their original meaning and new words such as monetization, thought leaders, et cetera are added into the lexicon of meaninginglessness. A recent turn has been to crown individuals as *father of a technical field.* We have a lot of such fathers, from internet to various technologies. This is the Matthews principle fallacy and a pollution of the mind with scientific and political undertones. Science and engineering are built on the shoulders of work by others. At some point streams of thoughts merge and a new integrative theme emerges. Even Einstein rode on Reimann's, Minkowski's, Grossman's and Lorentz's shoulders to his relativity. We give him great respect and recognition, but do not call him father of relativity. Take the word creativity that we have used from very early in this essay. It applies to self and culture. It is a way of expressing dissatisfaction with what is around one, whether technical or political. It used be the proposing of new hypotheses that got borne out, or finding new ways to proofs, or inventing some thing new or a new

Matthew's principle or effect is the preferential ascribing credit to one, the most known one, above all others who could lay claim. It comes from the book of Matthew 23:29. Matthew's principle is itself an example of Matthew's principle, an autological phrase.

experiment. The new meaning of creativity, although it doesn't affect science and engineering per se, is a pollution of language that is insidious. It is a form of pollution that is a disorder of our age.

Insignificant speech is a loss of clarity about science and art. It weakens both in a synthesis of opposites by appealing to the society that wants to be told that it enjoys all good things. In turn, there is loss of trust in science and engineering. So scientists and engineers hurt themselves by the constant practice of describing small steps as giant progress. This is where those in liberal arts can keep us in check.

University education should be a privilege for special ability. Entrance and continuation should depend on continuing advantageous use of the time by the student. Both the teacher and the student should be driven by curiosity and learning, so research and knowledge of work around the world is essential. New knowledge is the chief source of progress. Utilitarian knowledge needs to be fructified by disinterested investigations. It is these that help each one of us understand the world better.

Relativism, both in the liberal and the scientific sense, is necessary for openness. Just because one has learned arithmetic of addition and subtraction, or from day one religion has been forced in, does not mean that the world of knowledge and questioning must not keep opening up. The major questions in nearly all areas of sciences and engineering today are of completeness versus incompleteness. This is true through quantum uncertainty and in the classical from not knowing all. This is where relative truths arise.

Ancient worship is a common religion. One tends to build myths by suppressing faults and inconvenient story lines. As a child and a young youth, I read and noted all the European history industry, Toynbee, Taylor, Gibbon, the American history interpretation of Durant, the Indian interpretations from Majumdar or Nehru or the British scholars such as Basham. They were not satisfactory. They had in common either imperial or colonialist view or an adulatory view with much to question. Why and how and where from are important questions to understand one's trajectory which is what history is. I realize now that these books were what we will now call not scientific. All data needs to be looked at unbiased in case of the past even if that experiment cannot be repeated. It is still not possible to see a good discussion of why with all the great contributions in mathematics—hindu numerals and 0 without which one cannot even start mathematics properly being the most notable—pre-10th century and great arts and writing, the Indian science and the society declined so precipitously over the next millennium. The Greeks and Romans were not perfect, slavery abounded, so did blood games in the Coliseum, and the dark ages of Europe can be directly tied to the uncritical acceptance of Greek teachers put on a pedestal, or the Church sticking to a dogma. This is history of myth making. Native Americans had to be disposed of, shrunk into small pieces of reservations, and their former areas renamed and their languages banished to rewrite the history as Puritans so successfully did. There is even a holiday named Thanksgiving to celebrate the ``illegal˝ immigration. Now we build walls after Ronald Reagan is hailed for a speech calling ``Mr. Gorbachev, tear down this wall!˝ just a few decades ago. Slavery was hardly there in the books, nor a discussion why there were so many Scots in British East India Company and why the worst acts of brutality came from them while the company was establishing itself. In all this, people are just trying to improve their lot, and setting aside one's values to get ahead is a time-honored way. One can see it in the way how companies build dominance such as in the high technology industry. These acts are all around us. Nirad Chaudhary and Naipaul were right, Amitava Ghosh explores this well, and David Graeber and David Wengrows recent book and Thomas Piketty's brought it home. These are all books that took a more scientific approach instead of relying on dogma.

D. Graeber and D. Wengrows, *The dawn of everything*, Allen Lane, ISBN 978-0-241-40242-9 (2021)

T. Piketty, *Capital and the twenty-first century*, Harvard, ISBN 978-0-241-40242-9 (2013)

So one really has to be a skeptic. That should be the natural scientific way of conducting oneself. It helps with determining what is right in at least one's viewing through the enlightenment that the education gives one.

Learning has a relationship with the life of the community, not just of some refined delights. Disinterested learning, the learning of useless knowledge, is powerful. One never knows when one will use it, and it provides surprising powerful segues. This is a reason why search and online learning is not the answer. Both lead to channeling and rote. The pleasure of a library is not alone in the book one was looking for, but of others that one suddenly discovers as one peruses the shelves. So never lose your love for a good book.

Being human is to give one's love and care and companionship to people. A university is the place where the persona solidifies, and is the ideal place to reach across the spectrum—from the sweeper to the director—that makes humanity and be one with them. Music, besides its humane nature, also stimulates the analytic part of the brain. Music instrument learning is not easily accessible growing up. The institute is an ideal place for spending some time making this new friend. The same with learning a foreign language. Language flows from and in turn determines the thinking styles of communities. We become prisoners of language, something that media and management and governance exploits since it dulls our thinking and limits the scope of thinking. A foreign language—French or German, for example—will open the mind and reduce bias. There is tremendous Cartesian and Pascalian reading that can be very refreshing and keeps one young. Hesse, a wanderer of the soul, or Mann, an analyst of the soul, in German are more powerful than in translations. Same with Victor Hugo, an anti absolutist, or Albert Camus, an analyst of what it means to be free, in French.

While this essay has stressed liberal arts quite strongly—the rest of the essays kneel quite thoroughly over to the other side—it is prudent to hold on for later in life the bigger questions of the liberal theme. The discussion here is to encourage one to think broadly, open one's mind to the different ways of questioning since the world is an open system. Liberal thinking in science and engineering pursuit can be very helpful and keeps one human even if mostly practicing *scientiae lucrativae*. These bigger questions for the later stages of life certainly must include those related to ignorance, such as the dichotomies of reason and revelation, freedom and necessity, democracy and plutocracy or authoritarinism, good and evil, body and soul, self and other, city and human, eternity and present, being and nothing. These I leave for long walks.

What we now have in the new world is that men and women can hope to live in the same way and study the same things and expect the same from careers. This is enormous progress. For this we should be thankful. This is one of the great successes of liberal arts and sciences and engineering together.

2

Large and small: The problems of scales in semiconductor electronics

The ability to ``control´´ semiconductor structures at nanometers scale and integrate in multiple dimensions has made an integration of near-trillion scale possible using structures that are largely surfaces and quantum-mechanical-sized material. This is a non-random statistical assembly of near-classical objects. Information manipulation in this assembly must occur under constraints of energy and variability that has static and dynamic manifestation from the assembled object particles. Deterministic computing, which is largely the present paradigm, leads to a variety of consequences and constraints that set limits. Most of the modern themes—machine learning and neural networks in practice of artificial intelligence—are still subject to these since the implementations employ deterministic computation based on basic linear algebra subprograms (*BLAS*) even if dealing with probabilities. This essay is a discusion of the limitations that are far away from thermodynamic information capacity efficiency arising in such approaches, some common misunderstandings, and sets the context for what and which kind of problems under what constraints become amenable to exploration of alternative information-processing techniques.

Today, we are capable of making devices, structures, interconnects, and integrated electronic assemblies at the nanometer scale. We can pack controlled and reproducible forms that can talk to each other and operate on each other at an extremely high density on chip scale, changing connections on the fly, on package scale, and connect themselves further out in the cloud and all over the world. We can connect things that are at the Avogadro scale with the near-atomic sized small dimensions.

This is pretty much a statistical mechanics problem of predicting the behavior of a large collection of entities interacting with each other. It is similar to that of nature, which deals with it in the physical and natural world with atoms and collections of atoms into

Science and engineering in the world

molecules and solid, liquid, gas and plasma phases, inanimate and animate, with an energy flow through them. Later on, in this series of essays, I will stress that this is really a statistical problem of information. For now, just view information as a characteristic associated with the properties that we employ at any representational level to see how the physical and natural laws—the laws of evolution—show us the physical and natural evolutionary behavior in describing the system.

Event though we have arrived at the atomic scale at one end and the Avogadro number at the other end, what we are ending up doing is practicing the traditional deterministic way of computing at nature's scale. This enterprise has been a great success, but it is also a paradox in keeping the deterministic flame alive. Large numbers means that all the past hundred plus years of learning is informative and instructive. This deterministic paradigm—starting with Boole's and others' early work—will exist and continue to be enormously fruitful, but it is also undergoing an evolution to new non-deterministic directions. There are limits to what we can do deterministically. Plodding along step by step and consuming energy at each step has to have limits. We cannot just keep going down smaller and smaller ad infinitum because we would eventually get down to a single particle—be it photon, electron, or atom. Plodding and consuming energy through Avogadro-scale interactions too is not going to work. There are limits to energy as also the time it takes to complete any logical stepping. So, there is going to be a question of deterministic versus non deterministic computing that is inherent in this. I would like to build an argument over the first three essays starting with a discussion of deterministic approaches and their limitations and wondering what ideas really are there that are interesting to pursue. Ideas for the future perhaps or at least speculation about it.

A corollary to this statement of statistical mechanics of information—across nature to the computer realm—is also that this general purpose deterministic style is man made, nature is non-deterministic, so probabilistic computation to machine learning and artificial intelligence are also natural computational styles. This is tying of the natural and physical world on computation. The next two essays will dwell on these themes.

These first three essays are an exploration of the information engine and the technology in the world. I arrived at the *IIT* campus in 1971. It was a silicon bipolar transistor world by this point, we were taught from the book of Millman and Halkias, which was devoted to various analog circuits and their characteristics and also had a discussion of emitter-coupled logic (*ECL*) circuits. Transistor-transistor logic (*TTL*) circuits, which were much more fascinating because of their internal dynamics through a clever amalgamation of structure and action, was from the Texas Instruments handbook that professor R. N. Biswas used. Integrated-injection logic also got a mention, again another very clever twisting in structure-action form. Memories too appeared in the teaching. All these digital logic, analog circuits, and memory circuits were all based on bipolar transistors with a side

In discussing the future in the social and economic milieu of today in the last couple of lectures I will also pursue a few of the life lessons in this information journey and what it says to me for the science-in-society future cone that it unfolds. It too is a non-deterministic evolution in the midst of large number of interactions.

J. Millman and C. C. Halkias, *Integrated electronics: Analog and digital circuits and systems,* McGraw Hill, ISBN-13 978-0070423152

The TTL *data book for design engineers,* Texas Instruments, (1973).

discussion of *nMOS* transistors.

It was also in 1971—I did not know it then and I don't recall it being mentioned in any of the classes—that the first Intel 4004 4 *b* microprocessor arrived with a transistor count of about 2250 transistors, so an order of magnitude of a thousand, using *nMOS* technology. A few of the well-off kids had calculators. Prof. Ramakrishna and Prof. Oberai banned them. By the time a calculation was done, one could easily be off by a factor of 2 because there were enough logarithms and exponentials involved in the computation. The calculator made a difference and of course the computers made a difference. The computing world was opening up and it was not difficult to see this unfolding playing out. By 1980, when I got my *PhD*, bipolar transistors had almost gone from digital to analog usage meaning that *DEC* and *HP* and a lot of other companies were making *nMOS*-based real-time computers with direct memory access. Microprocessor had arrived and computing was starting to get democratized. So, within a decade of the first microprocessor, we had the Motorola 68000, with that number indicating a processor with 10^5 transistors. The processor was now a hybrid 16 *b*–32 *b* system. In a decade, one has moved from 10^3 to 10^5, while I was still undergoing university training. The education really blossomed at *IBM* Research, where there were many great scientists and engineers from whom one could really learn and develop the critical skills of questioning, distilling principles and meaningful laws, getting to either precise or good-enough answers with an understanding of the constraints and limits, and finding one's path in the midst of constant evolution.

In 2023 (today), the bipolar transistors still exist, but they exist in advanced technology as silicon-germanium bipolar transistors with some silicon bipolar transistors because all the multi-decade and hundreds of *GHz* transmissions, the *LIDARs* and everything else that are used in cars and other places for figuring distances and surroundings, and so on, need dense transmission and analog-to-digital conversion. Computing, by and large, is based on *CMOS* circuits that are mostly static and some dynamic depending on the needs. Central processing units (*CPUs*), graphical processing units (*GPUs*), tensor processing units (*TPUs*), artificial intelligence units (*AIU*), et cetera, all operating at high frequencies, but far more so operating in a systematic design that flows the entire processing at high speeds, are all *CMOS*. Video and data flow rules with all the machine learning forcing new designs where matrix calculations and propagations happen in ensembles without having to wait for information to accumulate. The problems have been projected in a different way. The most advanced *CPU* processors today are Apple's based on *ARM* at a 4 *nm* node. The *GPU*s and *TPU*s are at 10^{11} transistors such as for NVidia

Ramki—Prof. Ramakrishna—wore very bright colored shirts, yellows and pinks as I recall, had an incredibly flashy smile, and wrote the toughest tests I have ever encountered. Points were both positive and negative. The class average of the first mid-term test was -19. Fortunately I did well. Ramki insisted that only slide rules be used in the transport and thermodynamic calculations. So did Prof. Oberai. Prof. Oberai wore the whitest shirts I had seen until that point, management and Wall Street types came later. I still have the slide rule that I had at that time. In recognition of the personal feel that the slide rules gave to calculations and numbers as one worked through logarithms and exponentials, I have now amassed a large collection, thanks in no small measure from a gift from my father-in-law. They go well together with the abacus. The lobby display at *IBM* Research Center in Yorktown Heights, with models of Babbage's engine, and so many other creations of past history, can't be beat. These instruments are a tool to information manipulation, like function transformations, in a restricted space of symbolic mathematical manipulation.

In the early 1980s, the *IBM* mainframes still used current-model logic *CML* and *ECL*—speed, and not power, being of paramount importance. It takes at least ten years, if not more, for new technologies to establish. This is true here, same is true with the appearance of the digital infrastructure based phones, of internet-based commerce, of internet-based secure contracts and paperwork, and so many more things. All these new acts are based on the deterministic semiconductor computation and communication infrastructure.

I still have my first $IBM\ PC - XT$ bought a couple of years into my employment sitting in the basement and waiting for the day when playing the games written in *BASIC* will bring back lost memories just as the slide rule reminds me of Napier's enormous contribution to computation and the profoundity of exponentials and logarithms and infinite series in understanding simple everyday changes.

$H100$ and Biren $BR100$ at the chip scale. So, between 1980 and today, 40 years, one has gone from 10^5 to 10^{11} transistors. The processors are moving terabytes a second within the chip because of the way the buses and architectures are organized. These are $Tflops$ units, not consuming 20–30 W, but factors of 20 more than that. Just like the old mainframes of the bipolar era, they need a lot of complicated cooling and design.

These are all examples of deterministic computing. This is deterministic in the sense that they follow Boolean logic, or Leibniz's algebra of concepts, and De Morgan's laws, and even if one is making probability calculations—manipulating probabilities—it is be being done deterministically. Deterministic methods to preform indeterministic approximate determinations. *These are deterministic calculations implemented with semiconductor structures operating to deterministic limits.* Boolean logical transformations take the past states, operate under evolutionary laws coded logically and making a determination. *When they make predictions—speak to the future—it is determinism connecting a known past with an unknown future.*

2.1 *Deterministic digital computation and nature's computation*

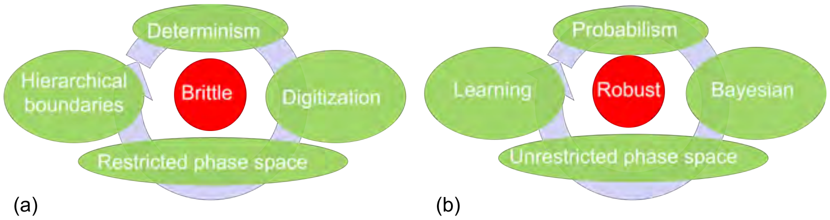


Figure 2.1: The deterministic—Boolean logical—and the non-deterministic—the behavioral, probabilistic, inexact and incomplete—universes that we compute our life in.

Computation, and life as a computed endeavor as we practice and live—occupies two universes. A highly simplified view is in Figure 2.1. This essay is about the world on the left. The following two essay will focus on the right.

In deterministic computation, we proceed step-by-step in a restricted phase space, where many states are blocked off by the computation's evolutionary law and the boundaries we build by extracting the most significant of the characteristics on which we build our computational hierarchy. We are now working with machine states that this machine may traverse. The restricted phase space is this machine state space—a collection of the logical states of all the elements in this machine whether it is in the memory or in the registers and the arithmetic logic units, and others—a state space that has been flattened out into a bit net configuration, with the path traversing the space in time. This state space is the computational analog of the deterministic phase space of Lagrangian or Hamiltonian mechanics. It is a different way—digital bit representations instead of continuous position and momentum variables—of looking at the machine universe. For some reason, if one somehow jumps out of this universe to a state that is not in it, one gets the blue screen of death. The enormously successful digital enterprise has been able to use this determinism, the digitization, the implementation of the Boolean logic in computation—the logical transformations are the evolutionary laws—in the operating system that implemented the computation's flow. The input data was our boundary condition of the starting states, and the computational software that combined both the flow and the data could be admirably implemented, used, and reused, by staying within the restricted phased space.

We make hierarchical boundaries so that only the essentials are to be kept track of. This makes computation manageable for problems that have to deal with only a very very very small amount of data such as transactions involving digitally measurable money, or digitally computable calculations, or writing and manipulating text in digital forms, and so on. The description on the right is of the natural world, of Avogadro-scale numbers that can not all be kept track of, or even measurable to utmost precision, and so are inherently non deterministic. We humans work with such incompleteness and indeterminism constantly, sometimes making a fool of ourselves, sometimes computing with aplomb. The worldly space is an unrestricted phase space. While we may be wrong a fraction of the time, it is a much more robust way of doing certain set of problems—the problems of inference in midst of incompleteness. If one has an understanding of how the probabilities are of jumping from one machine-state condition, to another further out in time, one can skip a lot of computation and get it right with some acceptable probability. This is what non-deterministic computing performed on deterministic apparatus can also achieve and is what goes under the moniker of artificial intelligence these days.

The deterministic computing may be brittle, but is very apt for a certain sets of problems. It is restricting itself to a small subset of states that can exist. The instructions, the data, and the precision decides what that space and what coarse- or fine-grainness is going to be as it moves along a machine-computed trajectory.

Data is not information. Information and knowledge are words that are loaded, scientists use it in one way by ascribing some precise definition to them, Shannon's being one example for information in a channel, but philosophers would explore them with a very different lens. This too, I will come back to in the last two broader essays.

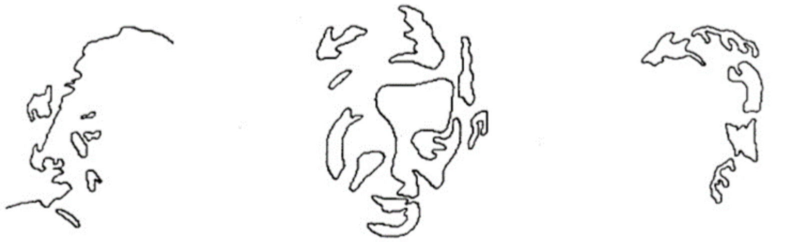


Figure 2.2: Three line sketches. What are they describing? What information are they conveying? Figure is from Olshausen (2008).

A good example to clarify how context and some pieces of data are more informative than others can be seen through Figure 2.2. What are these lines describing? Answers may vary from some abstraction of geographic visual, things floating on a surface, et cetera. Figure 2.3 fills regions, that is, identifies dark versus light in the sketches and suddenly they appear as faces. A small set of bits sufficed to give meaning to the data sketched in the drawings.

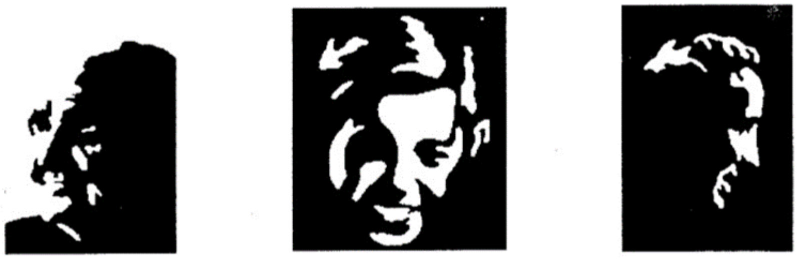


Figure 2.3: Using a few bits of information describing the regions to be darkened, one can see them human faces. Figure is from Olshausen (2008).

Try to get a computer to figure out whether it's a map or a face without that fill information. Just as for human, it will be a difficult task. This illustrates that there is certainly an interesting conjunction of the kind of problems one may wish to be working for in the future and the kind of problem one can do now. It is the second set that is more challenging, not that the first one doesn't need continuing progress too.

In deterministic computing, we abstract across scales, on time scales and on size scales, so that we get important abstracted characteristics right and eliminate the other excess information not relevant to the task. This is one way of taking tasks being performed on 10^{11} objects on the chip scale, and Avogadro number on the cloud scale to reduce down enough that they can be tackled deterministically.

Even with this reduction, statistical mechanics and thermodynamics are important. Try keeping track of 10^{11} pieces of information. Take the smallest size-scale devices with quantum-mechanical description important to understand the behavioral properties of the devices. Take a bit bigger, connecting devices, with signals stepping in *ps* and bouncing back-and-forth, or take these small signals and expose them to fluctuations between the devices programmed in at the time of creation, or taking place in time since energy causes change, or just the presence of thermal or shot noise, and we are in the statistical realm. 10^{11} device structures, with dimensional range from atomic scale to 10s of *cm* and earth-scale in their presence in the cloud, or switching taking place at *ps* to slowly drifting in *s* means that different physics models need to be incorporated depending on the context of the specific situation to be modeled. Models can be physical in how the device physical behavior is modeled—quantum-mechanical as well as various progressive approximations of methods of moments, structural—how chip-scale timing, routing, power, electromagnetics, and other matter are dealt with, and behavioral—the high level abstractions of system design. The quantum-mechanical effects manifest themselves into some emergent property that one needs to feed into the next level, for example, a quantum master equation manifests as a on-equilibrium Greens function description, which in turn feeds into hydrodynamic and Boltzmann form, eventually into the drift and diffusion. But even this grossest of simplifications will fail if one were trying to assemble the effects of hundreds of such devices coupled together in a sequenced chain. One needs compact models that tell us the current, voltage, charge and time behavior in manageable forms with some idea about noise and cross-talk. Add to this thermal effects because energy is being dissipated. Even this needs to be abstracted further for higher level design of large circuits, registers, arithmetic-logic units, memory, et cetera being put together. Hierarchy building on hierarchy, capturing abstractions, is still a building up of complexity, and approximations building on large numbers of units interacting.

This brings in the relevance of statistical mechanics and thermodynamics, which it is also another reason why deep neural network techniques become ever more appealing. There are parallels in complexity in the hierarchy and assembly of large numbers to the complexity in neural networks deploying nonlinear transformations and weights still implemented through *CMOS* logic gates. For some problems, the newer techniques may be quite inefficient, but for some, they will be more efficient The neural networks are behavioral, they employ probabilistic principles even if they approach all calculations deterministically in order to figure things out. Being a very

There is a false presumption that neural networks are enormously inefficient in energy. Consume large amount of power, something of the order of 10× for searches, for example, compared to conventional map-reduce based methods. Relevance of the search output also matters. The training of the neural network is an accumulation of prior learning, just as we humans are learning building on the past five hundred years of science and our own living.

different approach, they are useful for a range of complexity that the conventional deterministic methods are very poor at. This is the dichotomy between the two tuniverses.

Nicolas Lenard Sadi Carnot and Rudolf Clausius, among the first to explore thermomechanics, and developing the early understanding of energy conversion and efficiencies, taught us how putting energy in can lead to useful mechanical work, the relationship to heat, and the processes—time scales and exchanges through isothermal or adiabatic processes—interrelate. Carnot bequeathed to us the Carnot efficiency limit of conversion and Clausius the term entropy. Without dwelling on the enormous confusions and fallacies involved in terms like useful work and entropy and how one may interpret them, leaving them as nebulous terms, the underlying idea of engine of energy conversion has been profound.

All engines are engines of information. A mechanical effort is an informational change even if because of our natural development of how we look at these matters from an early age there is a physical transformation that we ascribe to the notion of work. The physical transformation is reflecting an informational change. Information is physical. To write a number 1 or a letter a, one needs a physical representation in chalk or binary representation in voltage of magnetization or current form. Same with smell in the olfactory structure. Same with music in the auditory structure and in how we convey it to be reproduced. A block moved from one place to another is an Avogadro number of particles in a new place. Even for mechanical purposes, this is a large information movement of what is important for mechanical purposes. For example the emergent property of mass that encapsulates it. All engines and all actions involve manipulation of information.

The reason mechanical work or heat are fallacious is because they are not really definable. Objective science demands that. A wheel turning doesn't move yet is doing work. So, we introduce pseudowork. Try understanding work in the context of a point particle like an electron undergoing a spin angular momentum flip! This same fallacy holds for heat. The pedagogical treatment starts putting a slash in the form of $đW$ and $đQ$ to call out path dependence. Energy conversion efficiencies now start depending on path. Do it slow enough—adiabatically—and one can be efficient and makes us face Zeno's paradox. Heat is not necessarily motion, sometimes it is, sometimes not. Spin glasses exist.

Hartley, Shannon, and Szilard, by exploring these conundrums taught us to look at all these energy transformations through an information view. Lack of information is increasing entropy. Not randomness, which is another vestige fallacy still taught. Heat is information content lost, and this places entropy in context with the vestige from the past.

The data is information is the same fallacy as embedded in useful work and heat. Whose work and how useful? What heat? Sometimes one needs warmth and it is useful. Same degeneracy issues pervade data.

The electronic engine is a apt place to show these notions at work. That data is not information through the example of Figure 2.2 and 2.3. Shannon's notion was related to useful transmission of a collection of content from one point to another. What content says and means what the interconnections in the content is is for one to figure out. To Shannon, faithful transmission of data was the core, and he labeled it as information. Even if information may also exist in addition and is not represented in his metrics.

Information manipulation leads to informational effectiveness in executing some action. This is what information engines let one achieve. There is energy involved since any change in information content, extracting informational value and leaving the state in an unknown state, involves energy. A state of maximum unknown is a state of maximum entropy. Extracting a single bit's informational value is $k_B T \log_2 2$ of energy following which one does not know what state it is in. The information engine is using bits encoded and achieved through a large ensemble, so plenty of Boltzmann and Gibbs statistical distributions also appear in the energy transformation.

We now have a picture of the emergence of features of this electronic engine. In steady state, the chip is undergoing activity at some factor α of the cycling. Let U be the energy expended in the active transformations and let A be the area of the chip. There is energy loss of informational unknown—the traditional Q—and one has a time constant of the engine response of

$$\tau = \frac{\alpha U}{QA}. \tag{2.1}$$

If there is single-dimensional flow of heat, such as dissipation occurring over arrays, semiconductors like Si have a heat carrying capacity of about $10^2\ W/cm^2$, which places $\tau = 5\ ns$ as a time constant. If one was fast in some isolated element, with heat flowing three dimensionally, $Q \approx 10^5\ W/cm^2$, and one can reach a $\tau = 5\ ps$. The former is a limit with activity over arrays, such as large matrix engines, the latter over clock generators, such as ring oscillators. This is an example of emergence of averaged properties. The specifics of the engine were irrelevant.

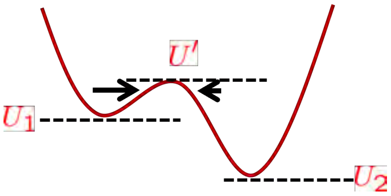


Figure 2.4: An energy-state landscape schematic where one attractor state is higher in energy than other and are separated by a barrier.

A simple understanding of determinism can be drawn from an energy-state landscape picture as seen in Figure 2.4. The abscissa is a generalized state coordinate. There are two attractor states, call one 0 and one 1, a low state and a high state. If one is transitioning from one state to another, one has to surmount a barrier. One has to remove this barrier—some active energy is needed—and let the state flush out and go into the other state, for example from left to right. For right to left, one would have to raise the energy of the state too. The figure is of two energy wells separated from each other. Changing means some combination of raising well energy and possible change of barrier. Change can also happen by fluctuations. A fluctuation taking place in this barrier—this system is in a reservoir of the universe at some temperature, say 300 K at the least.

Trapped in either of the wells, the microstates captured by the 0 or 1 macrostate, has a chance of transitioning to the other macrostate because of the statistical nature of the Boltzmann distribution describable through the entropy. There exists a rate constant of transitioning from one to another with an Arrhenius factor. It is this possibility of transitions, when not expected, that is an error. The rates are exponentially related to the energy barrier normalized by the thermal energy. They are inverse functions $\exp[-(U' - U_1)/k_B T]$ and $\exp[-(U' - U_2)/k_B T]$ with this factors representing the rate probability of transitioning from one well into the other. The temperature has shown up as a thermal source of excitation. We like the barrier energies to be very high is so that this error rate is much smaller.

This is such a simple picture, but so elegant and beautiful metaphor for the schema relating Boolean logic to deterministic computing. Details are of course more demanding. The $CMOS$ transistors form such a well structure in the inverter form and in the variety of multiple fan-in and fan-out forms. Gets complicated, but one of the important simplicity there too is that one can make very poor transistors, they may be off from what was planned, yet in a static gate form, they are going to have one state near the lowest potential, usually a ground, call it the 0 state, and the other near the power supply, a high, call it the 1 state. The electrochemical potentials of reservoirs primarily determine the two attractor states.

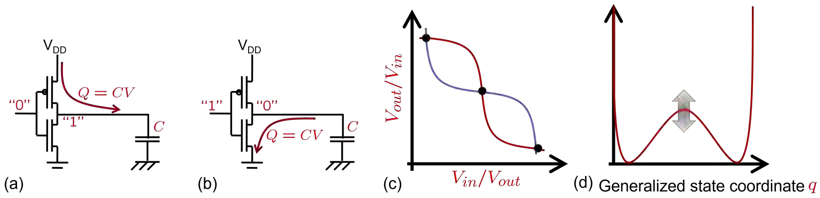


Figure 2.5: (a) and (b) show the switching of a $CMOS$ gate which takes charge from the supply and puts into ground dissipating the energy QV, (c) shows the butterfly curves of input-to-output and output-to-input, where the two ends are stable points of the two attractor states and the middle intersection is an unstable state. (d) shows a generalized picture of the energy landscape, where transitioning will requiring raising of the well energy and the changing of the barrier energy.

The shortcoming of $CMOS$ is of course that in one switching event, where the output is going from low to high, the load transistor places charge on the interconnect line represented here through a capacitor and then when it goes to a low state, the charge from the high state that was there on the interconnect line goes to the ground.

Much of the energy one extracted from the supply has been lost to the environment and only some used in the Boolean-coded information. There are only a very few places where one can actually recover that energy and those are ones where both of these terminals have to be floating clocks. Sometimes this is quite acceptable, but they are hard to design. You can see in this picture of *CMOS* drawn in a butterfly curve. With V_{in} driving the abscissa, one gets a V_{out} following the red curve. If this was fed in, so V_{out} is now inputted, one gets the second curve. There are two intersections and an unstable point where the device gain changes the energy barrier and pushes the states from one to the other. The stationary states here, one low near the ground and one high near the supply, with a separation and ability to recover since the large gain of driver and load (a negative large number) forms a window over which the low and high states can fluctuate and yet this state will remain pretty much the same. The Boolean aspect is going to be maintained and that picture is essentially this picture. There is a barrier in energy that one has regardless of which state this system is in.

If one takes two *CMOS* gates and connect them back-to-back together. one gets a reinforced stability, with each driving the other as logical inverters. This is static-access memory, using deterministic mechanism feeding on each other. It is bistable. It is stable in either of the states.

This picture for memory is instructive. Take ferroelectrics. They can have a spontaneous polarization up and a polarization down. These are are also stable states and there are some others similar stable forms too, for example, tunneling diodes attached back-to-back. By and large, we don't use these. There are too many shortcomings to list here. But, they are examples where bistability keeps them undisturbed except through the fluctuation mechanisms where they must overcome a barrier.

The contrast to this are random walk memories. They are much more common because they are much more dense. A transistor connected to a capacitor, with the capacitor connected as either a stack or a trench capacitor, with the leakage from the capacitor being controlled by pinching the transistor. A 40–25 fC charge, few tens of thousands of electron, which are periodically backfilled is the assignment of the state. Something similar happens with resistive memories. These are all structures where flow paths are being pinched. It is not a barrier in the path, but just an impeding of the leakage. If one makes a material more insulating so the current leakage is much smaller. An example is by transitioning to a more resistive amorphous phase. In the crystalline state, it conducts more. The change between the crystalline and amorphous forms is is slow in time. But we know from thermodynamics that of course crystalline will always become amorphous if you wait long enough. That amorphization is the arrow of time. This is the consequence of the second law of thermodynamics. It is a higher entropy state of that system.

The ubiquitous non-volatile Flash memories are similar pinching of leakage. One stores charge, but as a result of where the charge is and the higher electrostatic energy, it has propensity to leak. Except that the insulators are generally very high quality and leakage mostly small except in the poorest of the structure. The charge does leak out if one makes these memories poorly and there are leakage paths on the edges or elsewhere, we have to take care of that.

The spin-torque memories are magnetic memories that too are bistable. They become stable because whether if one thinks in terms of spin precession or one thinks in terms of polarization changes of the magnetization fluctuating under the statistical conditions, the picture is very similar. One has a time constant τ related to the propensity to appear in the other state by spinning over or random walking into it. The spin magnetization and what that barrier is and what that barrier does is related to the anisotropy. One can go down as low as 75 k_BT in energy and it will remain reasonably stable sitting there quasistically. If one keeps flipping it, it will now lead to more errors because one is energizing the system. Flipping more means more errors will appear.

So, even in deterministic conditions, whether staying quasistatic or changing, there are going to be errors that will arise in thermodynamic causes.

With logic, such as with *CMOS*, for the errors in deterministic Boolean logic, one is interested in not just the state persisting in time, but also in the expected change (or not) when inputs change. With the thermal environment under the electrochemical bias and ground, one can view this change as a fluctuating response curves as shown in Figure 2.6.

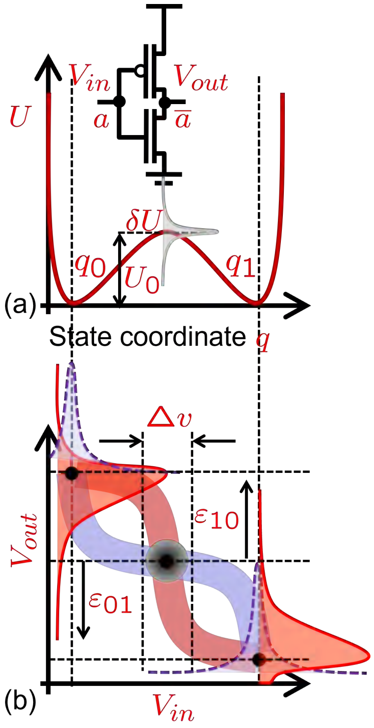


Figure 2.6: A pictorial view of a biased *CMOS* inverter at temperature T.

The barrier that was there is now fluctuating, the transformed equivalent picture of this situation is of course that there are fluctuations taking place in how that barrier is moving in between on those curves as the switching takes place. If the cause of fluctuation is thermal, it is independently random, the errors are going to be related to also how many times one is switching, and what the thermal interference is. Considered independent, this means that there are distributions, shown in the figure, where when 1 is expected a 0 is observed and the converse. The errors appear whenever one passes the threshold. With Gaussian distribution, a gain g in the transition region and a change as a function of input characterized by a sensitivity function S, the probability distributions and from these the errors of the integrated area under the curve on the wrong side of expectations is

$$\begin{aligned}
\mathfrak{p}(0) &= \frac{1}{\sqrt{2\pi}}\left[1 - gS(v, v_x, \Delta v)\right] \exp\left[-\frac{C(v - v_{inL})^2}{2k_BT}\right], \\
\mathfrak{p}(1) &= \frac{1}{\sqrt{2\pi}}\left[1 - gS^{'}(v, v_x, \Delta v)\right] \exp\left[-\frac{C(v - v_{inH})^2}{2k_BT}\right], \\
S(v, v_x, \Delta v) &= \frac{1}{1 + \exp\left[g(v - v_x)/\Delta v\right]}, \text{ and} \\
\varepsilon &= \frac{1}{2}(\varepsilon_{01} + \varepsilon_{10}) \\
&\approx \frac{2}{\sqrt{2\pi}} \int_{v_x}^{\infty} \left[1 - gS(v, v_x, \Delta v)\right] \\
&\quad \times \exp\left[-\frac{(v - v_{inL})^2}{2k_BT}\right] d\left(\frac{v}{2\sqrt{k_BT/C}}\right). \qquad (2.2)
\end{aligned}$$

The equations follow from the simple argument that there exists $(1/2)Cv^2$ of energy associated with the potential degree of freedom. The wiring is all in an environment into which the particles—electrons—flow in and out. This is a second power law and therefore it is $(1/2)k_BT$ of energy from classical equipartition.

The operating of this $CMOS$ gate means that there is a rate $\alpha\nu$ in some frequency—a clock frequency, for example—with an architecture and function dependent activity factor $\alpha < 1$ —at which these voltages are being sampled, and they must remain *deterministically correct.* This is the error arising entirely due to thermal fluctuations. With the error quantified, the problem can be reversed.

One of the beauties of $CMOS$ gates is their back-to-back coupling of inverters as bistable memories: the static random-access memory to which we add additional pass transistors for accessing and isolating from an array. These back-to-back inverters are self aware. One uses the same gate configuration for stable memory and by coupling to this same form as sensing and writing amplifiers, one can read and write through the access transistors.

What does Equation 2.2 say if one takes a chip with 10^{10} gates, a clock frequency of $1\ GHz$, a gain of $g = -10$ with $\alpha = 0.1$? The largest integration scale today is 10^{12}. The reversal tells us that to have only 1 logic computation error in 10 years of operation, an energy of $260k_BT$ per gate is necessary to overcome thermal causes.

There are lots of sources of fluctuations, some of them are even programmed in at the time of fabrication such as due to the various structural factors and the doping and so on. Arising in multiple causes, some independent, the net is again a Gaussian. It is of the skewed variety and is observable in the variations of the threshold voltages of the different transistors being employed. The energy put into the switching and the opening up of a large enough window for the tail to decay places limits on the scaling of voltages that can be employed. If one incorporate these factors in too, then the standard deviations necessary consistent with the biases that can be employed also show trends of the finesse demanded at high integration. Figure 2.10 shows a plot of this for a simplified but instructive model 10^{10} transistor calculation. If one wishes a high yield of six 9s, wants to use 1 V bias voltages, then standard deviations of 10 mV order are needed. One may do some adaptation and some corrections in order to overcome some of the errors, but one has reached a limit that is bounded by some function of integration, energy, bias voltages and the level of control one must be able to exercise.

This issue of fluctuations and the deterministic correctness expectation shows up in a very significant way in the random walk memory structures where the memory state is not a low energy minimum but a slow-drip unstable state. This is the issue of *variable retention* in dynamic memories and the that of testing and replacing columns of devices where a single device may be poor, and of more complex architectures requiring testing and pumping of more charge in to the quasi-nonvolatile charge storage memories.

The variable retention issue arises in the Poisson distribution of singular defects that happen ever so rarely. All one needs is one defect near the transistor with the transistor affected by trapping and detrapping.

The energy picture of variable retention is shown in Figure 2.8. There are good cells and ever-so-rarely there are poor cells. Because there is an electron charge stuck nearby, the pinching barrier changes causing the capacitor to leak much much faster. These random walk related variability issues are pernicious and are there across the disciplines when one looks for it.

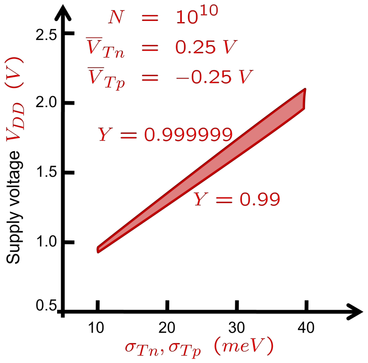


Figure 2.7: The bias-standard deviation behavior for an integrated circuit of 10^{10} transistors.

A 40 fC of charge on a capacitor is about 250, 000 electrons. If one wishes to refresh a cell say every ms at the worst, and yet integrate 10^{10} of such cells, then one is deep in the tail of a distribution. This means that the transistors have to have aA leakage currents—they need to leak at the rate of 10 electrons a second—so that the worst case can still be compensated for by refreshing in ms. It is an incredible feat that they work so well.

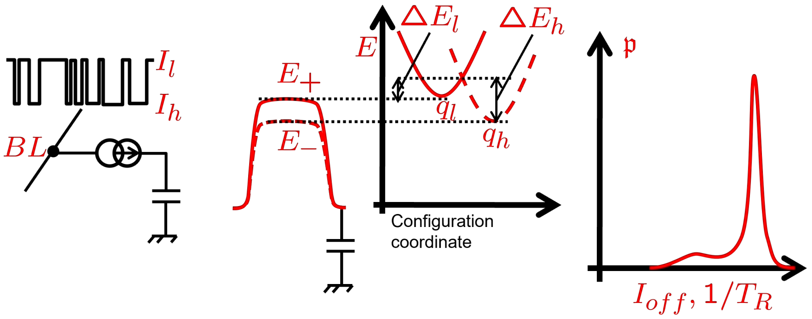


Figure 2.8: The problem of variable retention influenced by Poisson statistics of rare defects in random-walk memories such as the dynamic random access memory.

Its most immediate signature is random telegraph signal behavior. Random telegraph signal is observable in recombination in transistors whenever a shot-like behavior happens rarely. Defects trapping and then re-emitting ever so rarely is a shot-like behavior. In dynamic memories, you can push the barrier up and reduce the leakage a little bit by the back bias but you still have to cut the leakage currents significantly. Such memories have to be refreshed at least 100 times faster than what we would normally have.

Alternatives are always interesting, first for technical reasons since they exercise the mind, but second, because often some are so thoroughly unsound, yet, somewhat in desperation, we welcome them and quite often hide them behind the word-de-jour, interdisciplinary, for example. Rolf Landauer, who saw ferroelectrics and tunnel diodes as first such sorties at *IBM* during the late 50s and early 60s, the former deficient for reasons I will discuss, and the latter because of limits to voltages, currents, and the variability, but most of all for the absence of directionality in the computation, once remarked to a speaker, ``But you must give us a reason for us to leave our ship for your spaceship.''

These two examples illustrate the consequence of determinism to the traditional implementation of Boolean logic in deterministic form in semiconductor technology. Computation however is not just this data transformation, but also needs data flow, whether through short or long interconnections in all the different forms. This need now connects Shannon's description of transmission of information.

Shannon tells us what the channel capacity—the rate that one may not exceed—of a Gaussian channel given some signal to noise power ratio (S/N) and a bandwidth B is $C = B\log_2(1 + S/N)$. If one has a 100 GHz interconnect transmission channel, and one is employing low voltages such as in low-voltage differential signaling ($LVDS$), maintaining timing acuity means that a channel capacity of about $2B$ is needed. For a 5 GHz clock, this is a $S/N = 3$ for the $LVDS$ approaches, which with 10 % duty cycle needs about $14k_BT$ energy for each bit. This is another 10% on top of the thermal fluctuation constraint and the minimum supply constraint. *Everything adds up to increasing energy because of the need of determinism be it logic or memory.*

This entire discussion up to this point is from semi-classical consideration of constraints.

When faced with limits, numerous alternatives are forwarded, some of them become useful, and in the most special of circumstances, a new technology is born and takes over. The replacement of vacuum tubes was on of such fortuitous turns, but more often than not, many such forays are wishful and often filled with fallacies.

This introduction to the current state of computation with semiconductors in this technical sequence is an appropriate moment for such an exploration for its educational value.

2.2 *Transistors: Fallacy of placing quantum wells as another bottleneck in a transport path*

We start with quantum structures based ideas for electronics. Quantum as a central operational principle has been enormously successful with the now ubiquitous quantum lasers, and although jury is still out for quantum computing, there is much algorithmic and intellectual that has been learned and applied in machine learning that will be the subject of the third essay.

Quantum's success in communications, where individual device's

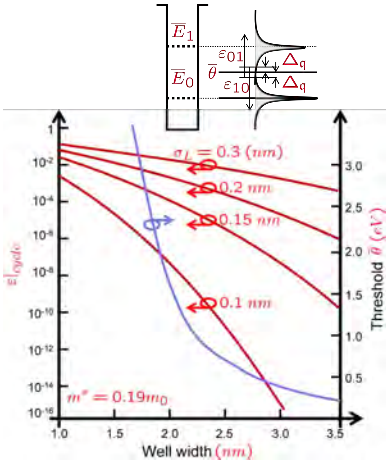


Figure 2.9: The threshold characterizing two energy subbands and relationship to any variations characterized by the standard deviation in the well width L. A threshold delineates the error of the spread causing errors in identifying the association of the operation with the binary operation.

fidelity is critical and one doesn't integrate at gigascale, therefore is not statistically constrained, is the first point to be stressed.

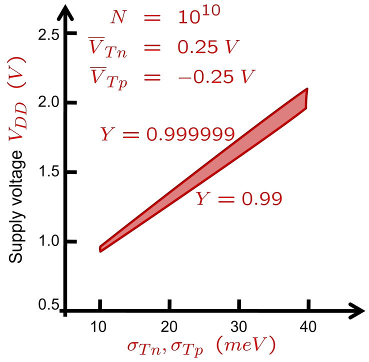


Figure 2.10: The bias-standard deviation behavior for an integrated circuit of 10^{10} transistors.

Use of quantum wells within device structures is the one such instance. At the simplest, a one-dimensionally confined well, with in-plane freedom movement has subbands. Take a structure with two subbands so that one can associate a binary form to them. For an idealized infinite well, the energies are $n^2\hbar^2/2m^*L^2$ with $n = 1, 2$ and m^* as the effective mass. There is an inversed square dependence on the confined region's width L. This large inverse polynomial dependence on the size means that the energies are going to change significantly, an estimation of which is in Figure 2.9. The equation for error, following similar arguments as those for the $CMOS$ gates are

$$
\begin{aligned}
\varepsilon &= \frac{1}{2}(\varepsilon_{01} + \varepsilon_{10}) \\
&= \frac{1}{2}\int_{\overline{\theta}-\Delta_q}^{\infty} \mathfrak{p}(E_1)dE_1 + \frac{1}{2}\int_{0}^{\overline{\theta}+\Delta_q} \mathfrak{p}(E_2)dE_2 \\
&= -\frac{\sqrt{C}}{4\sqrt{2\pi}\sigma_L}\int_{\overline{\theta}-\Delta_q}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{\sqrt{C/E}-\overline{L}}{\sigma_L}\right)^2\right]\frac{dE}{E^{3/2}} \\
&\quad -\frac{\sqrt{C}}{2\sqrt{2\pi}\sigma_L}\int_{0}^{\overline{\theta}+\Delta_q} \exp\left[-\frac{1}{2}\left(\frac{2\sqrt{C/E}-\overline{L}}{\sigma_L}\right)^2\right]\frac{dE}{E^{3/2}} \\
&= \frac{1}{4}\mathrm{erf}\left(\frac{L_{\overline{\theta}n1}-\overline{L}}{\sigma_L\sqrt{2}}\right) - \frac{1}{4}\mathrm{erf}\left(\frac{\overline{L}}{\sigma_L\sqrt{2}}\right) \\
&\quad + \frac{1}{4} - \frac{1}{4}\mathrm{erf}\left(\frac{L_{\overline{\theta}n2}-\overline{L}}{\sigma_L\sqrt{2}}\right),
\end{aligned}
\tag{2.3}
$$

where $C = \hbar^2\pi^2/2m^*$ characterizes the constant of energy scaling arising in the effective mass. The smaller one goes in size, the larger the energy spread, but far worse is the scaling in error since one is now at discrete size limits. A 1 nm width is three interplane spacings of common semiconductors. This is not a square-root limit for variability. It comes with large discrete jump.

This broadening effect is well understood as a homogeneous broadening in semiconductors. *It should be obvious that such structures are not even conducive to integrate a hundred devices, let alone the peta scale of transistor technology.*

Linewidths of bipolar lasers remain narrow since the effect is simultaneous for both the conduction and the valence band unlike the case of unipolar semiconductor lasers. The broadening arising in uncertainty is related to the leakage that transport through a constrained well region wishes to exploit. Being tied for this limit to $\Delta E \Delta t = \hbar/2$, any expectation of 10 ps time constant would be a requirement of 0.033 meV of linewidth. The homogeneous broadening will therefore prevail as seen in the standard deviation's impact in Figure 2.9.

2.3 *Dynamics and statics are different: Fallacy of Schottky barriers in conduction path*

Leakage of an off-transistor that needs to be turned on to operate at low energy is an important design issue in transistors. Having a transistor be well off is a constraint on the threshold voltage. So, proposals have been made to incorporate weak Schottky diodes as injection and collection structures. This is in a way similar to the use of quantum well in series, where flow through through the subband is being turned on or off. With Schottky diodes, it it is the contact regions as shown in Figure 2.11.

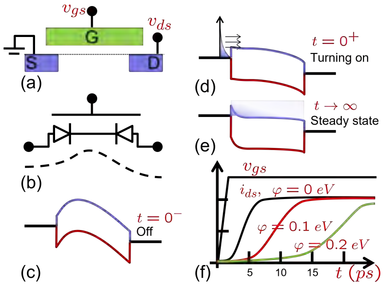


Figure 2.11: Placing a weak Schottky barrier in series at source and drain and its effect on the dynamics of the transistor.

Placing Schottky diodes does reduce current. It is placing an impedance in the path that is being controlled with gate modulating the region underneath it which includes the channel that is serially linked to the diodes. Using metal electrode instead of doped electrodes metal electrodes with barrier heights, 0.2 eV, for example, will certainly have some consequence of how much current can be passed through too under steady-state conditions. But, the more pronounced consequence is in the dynamics of arriving at the current conducting state. There exists an abrupt barrier right at that interface with the Schottky diodes. What one does when one places doped electrodes as contacting regions on either side of the channel and apply an electrostatic potential to them, is that one is controlling the electrochemical potential at the edge of the channel. The quasi-Fermi energy is nearly a constant at the interface. There exists enough thermal flow of carriers that only a small excess is sufficient to supply the current. A small perturbation is near constant quasi-Fermi energy. This says that there is no impedance arising in the interface. The actual current is being limited by the channel region—a p-type region for the n-type transistor drawn—through which the flow must take place and it has to be consistent dictated by current continuity. With doped regions, carriers can stream in and carriers can stream out from the source and drain contact regions.

With a Schottky barrier, even one with say 0.2 eV barrier, electrons coming in have to jump over this 0.2 eV barrier or tunnel through this barrier. This is now rate limiting. If the device is off, this tunneling does not exist as seen in the (c) panel of Figure 2.11. The reservoirs have limited connection, the barrier height itself has reduced the injection by a $\exp(0.2/k_BT)$ factor. Diffusion will not do. Turning the gate voltage on, such as in panel (d) has limited ability in controlling the channel region while it starts from the off state, and electrons can not be easily supplied since the source and drain reservoirs are impeded by the barrier. Most of the field tying the gate charge ends up in the metal contacting region or deep in the substrate. This is a situation more akin to the *MOS* capacitor than of *MOS* transistor. In a *MOS* capacitor, thermal times prevail, not the fast transport times. *This is the problem of the dynamics versus statics.* This structure looks perfectly fine if you look at it electrostatically in steady state. If we apply the voltages, wait long enough, it'll pass the current, perhaps with some reduction in current. But to get to that point is limited by a dynamic resistance of $k_B T/ei$, where i is small in off state, and for a 0.2 eV barrier, the limiting consequence of the barrier is $\exp(-0.2/0.0259) \approx 0.0004$. This device has latency with orders of 15 ps or more delay consequence arising in the short barrier height Schottky barrier. It arose because of the absence of a direct control of the electrochemical potential of the channel side of the junction.

2.4 *Dimensionality reduction has reciprocal-space funneling constraints*

The next example is that of attempting to exploit large mobility improvements, that is, reduced scattering, on reduced dimension structures. This problem too has statistical foundations. If one takes a silicon transistor or any other transistor that has a two-dimensional conducting channel, so a Si-like inversion layer or a heterostructure-like accumulation layer, and one places a a three-dimensional region for injecting charge adjacent to it, there is an interface resistance arising in the occupation of states in the confined region connecting to the occupied states of less-confined region. Fortunately, through suitable choices in doping in contact region, the impedance can be kept small and the contact treatable as ohmic. When one has a classical metal with metal contact, one of layers say few atoms thick, because of the large scattering that happens in metals, and at metal interfaces, means that the resistance will be low. A very heavily doped region of Si interfacing a Si inversion layer, again the large scattering in the doped Si region means that there will be a high probability of a stream directed towards the confined inversion region. The same is true in heterostructure transistors.

The Schwartz Christopher transformations give us a tool for modeling these situations of dimensionality changes and of sharp corners where the abruptnesses can be mapped so that one can view and model the flow patterns not unlike a large diameter pipe interfacing to a small diameter pipe. It is the flow in the extended regions that determines the net resistance arising in the presence of the interface and causing a three-dimensional electron velocity distribution to transform and connect to a two-dimensional electron velocity distribution. Rate-limiting region is spatially stretched out and can be modeled and it does it pretty accurately.

Now consider a metal and a high mobility, that is, one with very few scattering events reduced dimension region interface. Figure 2.12 shows the previous case of a silicon inversion layer—plenty of scattering—and the case of limited scattering. How do electrons in the metal enter and transport through the two-dimensional ballistic

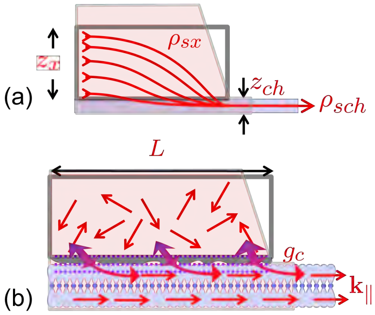


Figure 2.12: A three-dimension to two-dimension region interface, where the mobility is poor in both is shown in (a). In (b) the reduced-dimension region has high mobility and quantized conductance prevails.

region of atomically thin layer below? They have to scatter in. Some with low-angle scattering some with high-angle scattering, with the lower-angle more likely. Only some scattering processes allow some of these carriers to jump in. The scattering process is going to be a rate limiting step. A simple transmission-line model shows that a contact resistance of

$$R_c = \frac{(2\alpha\beta + \beta^2)^{1/2}}{2\beta} R_q \coth\left[(2\alpha\beta + \beta^2)^{1/2} L\right], \tag{2.4}$$

where $\beta = (1/2)g_c R_q$, and $\alpha = \rho_s / R_q$ exists spanning the diffusive and the ballistic limit. α characterizes the scaling factor of quantized versus diffusive transport. When in ballistic limit, $\lim_{\alpha \to 0} R_c = (R_q/2) \coth\left(\frac{1}{2} g_c R_q L\right)$, and when in diffusive limit $\lim_{\alpha \to \infty} R_c = \sqrt{\rho_{sch}/g_c} \coth \sqrt{g_c/\rho_{sch}} L$. The important implication of the behavior, physically viewable, is that, very long contact regions are needed to make a low resistance contact to the high mobility region since only a very narrow funnel of carriers with energy and momentum matching through the scattering can couple. *What is gained the improved transport properties of the lower-dimension medium is lost in the ability to couple.*. The devices may be short gate lengths, but they will be large devices from the large contacts that are needed. The carriers need to funnel in the reciprocal space.

2.5 *Transistors: Subthreshold swing manipulation is of limited utility*

Subthreshold swing manipulation via tunneling through the confined conditions is another theme of the past decade. We discussed quantum means to discriminating between two subband energies earlier. Subthreshold current swing manipulation are of a similar flavor. The proposition is that with no states connecting to the onset of connecting by making bandstates available through electrostatic manipulation may provide sudden current changes not subject to the Boltzmann tails. Tunnel diodes of old are based on conduction in the negative direction and positive direction around zero bias, and then a suppression through the disappearance of tunneling states, and then at higher forward bias the appearance of diffusive current. The proposition is that, examples are shown in Figure 2.13, by placing a tunneling structure at one of the contacts, one may cut off tails. One could do this in a variety of ways since staggered band lineups and other possibilities exist with *III-V* compounds.

Gallium antimonide and Indium arsenide structures, and others, offer different discontinuities in different direction across the interface. The issue is that ensemble fluctuations and homogeneous

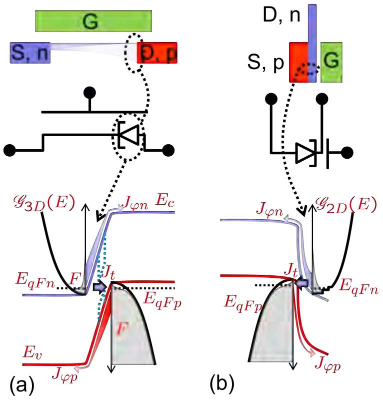


Figure 2.13: A three-dimension to two-dimension region interface, where the mobility is poor in both is shown in (a). In (b) the reduced-dimension region has high mobility and quantized conductance prevails.

broadening itself, as we explored earlier, is of the order of 10 meV. Add to that Gaussian fluctuations of the control of the dimensions of the well. This is similar in nature as the issue of bipolar and unipolar lasers. The quantum cascade lasers have much much larger linewidths because of inhomogeneous broadening compared to bipolar lasers, and this same behavior will appear in subthreshold swings in tunnel diode structures.

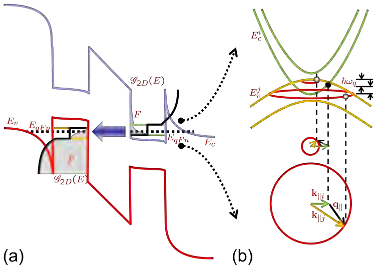


Figure 2.14: Tunneling between two quantum wells to reduce the subthreshold swing.

Take an example structure of coupling quantum wells to quantum wells to reduce linewidths. The tunneling in such structures must conform to

$$\begin{aligned} E_c^i + \frac{\hbar^2\mathbf{k}_{\|i}^2}{2m_c^*} \pm \hbar\omega_q &= E_v^j + \frac{\hbar^2\mathbf{k}_{\|j}^2}{2m_v^*}, \text{ for energy, and} \\ \mathbf{k}_{\|i} \pm \mathbf{q} &= \mathbf{k}_{\|j} \text{ for momentum.} \end{aligned} \tag{2.5}$$

Take the Lorentzian lineshape of an intrinsic quantum structure, and consider Gaussian fluctuation in thickness of Δ. One has

$$\begin{aligned} \langle\Delta(\mathbf{r})\Delta(\mathbf{r}')\rangle &= \Delta^2\exp(-|\mathbf{r}-\mathbf{r}'|^2/\Lambda^2) \\ &\propto F_{mn} \\ &= \sqrt{(\partial E_m/\partial L)(\partial E_n/\partial L)} \\ &\propto L^{-3}. \end{aligned} \tag{2.6}$$

Now the broadening is to the third power inverse of the dimension. This, in a similar vein as that in Figure **??** is of the order of 10–30 meV. It has become the same order of magnitude as the thermal energy and the current has been limited by the number of states coupling.

2.6 *The problems with ferroelectrics*

Ferroelectrics have found multiple cycles of interest since the 1950s. The property of being able to store energy in spontaneous electric polarization and its manipulation with small device compatible voltages with thin films is attractive. After all ferromagnets—spontaneous magnetic polarization—have found large-scale usage in disks and tapes for various hierarchies of storage. But, herein is the issue since early debates on this issue between Landauer and Merz: *Electric polarization is polar while magnetic polarization is axial. This leads to a variety of static and dynamic consequences with spatial intervention. See W. J. Merz, Physical Review,* **95**, *1, 690–698(1954), and R. Landauer, Journal of Applied Physics.,* **28**, *2, 227–234(1957).*

One can illustrate this problem through two examples. First, consider the issue of hysteresis and under what conditions is it observable through Figure 2.15. If one has a load line—defined by the slope

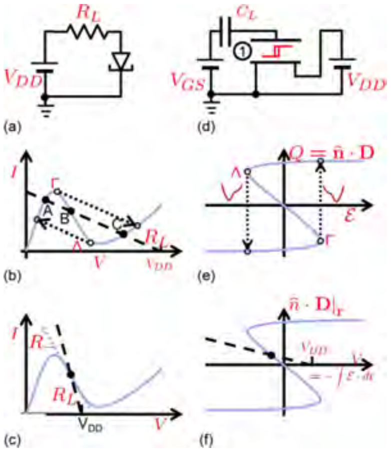


Figure 2.15: Load line effects in observations on tunnel diodes in (a) through (c). (d) through (f) shows the same for ferroelectric structures.

$1/R_L$—as in (b) and sweeps from low voltage to high voltage, one would follow the left (I, V) line till the Γ point at which point one would jump to the right (point C) and then follow the diffusion-dominated current rise. If one was dropping the voltage from up high, one would stay on this right curve till Λ and then jump to the point B. The negative dynamic resistance of the tunnel diode is buried in this hysteresis because the load line resistance is higher than the dynamic resistance. If one made it lower, one would actually see the negative resistance as (c) indicates so long as one has conditions that prevent oscillations such as damping through oscillation-induced dissipation.

Tunnel diodes, by forming the hysteresis window and being fast, were used for a long time in sampling scopes through this careful load line design. One could observe fast transient phenomena using an even faster transient from the tunnel diode.

The equivalent for ferroelectrics in Figure 2.15(d–f) for a transistor is the introduction in a gate. (e) here in the charge–field picture shows the hysteresis jumps arising with the order being largely one way on the left and the opposite way in the other. The charge–field implies an equivalent capacitance load line, and one can imagine seeing the negative dynamic behavior as in (f). The problem with this description is the following. Being polar, spontaneous polarization can be pointed up or pointed down and this can happen in adjacent cells. The spontaneous polarization happens with a subtle very very small fraction of unit-cell size movement to form dipole one way or the other. Movement is crystalline change. In order for the movement to progressively change, the polarization needs to also progressively change. But, a change in crystalline movement couples laterally and propagations happen. In the case of tunnel diodes, there was an external damping mechanism. Ferroelectrics need something equivalent, and that is clamping, that is, preventing size change. But, the transistor gate structure is free. So, the structural effect will lead to ferroelectric domain propagation—an oscillatory phenomena—as shown in Figure 2.16. In this figure, a paraelectric interface layer is also shown since most ferroelectrics employed are not crystalline heterostructures, and the oxygen in them, in time will form the more stable oxide interface layer.

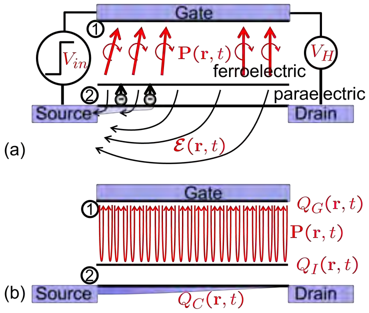


Figure 2.16: The problem of ferroelectric domain propagation, similar to that of tunnel diode current oscillation, unless clamped.

What these illustrate is that negative dynamic parameters, while theoretically may appear as stable points, the dynamics of the change is to make them unstable, and this dynamics is from within the phenomena, in this case the polar nature of the polarization. Any fluctuation will cause a bounce. Changing the gate voltage is a change that will lead to large instability, and the response time of that instability is at sound velocity through the solid's response. This is a *ns* time constant.

2.7 *Thermodynamics and statistics rule*

The lesson of all these examples is that thermodynamics and statistics must be kept front and center. Their consequences for deterministic processes is a price in energy and a limit to the efficiency, no different than that within Carnot cycle.

An interesting illustration of this is shown in Figure 2.17, where very small subthreshold swings can be seen to appear, but there is energy price to it, and there is a speed price to it. This is a double-gate transistor with two separately accessible gates. One gate can therefore be used to change the threshold characteristics of the transistor dynamically. By changing the voltage of the other gate by putting energy into the system and if you change the voltage of the other gate you can change the threshold voltage of the channel. If you can change the threshold voltage of the channel you can move along another path so you can get a sub-60 mV swing. One needed to dissipate energy in the amplifier, the speed will be reduced, and there is more space needed by the entirety of the structure.

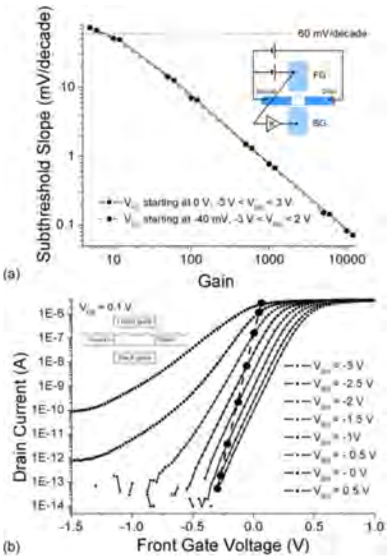


Figure 2.17: A double-gate transistor with the second gate driven voltage amplified from the first gate. As the first gate is turned on the second gate is reducing the threshold voltage of the transistor. The result is a sampling along a curve of lower threshold swing so long as the entire structure can respond with the quasistatic speed.

These are all illustrations of how thermodynamics enters in nearly all the important considerations of structures at the vast scales of integration in the deterministic approaches. A Carnot-like encapsulation of this picture of electronic engine is in Figure 2.18. Consider a circuit of 10^{10} gates, with a fan-in and fan-out of 4, and 1000 terminals. It has a maximum information content of 1.5 terabits. This is its configuration volume. If this chip accesses about 256 Gb of data from an on-chip memory, the maximum convertible negentropy is 5.8×10^{-9} J/K for the chip and 3.5×10^{-12} J/K for the data, that is, a total capability of 6.9 $nW - s$ of information engine capacity. If it performed all this work in a second, it would consume 6.9 nW, if it performed all this work in a ns, that is, about one clock cycle, it would consume 6.9 W. Real microsystems consume nearly orders of magnitude higher power. The Carnot-like efficiency of useful information work compared to energy inputed tends to be a percent of less. This is not unlike, what we found with a single gate, a deterministic inverted needed about 250 k_BT of energy for 1 bit of manipulation, which is informationally $k_BT \ln 2$ away from randomization. So a ratio of $(\ln 2)/250$ of efficiency.

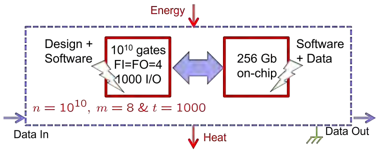


Figure 2.18: The highly integrated semiconductor circuit as an electro-information engine.

The next essay will take a stab at getting far higher efficiencies by relaxing on the deterministic constraints, and by thinking probabilistically, in the next essay using conventional and unconventional forms towards these limits by exploiting the native probabilism of nanoscale and of architectures that can connect this scale to the real world scale by debating the underlying deterministic and non-deterministic approaches and conventional and unconventional device forms.

3
Non-Turing machines: Stochastic and probabilistic learning circuits

The world is probabilistic, whether classical arising in the incompleteness of the classical unknowns or of the natural randomness as in quantum-mechanical fluctuations or their spontaneous classical manifestations. The Turing machine is a computational device that explores the extent and the limit of what can be computed. A simple view would be that it sets limits for implementation of deterministic logic implementation in a computing engine. Boolean, von Neumann, for example. Probabilities, which have within them the objective versus subjective conundrum, not unlike the natural world we inhabit, provide a non-Turing means to computation as one learns. The Bayesian reconstitution of the probability with new information is the subjective tool for this learning. This makes stochastic and probabilistic learning circuits possible, compact and specific, that can operate rapidly and at low power in real time on real-world problems. This essay discusses the underpinning of the computational approach and develops and gives examples of implementation in circuits, where the probabilities are derived using the low-power randomness from superparamagnetism.

We have built an argument that deterministic correctness in computation exacts a large thermodynamic energy penalty. At its best in *CMOS*-based deterministic implementations in an Avogadro-scale environment one would expect under the best of conditions $\sim 300k_BT$ dissipation to overcome fluctuations arising in thermal environment, fluctuations programmed in during fabrication, and changes that ensue during usage. This is excluding the standby energy because off is not really off. This integration number correspond to current state of the art of the densest processors employed for machine learning extended to the cloud. The local processing unit—an artificial intelligence unit (*ALU*), or a tensor processing unit (*TPU*), or other forms that consume much more power since speed is desired. Thermodynamics is quite complete in describing these,

Science and engineering in the world

including their information engine efficiency of less than a percent. Although highly inefficient, the deterministic computing has been an extremely successful paradigm that we will continue with. A practical issue with continuing past practices is that this element—the transistor—is now at nanoscale, abounds with surfaces, and that these very small structures are essentially surfaces with numerous quantum-confinement contributions from what little there is of the bulk has a lot of variability.

In thermomechanics, the Carnot efficiency is of the order of 37 % for conditions of the combustion engines, such as in cars. They come pretty close to it.

The Avogadro scale number of devices and structures put together when one looks at the cloud-computing level makes this enterprise entirely a statistical problem. The last essay argued that the physical phenomena that is being exploited in devices and computation in various ways is all a statistical problem of evolution under the evolutionary law that we prescribe for the computation. Materials could be silicon, could be any compound semiconductor, could be atomic layer thickness, maybe there will be some additional implications in all these various material forms, regardless, the statistics and thermodynamics tells us what we are going to get out of it in terms of an information-centric objective if we think that it must be always correct, that is, that it has to be error-free prediction and inference. Error free by itself puts energy constraints per bit to overcome the statistical and themodynamic matters of knowns and unknowns.

Let us therefore step away from this correctness limitation and explore what one could do if one could actually exploit randomness and probabilities. Maybe there are some interesting low energy things one can do with this twist. This may take us in a non-Turing direction, that is, not being formally complete in the same way that Turing tells us how it should be. As before, we will employ simple toy models to gain an understanding. Toy models are useful tools to clearly see through given the constraints.

Taking the argument of relaxing how accurately one wished to compute, let us first see what may be achievable. Our assumption is that often, being off in the the lowest significant bits is not as big an issue for several computations. For most of nature's species—birds and animal, getting the order of magnitude right is good enough.

Economics, businesses, much human enterprise gets away with such approximations pretty well. As an infant, we start knowing the count of 1 and 2, but more than that is many. As we get our number sense perfected, we may get to the point of single or low double digits. The reason for 10s, 20s, 100s, 1000s et cetera, as being significant markers is that they are a general number that are useful and descriptive. We don't need to be specific to 997 when 1000 is good enough.

An ensemble of $n = 10^{10}$ gates with $m = FI = FO = 4$ as inputs and outputs, and $t = 1000 \;\; I/Os$ can access a configuration space of N^N, where $N = nm + t$. Memories being a very specific intersection organization of the binary possibilities, on the other hand, can access M^2 configuration space . This ends up being a 7.3 $nW \cdot s$ work capability. Thermodynamics tells one from this that waiting for a ms for some big computation to finish should be doable with just 10^3 times 7.3 nW power. Reality is worse since current technology is far off the limits of few hundred k_BT, being about $10000k_BT$

per bit and the computation is sequential and not en-mass. One can see in this estimating all those ways that the thermodynamic notions came in. If one could relax the ability to the least significant bit, that ism let them be more error prone, one could change the thermodynamic activation slope of the curve for any limits one places on the error. An example of adder, where one allows the least-significant bit (*LSB*) to relax is shown in Figure 3.1. The obvious thing to note that there is an exponential dependence arising entirely from the simulation. Thermodynamic expectation of form matches a quantitative estimation from device up. One can live with some of these errors for certain applications. A daily one is the the usage in compression techniques, that is, soft techniques, for example in the compression one does is to reduce the size of images.

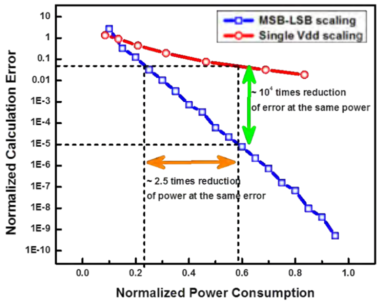


Figure 3.1: Activation energy of model adders where scaling of voltage is employed to relax least-significant bit errors.

The problem with all this estimation by relaxing error requirements, while still computing deterministically, is that one could reduce power by a factor of 2 for a factor of 10^4 in the errors. This is important to note. The toy model is a beautiful ways of understanding and not have to do simulations all the time because simulations always go into a very very specific niche.

One can expand on this to emphasize the power of toy models.

The opposite point also needs stressing, that is, the power of tackling complex models with machine learning. This is what Kac called a deep truth, which is a statement, and its opposite, both being true. This is duality, which shows up in so many places. Even in simple philosophical matters. To know that one is happy or satisfied, one needs to have experienced unhappiness and dissatisfaction. A quantum superposition in a two-level system can be one or the other, from which the power of entanglement as non-product states appears. Deep truths are the counterfactual of opposites.

3.1 *Toy model of deterministic approximate calculations*

Consider a floating point inexact addition, $A + B$,

$$A(\equiv a_{n-1}...a_1a_0) + B(\equiv b_{n-1}...b_1b_0) = S(\equiv s_{n-1}...s_1s_0). \qquad (3.1)$$

Take a sum S' within a probability distribution $\mathfrak{p}_\sigma(S' - S)$ centered at S of width σ, all expressed in normalized form, for example, in units of least precision (*ulp*s). If $\sigma \gg 1$, bits less significant than σ can be ignored and the circuit truncated, that is, those parts shut off, to make the $LSB \sim \sigma$. This is reducing power, where one has employed $\sigma \approx 1$ as a useful marker for error. Now consider distribution of sums, $\exp\left[-\pi(S - S^{'})^2\right]$ for Gauss-normal, or $2/\pi[1 + 4(S - S^{'})^2]^{-1}$ for Cauchy-Lorentz like distribution. For $|S - S'| \gg 1$, that is, in tails, the discreteness can be ignored and what is important is the exponential or polynomial tail. Now take arithmetic circuits composed of elements $\alpha = 1, 2, \ldots, M$. Adders come in many forms, but carry-select, a common one using half adders generating partial sums and Kill/propagate/generate (KPG) signals is a popular one that combines blocks to form carry propagation tree. Multiplexers at the output select the partial sums based on carry-propagation trees.

In this arrangement, for every element α, let U_α be the energy dissipated per computation using design/voltage scaling, ε_α the

the probability of error, and w_α the weight for each element that quantifies mean magnitude of numerical error in the answer caused by incorrectness in the element while all the others are correct. In a multiplexer corresponding to the bit k of output, the weight is 2^k $ulps$. This is the weight contribution to the error in a complex assembly.

Now assume that the errors of each element are uncorrelated. When an element, $\alpha \in W \subset 1, 2, ..., M$, is incorrect, it causes an error of $\langle (S - S')^2 \rangle_W = \sum_{\alpha \in W} w_\alpha^2$. Accumulating W as the set of erroneous elements with probability $[\prod_{\alpha \in W} \epsilon_\alpha][\prod_{\beta \notin W}(1 - \epsilon_\beta)]$, the error in the final result would be within the constraints of the probability distribution $\mathfrak{p}_\sigma(S' - S)$ so long as

$$\left[\prod_{\alpha \in W} \epsilon_\alpha\right]\left[\prod_{\beta \notin W}(1 - \epsilon_\beta)\right] \leq \frac{1}{g_W}\mathfrak{p}_\sigma\left(\sum_{\alpha \in W} w_\alpha^2\right) \forall W \subset 1, 2, ..., M, \quad (3.2)$$

where g_W is the degeneracy, that is, different subsets W that lead to the same mean-squared error $\sum_{\alpha \in W} w_\alpha^2$.

In a more rigorous analysis the right hand side will be eplaced with a sum of eigenvalues of the covariance matrix.

For Gaussian distribution, the constraint on error is satisfied if

$$\epsilon_\alpha \leq \frac{1}{g_W} \exp\left[-\pi w_\alpha^2\right] \quad \forall \alpha. \quad (3.3)$$

For the Cauchy-Lorentz distribution, the equivalent requirement is

$$\epsilon_\alpha \leq \frac{2/\pi}{g_{w_\alpha}(1 + 4w_\alpha^2)} \quad \forall \alpha. \quad (3.4)$$

It is gratifying to see that this toy model follows to the same results as that of the last essay. We end up in this formulation, following Gaussian distribution, of $U \propto -\ln \varepsilon$, same as what was estimated on thermodynamic grounds, and a proportionality constant that depends on the source function of errors. For threshold voltage variations, the energy is $\sim CV_{DD}\sigma_{V_T}$. For thermal noise as the source, it is $\sim k_B T$.

This methodology lets us look at cases of interest. One is the carry tree linear chain: an inefficient ripple-carry adder in carry-select. Take an adder, a KPG unit, 1 multiplexer per bit, except for the extrema, with elements at bit level k each weighted as 2^k. We have a degeneracy $g_{w_\alpha} = 3$. The energy dissipated per computation for an entire n-bit adder under Gaussian distribution errors is

$$U \propto -4\sum_{k=0}^{n-1} \ln\left(\frac{1}{3}\exp(-2^{2k}\pi)\right) = 4n\ln 3 + \frac{4\pi}{3}\left(2^{2n} - 1\right), \quad (3.5)$$

and the Cauchy-Lorentz distribution errors of

$$U \propto -4\sum_{k=0}^{n-1} \ln\left\{\frac{2/\pi}{3[1 + 2^2(k+1)]}\right\} \approx 4n\ln\left(\frac{3\pi}{2}\right) + 4n(n+1)\ln 2. \quad (3.6)$$

$U \propto 4n \ln\left(\frac{1}{\epsilon}\right)$ for the flat errors in exact arithmetic. Between this set of equations, one can now tell where the approximate adders can be more efficient given the Gaussian or Cauchy-Lorentz distribution. The lesson of this calculation is the same as one found through adder simulations and a set of results are shown in Figure 3.2, one for adder and one for multiplier. Factor of 2 is about the limit to the savings, whether one considers adders or multipliers. *There are only a limited number of tasks where traditional deterministic computation calculations relaxed for constrained errors can be gainfully used. And that too will only be factors of* 2. This is simple analytic analysis leading to the same result as a more complex model-based simulation.

So we must look elsewhere.

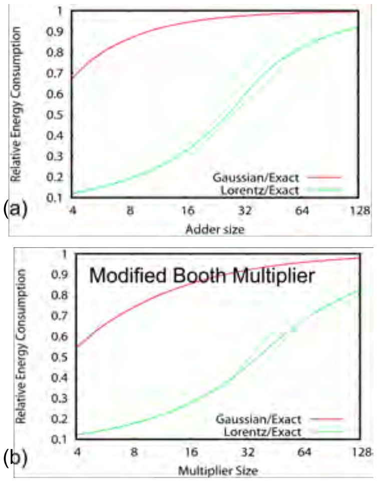


Figure 3.2: Relative energy consumption in adders and Booth multipliers under various distribution functions of error. Factor of 2 is about the saving the energy in best of cases.

3.2 *Using randomness and Turing approach*

The history of information mechanics engine is an interplay between algorithms and physical platform. Hardware and software together, software drives hardware, hardware drives software, and one finds symbiotic approaches depending on what is constraining and needs to be addressed. This is linear perturbation and a linear approach. For example, today everybody is doing stream processing with graphical processing units because that way computation is mapped to streaming, thus not waiting for information to arrive to continue a computation or passing a lot of information around on a chip. This is efficient, both in timing by disposing of waiting requirements and energy since there is not much shuffling of movement of data. So, Compute Unified Device Architecture (*CUDA*), which got its start a couple of decades ago went mainstream, and today the fastest processing is on graphical processing units or tensor processing units or artificial intelligence units incorporating a lot of streaming.

However, we must still recognize that the non-determinism is tied to complexity. *CUDA* and all the graphics applications are still deterministic computation of probabilities performed on deterministic systems and still subject to all the statistical arguments made to this point. Neural networks, for example, use randomness and approximations obtained through affine transformations followed by nonlinear thresholding. Underneath the various approximation approaches is the probabilities. One is in the end making a probabilistic match to whatever pattern one is trying to figure out through the weights and the divisions in whatever way one parses, such as the sequential transformations of Basic Linear Algebra Subprograms (*BLAS*) across the hidden layers of neural networks. So all the energy

Machine learning, with tensor processing of back and forward propagation in neural network is propagative, and particularly suitable for the streaming hardware. Artificial intelligence units are a higher order perturbation on that. What is fascinating is to wonder whether machine learning's rise is due to progress in understanding the perceptrons and deep networks, or is it that the computational platform turned out to be the perfect medium for playing and thus caused the development of all the different neural networks.

considerations still exist since determinism is how the computation is locally. This holds for NVidia chips with probability calculations performed deterministically. It is therefore no surprise that these are thousand Watt chips because of the needs of deterministic computation. In the next essay, we will explore neural approaches. They are extremely powerful, useful, and open up new frontiers in complexity despite this shortcoming of high energy dissipation, but first, we explore using randomness at a simpler level. This lets us understand the fundamental ways of using randomness, and to approach learning, therefore past as in priors to future and inferencing through Bayesian methods. This is also applicable to neural networks.

This essay explores this randomness—as a a foundation for computation—subject through more grounded simple methods that are useful in edge-of-the-network usage, where being mostly correct and using much lower energy is useful. This direction grew as a diversion within the last decade when the statistical limits intersecting with determinism became too clear to ignore.

Mathematics has much to teach us wth statistical mechanics seen as an offshoot and probabilities as a way of seeing what is not known as entropy, whether it is in the old Claussius sense, or in the modern information sense. Bayesian approach is the evolutionary law that one has to follow through on state description even if it is states that we don't know. Bayes'

This next segue emphasizes the dynamic evolution of learning in the midst of unknowns and in the next chapter exploiting randomness using neural networks to tackle the complexity of the physical world. This essay emphasized Bayesian methods as low energy and intrinsically probabilistic methods that can also be handled non deterministically for computational information processing tasks.

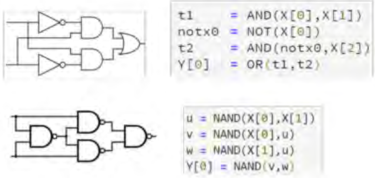
Turing machines are hypothetical abstract devices that yield finite descriptions of algorithms that can handle arbitrarily long inputs. Turing placed in machine, therefore an automaton-like form, Church's thesis. It instructs us how an algorithm can evaluate functions of every input length. Take the Boolean set, and some variable x that is being mapped to a function, a simple deterministic calculation as shown in Table 3.1, in a bounded form. We know this is a complete specification and one can map this to a circuit form, two examples of which are shown in Figure 3.3 for some specific $f(x)$. This is straightforward.

Now consider an unbounded problem for $x \rightarrow F(x)$ such as of Table 3.2. It may be that one can tackle this in an automaton, or maybe we cannot. Turing taught us how to think about this objectively (see Figure 3.4) through a tape and a finite state machine that decides the moves of the read/write head that operates on the tape. This is a

While we give exclusive credit to Bayes for formulating the prior to posterior description, a significant credit to Laplace is also warranted. Laplace had his rule of succession—the Bayes analog—and applied it to calculate the probability that the sun will rise tomorrow, given that it has risen every day for the past 5000 years.

$\{0,1\}^n$	x	$f(x)$
	0000	1
	0001	0
	...	...
	1111	1

Table 3.1: A mapping of a Boolean variable x to a function $f(x)$ of the Boolean variable.



```
t1     = AND(X[0],X[1])
notx0  = NOT(X[0])
t2     = AND(notx0,X[2])
Y[0]   = OR(t1,t2)
```

```
u = NAND(X[0],X[1])
v = NAND(X[0],u)
w = NAND(X[1],u)
Y[0] = NAND(v,w)
```

Figure 3.3: Two example circuits of a bounded problem of finite specification describable through the corresponding programs.

$\{0,1\}^*$	x	$F(x)$
	0	1
	1	0
	00	0
	01	1
	10	1
	...	...

Table 3.2: An unbounded mapping of a Boolean variable x to a function $F(x)$ of the Boolean variable for contrasting to the bounded problem of Table 3.1

form of λ calculus driven by a program for calculating the moves of the machine. A solvable problem halts. It leads to finite descriptions of algorithms that can handle arbitrarily long test so one may design, for example, a finite state machine with a certain number of gates, et cetera, in order to do some computation. Turing machine is a way of telling how that finite state machine can take on a bigger and bigger problem and remain finite and tell one how to solve that particular problem. The Turing method has added looping and a way to reduce the λ calculus to a machine form. A simple example is that to calculate $n!$, let $(m = 0)! = 1$ and then for $(n = m + 1)! = (m + 1) \times m!$. One now has a recursive relationship in the λ form. One can see in this formulation the power of the Turing approach in reducing solvable questions to a machine form, even if there is no way of being able to tell in general if the machine will halt with the problem solved.

We view this Turing device as hypothetical abstract construct just like the use of statistical mechanics as a hypothetical abstract device. It is a generalized recursive program that allows one to perform λ calculus.

What has this got to do with randomness, probabilities, and information processing in non-deterministic way in the presence of unknowns? Randomness means different things to different people. What just happens by chance, that is, was unpredictable. Sometimes because we don't have the time or the tools or the resources to remove that unpredictability. Sometimes, as with quantum uncertainty, it is intrinsic not knowing specified by nature.

From an information perspective entropy is what is not known. Randomness is sometimes equated with or connected to entropy. *This is a fallacy.* There are plenty of random things that have low entropy, there are plenty of ordered objects that have large entropy. A variety of interesting metal-insulator transitions owe their origin to ordering or disordering of spin, with some of the order actually arising with increase in temperature.

The problem of information and entropy and what is not known are all connected to each other. If one peeks into a fire kiln that is hot through a small pinhole, all one will see is largely red hot, and maybe one may make out a shape, or maybe not. But shine a light from outside source, so a different distribution of photons, and one can see more clearly what is in the kiln. When one is near thermodynamic equilibrium, a high entropy condition of lot more of unknown possibilities, an external energy resource can reveal information.

The hidden an be revealed using proper tools for observation, that is, information gathering.

Two systems out-of-equilibrium with each other can enlighten.

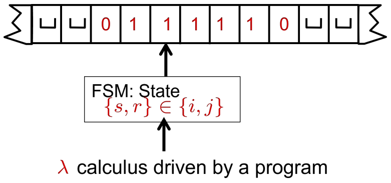


Figure 3.4: A Turing machine consisting of a tape with information on state, a finite-state machine that determines the action (an evolutionary law) and a program, so a finite memory that tells the finite-state machine the action driven by λ calculus.

The name Turing machine has panache. It is name dropped. I too am guilty of it by calling this essay Non-Turing machines. They are non-Turing only in the sense that the original Turing construct is working with the Boolean representation to perform a deterministic evolution. A state transforming to another state using an instruction from a state machine, which may or may not have memory, is a Turing machine. It is just a mapping of state evolution. There is nothing that says that one could not do this with probabilistic and therefore a fuzzy tape and a fuzzy state machine. Bayesian rule describes how it should proceed. The *AI/ML* community uses non-Turing and Turing with abandon, with the original construction now quite lost. I am reminded of Kurt Vonnegut's edict on using semicolons. ``Do not use semicolons. They are transvestite hermaphrodites representing absolutely nothing. All they do is show you've been to college.˝

The sun, a pretty decent blackbody object at 6800 K, helps us see things on earth. If one wants to see what molecules are in the air, just look at the specific lines of absorption and radiation of these molecules in the incident radiation on earth. This despite us not really having to know necessarily what makes the sun radiate. It may be fusion, or it may be composed of gold!

Both cause and chance have entered in this description as did the classical and the quantum. Causal evolution in quantum is an evolutionary law, no entropy involved, that describes how states change. The chance of one of many possibilities comes from one of the many possibilities that the quantum state, whose state function encapsulated the possibilities, is found in. This is now a reduction to classical information. In the classical description, it is either the large number of possibilities that one needs to keep track of, or the inability to actually figure it out through the classical description, that is chance. Cause is a relationship. Chance is the unknown or the random intermezzo between the unknowns. Also, when we compress, that is, approximate, chance will appear. Entropy is therefore appearing in multitude of ways in indeterminism. It connects to the probabilism of inference. *Any change in entropy is either acquisition or loss of information, therefore involves dissipation of energy.*

One can shine light on the what is not known, represented in entropy, and learn and do things with the information. Energy was involved, probabilities changed, and our information changed. In all this, what we are doing is figuring out what is caused and what is chance and how did they interact. So, cause and chance and randomness are connecting to each other.

How can one place some objective weight on a picture surfeit with this subjectiveness?

Kolmogorov, the grand master of probability, stated it best for information in *a sequence of bits is random if the shortest computer program for generating the sequence is at least as long as the sequence itself.* This represents the Kolmogorov notion of complexity too. It is a self referential statement. Often the best descriptions, because of inherent complexity, of the most difficult questions are self referential. Kolmogorov's is a self-referential definition of randomness. Pseudo random numbers immediately fall off from randomness ladder. They may take long to figure out, but they can be figured out, and new tools may make that process much faster. They are just stretched out by elliptic functions. Only quantum uncertainty—Heisenberg's great insight—is truly random, and whatever else at the quantum scale that appears in the real world in an objective way from this randomness. We will look at this shortly using superparamagnetism.

But, pseudorandomness can still be useful. It has been deployed in

What is life is one of these difficult questions. Ask a biologist and you will get many many different answers all trying hard to differentiate molecules, cells, viruses, trees, animals, and so on depending on where they wish to draw a line, a bit like distinguishing between the non-vegetarianism of eating dogs, cats, pigs, cows, plants, bacteria, animal products, and so on, or it may be a more pragmatic and nuanced question and answer session involving many questions, which has the entropy-like yes/no question view towards information content. But that is categorization and asserting an objective measure of what is not known. To me the self-referential, ``Life is what life does.'' is sufficient to describe the complexity. It is information and what is not known both merged in it.

algorithms and in communications and in cryptography and safety as a resource since the dawn of computing. Even the *TCP/IP* communications employs it to reduce message collisions.

3.3 *Randomness, compressive sensing and neuron spiking*

An example of the use of randomness is what one accomplishes through compressive sensing. Much of what we work with does not appear ordered.

Tree leafs, signals that appear and disappear in short times, or small spaces, et cetera are all sparse vectors. If one samples vectors directly without knowing what is active and what is not, we will sample nothing most of the time. As an electrical engineering student in an early class, the answer would be to deploy Nyquist theorem and use twice the amount of bandwidth. This will capture rise and fall. But that answer is silly in locating the object efficiently. Use of pseudorandomness, that is incoherence, with pseudorandom bit stream, has a higher likelihood of picking up a little bit of information. This is used extensively in compressing representation of large data, such as *MRI* images, et cetera. Compression is more efficient. Take multiple pseudorandom sampling signals and one can use them to project the data as $\mathbf{A}\mathbf{x}$ where $\mathbf{x}$ is the basis set. This picks up the information because the signal itself is not ordered. Using unordered sampling actually is helpful. There are definitive rules—mathematics tells us those—and we end up with the $l1$ norm, instead of $l0$ or $l2$ for extracting information.

This idea of compressive sensing or compressive information coding, dependent as it is on pseudorandom patterns, is interestingly exhibited in the neural information system. The action potentials of spikes are based on electrochemical potential changes arising in leaky ion channels. The spike trains, by and large look random, and the signal is of about 70 mV magnitude. The information is embedded in spike trains, the rates are connected to information, and one can see how energy and information is near-ideally handled through statistical mechanics applied to the physical world.

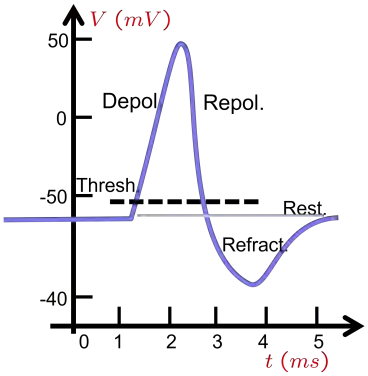


Figure 3.5: The action potential of a neuron spike.

These spikes have thresholds. When the signal rises above a threshold from summations of signals coming in, the neuron fires. And then it decays out, following which it gets back to a rest state. There is much of physical interest in this despite the spike carrying considerable energy in its *ms*s time constants, and despite the *ms*s, the information processing once everything—much of which we don't understand—is said and done. Ultimately it results in a net low energy in the inferencing task.

The brain is 10–20 W engine. This is low by silicon circuit standards, but high by human standards. The body utilizes similar power for mechanical work. The brain needs a lot of blood flow in its vicinity, chemicals to be moved in, and all the oxygen for electrochemical processes.

The membrane signaling process is a capacitance based and signaling through the capacitance-conductance pathway of channels manifested in the voltage spike.

The action potential in spiking is viewable through the constraints of thermodynamics and living biology. The action potential moves down an axon whose simplest model is that of a capacitive membrane across which an electromotive potential exists due to the ion concentration differences causing the conductance channels to open and close. This signaling is dissipative. But, absence of signaling too is dissipative. So, the potentials, currents and times need to be consistent within the energy constraints. The spiking and the noise together make this signaling mechanism effective. The equilibrium potential is calculable from the ionic concentration across the membranes, with K^+ dominating, but Na^+, Ca^+ and Cl^- also present. K^+ concentration is higher outside the tubule, while in case of Na^+ it is higher inside. The biologically sustainable concentrations are in the order of few mM to few $100s\ mM$. For example, for K^+, Na^+ and Cl^- inside/outside these are 5/140, 140/12 and 20.

The diffusive and electrical flow balancing establishes the reversal potential, which is the voltage across the specific ion channel during its operation. This reversal potential, for K^+ flow in its channel, is

$$V_{Rev} = \frac{RT}{zF}\ln\frac{[K^+]_{out}}{[K^+]_{in}} = \frac{8.314\times 310}{96845}\ln\frac{5}{140} = -88.7\ mV. \tag{3.7}$$

Here, R is the gas constant (8.314 $J/K.mole$), T is the body temperature (310 K), z is the ionicity (1 for K^+), F is the Faraday constant (96485 $J/V.mole$) and concentrations of ions is a ratio in identical units. Cl^-, which is not actively pumped, settles at a reversal potential close to the resting potential determined by other ions. Chlorine is also highly impermeable. This resting potential, absent any activity, is a balance of concentrations and the permeabilities of their channels, which following the Nernst equation approach is

$$V_{rest} = \frac{RT}{F}\ln\frac{\sum_i \pi_i[A^+]_{out} + \sum_j \pi_j[B^-]_{in}}{\sum_i \pi_i[A^+]_{in} + \sum_j \pi_j[B^-]_{out}}. \tag{3.8}$$

The size of the ion ($Na^+ < K^+(0.138\ nm) < Ca^+$), for example, and the size of that ion's pore matter for this resting potential. $\pi_{Na}/\pi_K < 0.01$. The resting potential is maintained by active ion pumping to compensate for leakages. The pumps—marvels of near-ideal electrochemomechanical coupling—and the permeability lead to smaller resting potential, which for these parameters, including leakage of K^+, Na^+ and Cl^-, are $1\times 10^{-6}\ cm/s$, $2\times 10^{-8}\ cm/s$ and $5\times 10^{-10}\ cm/s$. This resting potential calculates to $-78\ meV$. This is in the range that is measured across species and cell types. The

B. Hille, "Potassium channels in myelinated nerve. Selective permeability to small cations," J. Gen. Physiology, **61**, 669–686(1973)

P. Ronald and J. MacGregor, *Theoretical mechanics of biological neural networks,* ISBN 978-0-12-464255-3, Elsevier (1993)

$+40\ mV$ peak—of the order of k_BT/e—in action potential is significant in its role in how noise, spiking and information processing interact. This spike contains $\approx\ 0.5 \times 110\ \ mV \times 10^{-3}\ \ s \times 3\ \ pA\ \approx$ $165\ aJ \approx 40000 k_BT$ of energy.

The simple model then is that there is a membrane, there are chemicals both inside the axon and outside the axon, potassium is the one that is really the one that moves, the large sodium doesn't do that much, but these ions are leaking and these opening and closing of the pores is what causes this electrochemical potential to suddenly change so there's a reversal potential and that reversal potential is related to how the potassium moves.

What is most interesting is the following. One can make a small model of this as a FitzHugh-Naguno (FN) neuron for potentiation and flipping.

$$\begin{aligned} \tau_{\uparrow} d_t u &= u(u-0.5)(1-u) - v \\ \tau_{\downarrow} d_t v &= u - v - \beta + \varepsilon \sin(\omega t) + \zeta_n, \end{aligned} \tag{3.9}$$

where $u(t)$ is an action potential, $v(t)$, a refractory variable, β a bias, and ζ_n noise ($\sim\ 30\ \ mV$). One can apply the Fisher information statistics to it. Fisher information is about the connections within the data and the continuity of the data. Applying these statistical methods to the FN neuron says that about 40 mV is the peak of the signal, and the spike has an energy of the order of $40,000 k_BT$. This is more than 20 times what exists in $CMOS$ logic. Even then, then it is a low energy process for inference. It is so because of how it works is a mutual information driven process. Maximum information content is preserved during transformations. Information processing itself then is efficient.

This ability to achieve efficiency is throughout us. It is using of the right speed and techniques in the manipulation to achieve an objective. We convert nearly 20 lbs of ATP and ADP into each other during day to manage the proton motion for all the mechanical energy we expend similar to the brain's computation energy. These ones are low energy 100 k_BT steps, slow and steady. Ribosome and the $mRNA$ translation into proteins, a transduction that is an essential step to living, happens largely without errors, although our last chapter's discussion would say otherwise, since there is also a ratchet mechanism, a mechanism that checks the correctness, and if it is wrong, it steps back and corrects The slow timing, both in neural and the proton engines, are state-to-state change with very little irreversibly loss involved in the process. It is a little more elaborate Turing machine correcting errors.

With this discussion of randomness, sparsity, the time and the information coding and the energy transformation in the mix together,

Fisher information is a statistic of parameterization of a distribution to tell us how one may capture the most information in the distribution in the parameterization. This information content that is capturable and errors are then relatable, and that is what the Crámer-Rao bound is.

There is an equivalent of this same idea useful in neural networks. Neural networks work best by handling information bottleneck of this mutual information preservation best. They need appropriate hidden layers and numbers for that process to take place. If one wants to preserve the most information in the statistical parameters of a large data, then maximizing Fisher information through neural networks is another of one such common techniques.

There are a few places in computing, specially the specialized high performance computing, where this method of using checkpoints, and stepping back and doing again is employed. The brute force way, such as in the Apollo machines, which had to handle all the high energy particle induced errors in outer space, was to have three computers compute, and then vote. Democracy in computing!

it is useful to see the power of these tools that we have in nature and in our own creations that achieve efficiency and low energy and appropriateness in technology.

3.4 *Nature's use of fluctuations and synchronization*

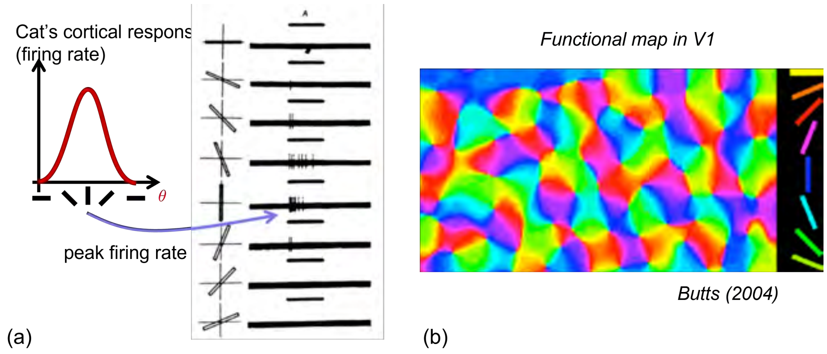
Fluctuations are the appearance of a perturbation randomly. The eye uses the flickering of the retina—the movement of the photoreceptors (rods and cones)—that accomplish the first-order detecting of edges and in turn allowing the axons to feed the information to the $V1$ area of the visual cortex. This flickering is basically the use of a kernel in a convolution for edge detection. It can also be seen as corresponding to the Green's function techniques of response theory.

The perturbation that we have mostly used in the physical experimental work is dithering—a small-signal ordered fluctuation—as in lock-in techniques to improve sensitivity and reducing noise in a bandwidth for repeating signals. Analog communications uses heterodyning for improving sensitivity. All of these are low-power and low-energy detection techniques that work by reducing the state space in which one is probing by working in a narrow band. They use ordered signals.

We also use the random perturbation. Compressed sensing is the use of linear projections onto random basis just like small-signal perturbation is a linear projection on an ordered basis. That the approach is using randomness lets it look at sparse unordered signals of edges and changes and lets one reconstruct via nonlinear processing.

We will see this interesting property in the neural networks that use randomness for achieving robustness and accuracy and avoiding overfitting.

In all this, the random change–flicker—is different—noise like—from the wave like continuous method. But, flicker can also be subject to synchronization in a window because of nonlinearities



Functional map in V1

Butts (2004)

Figure 3.6: (a) shows Hubel and Wiesel's results of a cat's cortical response to edge in the form of a tuning curve. Horizontal edges had lowest rate and vertical edge had the highest rate. (b) shows the functional map of $V1$ (Butts (2004)) from the differently positioned edges being shown to the eye.

The most powerful example of randomness' use in nature is our eye. The eye uses a tuning curve (a window) to capture the visual information. An experiment by Hubel and Wiesel dating back to 1959 showed the spike train coding, that is, a tuning power spectrum curve response from a cat as seen in Figure 3.6. The tuning curve is a power spectrum that captures the visual information and passes it on in the form of a spike train to the visual cortex, where it is further refined. The tuning curve actually peak when it is oriented vertical and goes down on either side with an angular dependence. Butts—nearly fifty years later— in 2004 showed this same response in the functional map of $V1$ as seen in the (b) panel of the figure.

One of the interesting part of this firing on edges is that it aids in the energy reduction. If one is looking at a blue sky, it has little information content. One need not spend energy for each and every elemental volume of the sky one is looking at. A model of just assigning blue to that large volume suffices. But, if there is bird flying, the eye fires away looking at that edge change, the energy in the eye and the brain is being used mostly for this information content, and one can follow the motion of the bird because the brain computational machinery has the capacity to do that with the efficient coding of firing. This even works at an even higher and deeper level. When one watches somebody walking away in the distance, one may recognize and assign an identity by the bearing and gait, and past memory of this somewhere in the brain through some domain integration. Yet, after a moment's passing, we may conclude that it is not that person because some other information was pulled out of the brain's archive. Two different time scales were pulled together for the inference. All this inferencing has to depend on energy-efficient techniques matching to the fast and slow needs of the circumstances.

What do flicker and synchronization have in common that makes them work? Flicker works with nonlinearity. If one has a threshold, so a nonlinearity threshold, then signals of one side are accentuated and the other suppressed. The threshold amplifies a difference that exists at these edges. A random sequence of sensing pulses are probing these edges, and are efficient since the edges or the signal are not a repeating functional pattern. Signal may be sparse, but can now be found through the chance and judicious use of randomness. We will see an example of this usefulness. It can be used to unveil information that is buried in, and in Chapter 4 it is being practiced for similar purposes in autoencoders. Synchronization can be seen as a way of building energy. Synchronization and flicker put together can become an even more powerful information extraction technique for useful circumstances.

The famous examples of bridge collapses or instabilities with people walking on them is an interaction between a swaying bridge and people's reaction to it. A large number of people inputting energy in synchrony with the swing causes more and more energy to be coupled to the swinging, and voilà, catastrophe.

This is also convolution, the technique for measuring how self similar two signals are and where they differ. Displacement and detecting change is accomplishing the same thing that the eye's flickering is. Except that the eye is using a random basis.

The response curve's information view can be seen through Fisher information. Fisher information is a measure of information content in the parameterization of data. Unlike Shannon's information measure that integrates the surprisals in a data stream, the Fisher information metric measures the relationships as

Surprisal in Shannon way of thinking is that it is the unexpectedness that relates to information. If one knows, there is nothing surprising, and the information content vanished. When first encountering this word, an old *Mahabharata* tail immediately came to me as I nodded internally. Yudhishthir's four siblings had been struck dead because they refused to answer them before drinking water from the lake guarded by a *yaksha*. Yudhisthir too faced the questions to bring them back to life. One question was ``*Kim ashcharyam?*''(What is surprising?). The reply, ``Death is inevitable but we take life for granted and live as if we live for ever. That is *ashcharyam.*'' This is Zeno's paradox, arrow of time, and information rolled into one and the same. Yudhisthir did not posit the natural logarithm of probability as the surprisal, but that is only an indication of how much surprisal there is in what we don't know.

$$\begin{aligned} I(\mathfrak{p}) &= \int [\partial_\theta \mathfrak{p}(x_i|\theta)]^2 \mathfrak{p}(x_i|\theta) dx_i \\ &\equiv \int [\partial_x \mathfrak{p}(x)]^2 \mathfrak{p}(x) dx, \end{aligned} \tag{3.10}$$

where one may view the data set $\{x\}$ with individual elements as $x_i = \theta + \varepsilon_i$. The Cramér-Rao bound relationship

$$\langle \varepsilon^2 \rangle \geq \frac{1}{I(\mathfrak{p})} \tag{3.11}$$

tells us the information content's relationship to the error bounds of the estimation. The best estimate of the parameter θ then has a mean-square error of $1/I$. Fisher information projects smoothness. A normal probability distribution $\mathfrak{p}(x)$ has a variance of σ^2 and Fisher information of $I = 1/\sigma^2$. If I is small, error is large, so smoothest $\mathfrak{p}(x)$ consistent with additional information is the more likely fit.

The tuning curve seen through the Fisher information view (Figure 3.7) is.

$$I(\theta) = \left\langle \left[\frac{\partial \ln \mathfrak{p}(r|\theta)}{\partial \theta} \right]^2 \right\rangle_r = \frac{1}{\sigma^2} [f'(\theta)]^2 \tag{3.12}$$

The Fisher information vanishes at peak and at no firing rate. Large and low firing rates have low information. It is the angular positions that needs the most acuity for detection, and they are the the ones where Fisher information peaks. This turns out to be a highly efficient coding method. It is speculated that this is the way the brain's $V1 - V2 - V3 - V4$ system works.

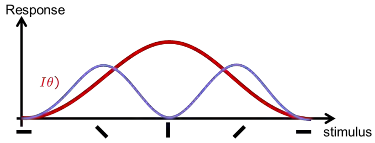


Figure 3.7: The tuning response curve and its Fisher information.

3.5 *From nature to physical*

THE TUNING CURVE AND ITS EDGE RESPONSE is a reminder of the magnetic memory errors and probability discussion of Chapter 2. It is the magnetization vector flipping from one direction to the other, when it should remain in the prior direction, is what we see as an error. If one was at the peak of the probability curve, no error, but as one reached the tail edge, errors appear. Errors are surprisal. They

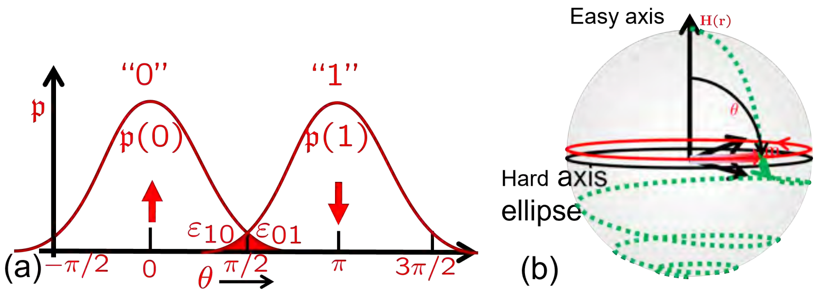


Figure 3.8: The magnetic domain as a prototype for flicker behavior in superparamagnetic limit as the magnetization flips across hard axis. (a) shows the probability curves for the binary states. (b) shows how in superparamagnetism, one is in the tails of probabilities causing precession to flip between up and down ever so often.

are meaningful information and not unlike what the eye tuning curve is also speaking to. This flipping happens across the hard axis, so it shows up as a shot-like, that is, a flicker-like like behavior.

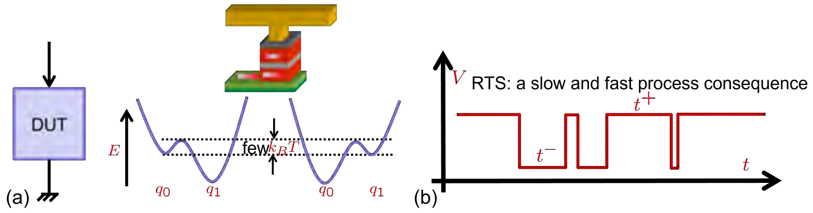


Figure 3.9: The use of superparamagnetic limit in spin-torque structure to create flicker behavior by current drive. In (a), this is shown as a flipping in the energy landscape. the consequence is is a junction a response that reflects the resistance difference between aligned and anti-aligned magnetism.

This can be reduced to practice using spin-torque current-driven structures where the free layer is made superparamagnetic as shown in Figure 3.9. Current drive can be used to modulate the energy landscape so the probabilities of transitions can be manipulated. One polarization—the aligned one—has low resistance and the opposite alignment increases resistance. So one sees a flipping induced shot-noise like behavior in the response proportional to the resistance. A random telegraph signal appears arising in the Poisson low probability as

$$\mathfrak{p}(k,\nu T) = \frac{(\nu T)^k}{k!}\exp(-\nu T). \tag{3.13}$$

This corresponds to dwell times in the two possible states (high and low) with probability

$$\mathfrak{p}(t^\pm) = \frac{1}{\tau^\pm}\exp\left(-\frac{t^\pm}{\tau^\pm}\right), \tag{3.14}$$

where the τs are time constants that can be manipulated by current.

This response is flicker. The spike rates are related to state transition and an important analog to the discussion of eye's flicker as a Fisher information efficient algorithm for detection. We have to make this barrier smaller and smaller of the order of few k_BT so that one can induce low-energy flipping back and forth. One way is to make the structure small and isotropic and then deploy the bias current. Super paramagnetism is achievable in about 10 nm sizes. It is compact. This is all very low energy, quite unlike the implementation employed in the thermodynamic tail of deterministic computing. This is potentially a way to achieve tuning curves that can be modulated.

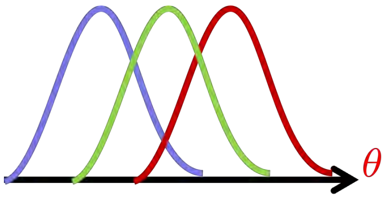


Figure 3.10: Multiple tuning curves through multiple junctions biased at different currents.

Let us see how to exploit this randomness using something similar to the eye. We can use a number of such paramagnetic junctions (Figure 3.10), whose weights are modulated to sum and achieve a signal $H(\theta) = \sum_{i=1}^{N} w_i r_i(\theta)$, where the population coding is achieved by tuning r_i. The r_i are the probabilistically-coded composing states of the net response that can be seen as a spectrum of many tuning curves as seen in Figure 3.10, each one the curves being of the form

$$r(I) = \frac{\bar{r}}{\cosh(\Delta EI/k_B T I_c)}, \tag{3.15}$$

with I_c as a critical current, and the cosh function arising in the two exponentials of two dwell states. Different tuning curves are like different basis sets. From the individual response, to the summed response, normalized, is

$$\begin{aligned} r_{j,out} &= \sum_{i}^{N} w_{ij} r_{i,in} \\ \therefore \;\; R &= \frac{\sum_{j=1}^{N} I_B r_{j,out}}{\sum_{j=1}^{N} r_{j,out}}, \end{aligned} \tag{3.16}$$

which can be given physical meaning via information-maximizing construction from a basis. This is the retina model now. We use feedback to control the currents and use the Fisher measure to achieve maximum information content. Classification follows. One has achieved a function $H(\theta)$ that codes the population, one can have multiple tuning curves using different currents, one cam minimize errors by maximizing information measure and it is now the tool.

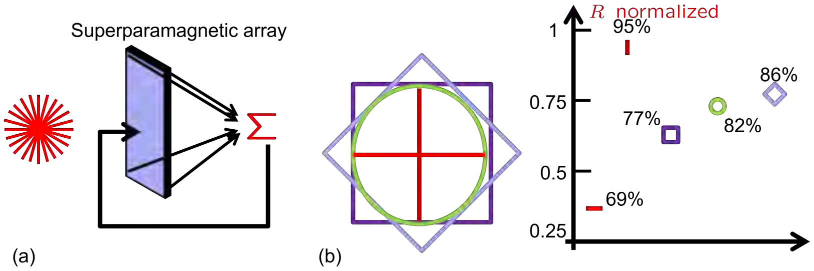


Figure 3.11: Using superparamagnetic array to train lines in various orientation and maximizing Fisher information metric in (a), and then employing it to test various shapes consisting of shapes in (b).

Figure 3.11 shows the training using edges in the toy model described above, and then using the trained simple network—there is no thresholding here, just a basis-based information maximization—to look at various shapes composed of horizontal, vertical, and circular edges. Even if simple, the worst case classification error has 69 % accuracy.

This is very low power because superparamagnetism is in a very small device structures, the currents are extremely small, and one can see in this case that it works pretty well. The worst case is the horizontal one where the crucial information is at a very low state and therefore probably the errors in minimizing that are also at the very very low state. Fisher information as an accuracy metric for parameter estimation has given a method to take advantage of the superparamagnetism in mimicking what the eye is actually doing through the tuning curve.

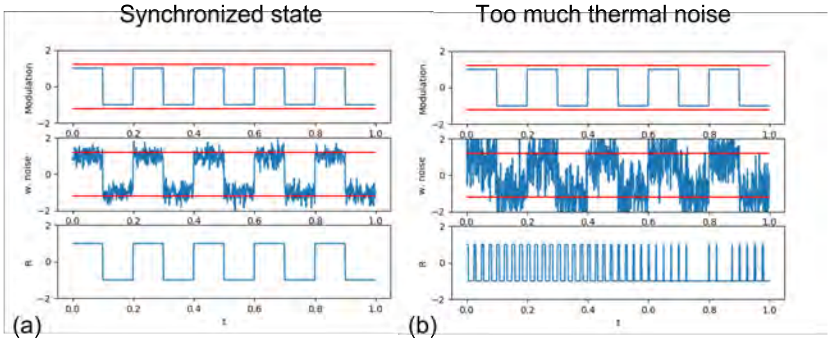


Figure 3.12: Thresholding with a synchronizing signal and thermal noise. With high and low threshold that causes the noise to bump past the threshold leads to a recovery of the signal as in (a). If the nose is too large, a spiking behavior appears as in (b).

To show the power of appropriate thresholding with thermal Gaussian noise to recover a synchronous signal, here is an example that can be understood through the toy model of Figure 3.12. The square wave is a periodic signal. If it has thermal Gaussian noise it looks like the signal in the middle panel. If the threshold is suitably chosen with the noise not too large, an up transition and a down transition can both be found and a cleaner signal recovered as in Figure 3.12(a). If too high a noise, one would get a fairly random spike pattern.

The noise here has behaved like dithering with linear superposition on modulation bringing out the synchronized signal. The synchronization is nonlinear and the signal fidelity improved by nonlinear removal of the thermal noise. It is noise, and its energy, that aided by making the signal cross the threshold to make the synchronization happen.

The reader may now want to think through how this toy model also corresponds to why the bridges swing with large crowds walking on it, and why that could be disastrous if poorly designed or made.

Figure 3.13 is example of signal unveiling using this method. Prof. Martin Luisier and Prof. Juerg Leuthold were my hosts during 2021–22 sabbatical leave at *ETH* Zurich. On the left is a poor picture—

too dark, but with signal buried and noticeable—, which by adding Gaussian noise could be made more recognizable. The gray-scale picture thresholded so that each pixel becomes either dark or light depending on whether it is below threshold or above threshold. Some dark pixels may become light and some light ones dark. But our eye performs local averaging over pixels. This leads to a gray-scale impression despite it being a white or black pixel. With an optimal amount of noise—*it is that noise that is useful*—and our eye—*it was the grayscaling by the eye due to its limited pixel resolution*, thresholding, and adding of noise that has improved the visual perception.

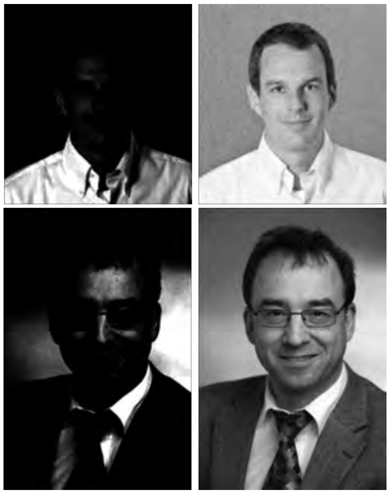


Figure 3.13: Prof Martin Luisier and Prof. Juerg Leuthold in poorly recognizable form—too dark—and then with Gaussian noise added so that eye's pixel grayscale averaging leads to a clearer image.

I would like to see how this same superparamagnetism can be employed in random number generation. One needs lots of independent random number generators if one wants to compute with probabilities. Figure 3.14 outlines some of the issues, the problems of correlations, and an example for how one may create an acceptable stream. Figure 3.14(a) shows the issue of initiation, that averages take time to settle, and consecutive bit correlation by pacing bits through an XOR gate. Are successive bits truly independent? Turns out not so, which is to say that while the phenomenon is physical, it is not naturally random such as that arising in quantum uncertainty. But, by partitioning the stream it is possible to achieve acceptable randomness. One breaks the signal in chunks, and then XOR putting them together to remove any residual correlations.The first few bits are highly correlated, and it takes at least 5 bits for the signal to settle at an objective of 1/2 probability here. If one takes 4 random generators and puts them through three gates, one still sees issues, it takes eight generators, and three levels of compositions to obtain a whitened stream.

Superparamagnetism requires about 20 fJ per bit. Good useful random number generation does require this whitening, but with some penalty in area from the circuits, it is possible to produce a white stream as shown Figure 3.14(c). This is now useful. With Muller C elements, which are flip flops with hysteresis, it becomes possible to perform non-Turing on-the-fly computing through probability manipulation. For a simple toy example, take a dictionary of known words with their associated occurrence rates in spam and non-spam messages. Associate each word of the dictionary with a probabilistic random binary generator whose probability of drawing a 1 is set to different values depending on the presence (or absence) of the word in the presented sentence. Create multiple binary random generators and use Muller C elements for Bayesian inference, and it becomes possible to classify message streams based on the occurrences of key words.

We now have a flickering superparamagnetism-based pseudoran-

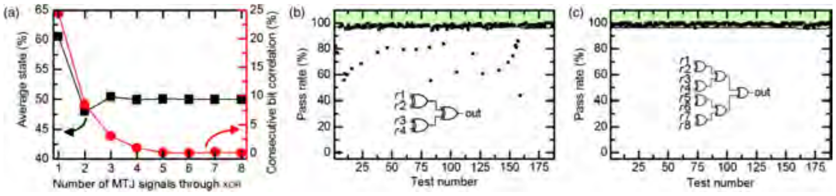


Figure 3.14: (a) shows average state probability following passing of number of superparamagnetic junction signals through an *XOR* gate to see when it becomes truly random. (b) shows the testing of randomness through two compositions and (c) through three stages of composition of chopped up bit stream to get relatively clean random signal. (Phys. Rev. App. 8, 054045 (2017)).

dom generator, we can create a whitened stream, the junction with current-driven stream (Figure 3.9 is one example) gives one probabilities when combined with *XOR* gates, a $20 fJ/gate$ energy, this gives one an ability to implement inference and computing—a non-Turing form of computing—using stochasticity as an intrinsic part of the edifice.

There are other methods too for exploiting randomness for probability generation. The earliest examples have been based on the instability arising in inverter biased in its high gain region that separates the two more stable and broad states. Such a gate driven by amplified noise will have a response of a stream of highs and lows, whose averaging with normalization can be seen as a probability, with the probability magnitude modulated. However, such an approach is very high energy, far more than that of using superparamgnetism.

Recall the Figure 2.3 used to emphasize how certain pieces of information are more important than others, and how even just a very small additional, yet incomplete, piece of data, turns out to be very informative, and suddenly the situation entirely crystallizes. *There is caution needed with this emphasis. It is in context, and one doesn't know the context before the inference.* It is a surprisal. Figure 3.15 stresses this by contrasting it with Figure 2.3. This is an example of a Mooney face. Strong light places some of the face in saturation and some black. They may look confusing, but suddenly they become realistic. In opposite contrast, they still have a problem with interpretation. We have information from the past of what faces look like. This is our learned prior model. We can fill missing information as we did with (a) panel of this figure, but with reversed contrast, the learned prior model doesn't really exist, and the priors cannot be related to the new image. This is our prior model having a say in what the posterior is.

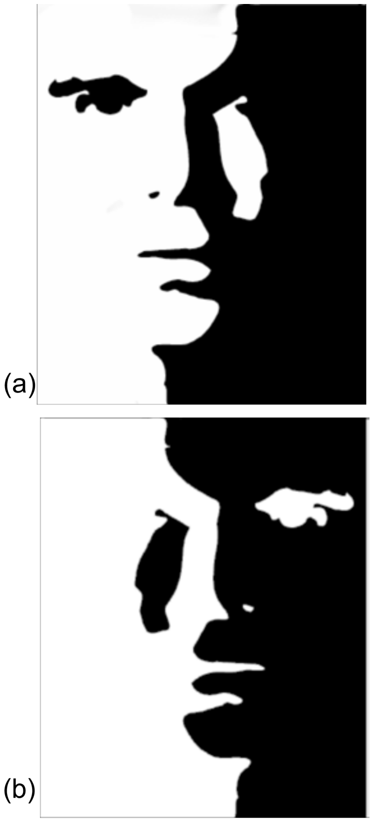


Figure 3.15: Information provided through data in context, and the data's reverse do not provide the surprisal. In (a) here, a face showed up, but in (b) with reverse fill, it is not as clear. Posteriors in the two situations are of different clarity.

The brain learning and processing information from the data is multiple steps of coding, encapsulation, and transformations. While is all mostly physical, as in the kiln viewing till the photons hit the retina, after that comes another set of transformations, electrochemical spike signaling, piping of this information to the visual cortex, passage through $V1$, $V2$, $V3$, and $V4$ hierarchical system that progressively again transforms and encapsulates in forms that emphasize information maximization, but now all this has nothing to do with photons. It is in a brain-appropriate electrochemical spiking form and lots of rhythms and cycles are playing.

That this hierarchical building is important and powerful. The brain's transformations interfere with the direct actions on the basis of which a model has been incorporated in some abstracted form in our brains.

This hierarchical building and information transformation and encapsulation, quite independent of what physical modality it had in some initial stage, is a very powerful idea that I have stressed elsewhere. See S. Tiwari, ``Coming of age with the transistor,˝ www.ieee.orgnsperiodicalsEDSEDS-JANUARY-2023-HTMLindex.html for the story of Hora and Tempus. In the brain, this abstraction means effective sensory fusion. A young ferret, if it loses connectivity in its vision system in its brain, can rewire to use the audio part for partial recovery. Both the video and the audio have similar information processing themes, and the brain has it figured out. Something similarly powerful is happening in many of the new developments of deep neural networks that are appearing in so many forms. From words to equations to deep ideas that they relate is all information. That there is hierarchy and transformation buried in there is most powerfully seen in our viewing and acting based on that of ourselves in the mirror. It is trivial to cut another person's hair. Try to cut your own hair and getting all the rotational and translational transformations right looking at yourself in the mirror. More difficult yet, look in the mirror and try to tie a bow tie.

This prior to posterior relationship based on something new learned is what Bayesian rule and methodology provides a tool for. Unlike Fisher-like thinking of parameterization of a repeatable statistic, which can tell us parameters, and perhaps also sometimes tell us how much confidence one may have in them, but only probabilistically, and in the process allowing false positives and true negatives to also pervade and cause serious conclusion problems, Bayesian is a truly powerful way that allows posteriors to be developed

One way to contrast this with the Fisher parameter statistics is to say that parameters are not something to be found, all data leads to newer inferences since they are new information.

With observations x_0, hidden variables x_h to be inferred that one doesn't necessarily know or recognize, and contextual variables x_1 that one knows about,

$$\mathfrak{p}(x_0, x_1|x_h) = \mathfrak{p}(x_0|x_1, x_h)\mathfrak{p}(x_1|x_h), \tag{3.17}$$

where $\mathfrak{p}(x_1|x_h)$ is the prior. Since

$$\begin{aligned} \mathfrak{p}(x_1|x_0, x_h)\mathfrak{p}(x_0|x_h) &= \mathfrak{p}(x_0, x_1|x_h), \\ \text{it follows that } \mathfrak{p}(x_1|x_0, x_h) &= \frac{\mathfrak{p}(x_0|x_1, x_h)\mathfrak{p}(x_1|x_h)}{\mathfrak{p}(x_0|x_h)}. \end{aligned} \tag{3.18}$$

In this last equation, $\mathfrak{p}(x_0|x_h)$ is independent of what is hidden. It is a normalization factor. x_h can be marginalized out. We have now, independent of what all can affect and is not known x_h a new probability of what x_1 should be expected to be, given the prior, and observations of x_0. This relationship form allows one to maximize $\mathfrak{p}(x_1|x_0, x_h)$ by a posteriori estimation of x_1. This can be done at several hierarchical levels to arrive at inferences, such as matching patterns, which means that it is useful for a large class of difficult computational problems.

With probabilities, we can now compute with Bayesian operators manipulating the probabilities. Bayesian multiplication is multiplication of probabilities, which is the $\mathfrak{p}_{AND}$ gate with

$$\mathfrak{p}(Output) = \mathfrak{p}(Input_1) \times \mathfrak{p}(Input_2), \tag{3.19}$$

and Bayesian addition, which is the $\mathfrak{p}_{ADD}$ gate with

$$\begin{aligned} \mathfrak{p}(Output) &= \mathfrak{p}(Input_1) + \mathfrak{p}(Input_2) - \mathfrak{p}(Input_1) \times \mathfrak{p}(Input_2) \\ &= \mathfrak{p}_{OR} - \mathfrak{p}_{AND} \end{aligned} \tag{3.20}$$

This to eliminate the 1 probability for both of the inputs, which could alternately be accomplished in a circuit form.

The probabilistic inferencing can be mapped. Algorithms are precise description of the state and the state change's evolutionary law. Flow charts are graphical representations of this same prescription of change. The only thing different in probabilistic way of thinking is to consider these as changes over state distributions, where one doesn't have complete deterministic information. So, it is a probabilistic change. The evolutionary law is the Bayes' law.

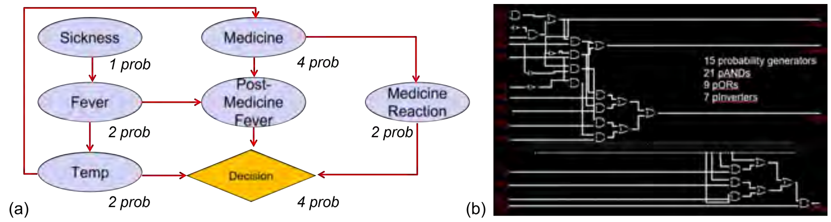


Figure 3.16: (a) shows a graph of an example decision making in a medical problem. This graph can be reduced based on the Bayesian inference mapping (without the feedback look which is an issue for Bayesian circuits) to the circuit shown in (b). This circuit has parts that can be broken, without significantly affecting the response.

As a toy example. take a trivial medical situation: some observations such as of temperature, a prescription of medicine, looking at the response and then modify. This is all mappable using Bayesian probabilities, and as more data accumulates, the inference mechanism improves. The probabilities are in a pulse stream whose average represents probability. Figure 3.16 shows the implementation of the simple graph that takes into account the beliefs and the evolution of the changes in those beliefs. There are plenty of probabilities that are needed that one can generate and feed. All one has to do is now average out and obtain the probability at the output for the inference.

What is most interesting, apart from the low power, is the ability of this approach to become a little more robust to errors. There are parts of the wiring and probabilistic gates that could be broken, yet the system will produce a result that will only be degraded, some only slightly. This is not so in deterministic approaches generally. This is one of the intrinsic powers of such implementations where probabilities are central. So to an extent this also works with the neural networks too. We see in this that the typical problems of decision making, all based on incompleteness unless highly circumscribed, can be implemented.

The *TCP/IP* protocol referred to earlier uses such a probabilistic method to avoid conflicts in between multitudes of information transmission streams passing through a network. Randomness built in minimizes the conflicting instances.

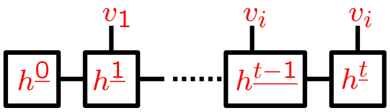


Figure 3.17: A hierarchical Markov chain.

Probabilistic methods can be applied to Markov chains. In Figure 3.17, a traditional Markov chain now written in distributed state probabilistic representation, one can use the Bayes' rule of updating beliefs and apply it on conditionals.

$$\mathfrak{p}(f_l^t|s^{1:t}) = \frac{\mathfrak{p}(s_l^t|f_l^t)\mathfrak{p}(f_l^t|s^{1:t-1})}{\mathfrak{p}(s_l^t|f_l^t)\mathfrak{p}(f_l^t|s^{1:t-1}) + \mathfrak{p}(s_l^t|\overline{f}_l^t)\mathfrak{p}(\overline{f}_l^t|s^{1:t-1})}, \tag{3.21}$$

with the state probability parameterizable (α and β are two probability measures related to s and $\bar{s}$).

$$\mathfrak{p}(s_i|f_i) = \alpha\mathfrak{p}(s_i|\overline{f}_i) = \alpha\beta\mathfrak{p}(\bar{s}_i|f_i) = 1 - \alpha\beta\mathfrak{p}(\bar{s}_i|\overline{f}_i) = 1 - \alpha\beta. \tag{3.22}$$

This is now a tool for a dynamic system, which can update, and quite a good way for usage in situations where some error is acceptable, specially when the timing and decision making does not leave room for deterministic correctness.

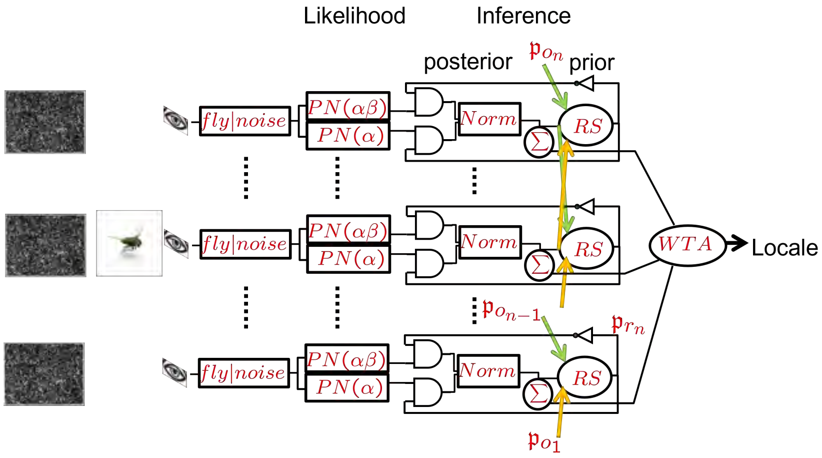


Figure 3.18: A probabilistic Bayesian inference implementation where the presence of a fly and its motion is detected and used to follow the locale.

The eye's flicker-based detection now can be mapped in a non-Turing platform as shown in Figure 3.18. The fly's presence is determined probabilistically through the background of noise, and prior and posterior calculations performed to determine the fly's locale as shown in Figure 3.19. There now exists a tracking ability in real time.. This means that this mechanism can now be a controller for other usage where the tracking information can be employed.

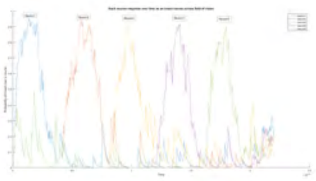


Figure 3.19: A hierarchical Markov chain.

Iteration in Markov chain can also let us do contraction mapping using iterated functions for various usage. One of image recovery is shown in Figure 3.20. One started with a poor image, but by iterating beliefs through the Markov chain, one could keep updating beliefs as one cycles, and obtain in 10^4 iteration a classifiable and clean-enough image for recognition purposes.

3.6 *In praise of randomness and indeterminism*

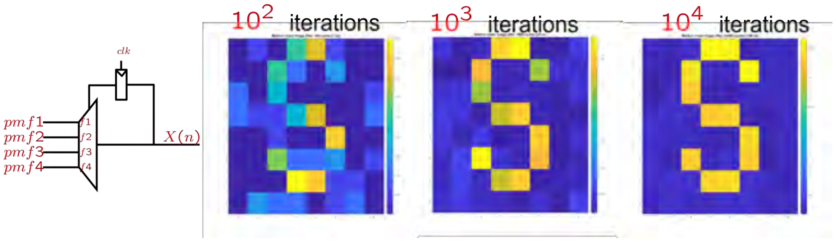


Figure 3.20: Iteration in Markov chain by clocked tuning to achieve image recovery.

I HOPE THAT THIS DISCUSSION and the different examples have been able to get across the message that accepting indeterminism, incompleteness, even negative capability leads to many useful directions and that the approaches of the lowest significant bit relaxation of deterministic computing, or of just using less bits to approximate a probability, are of limited utility. Finding ways to use randomness and probabilities that are low energy makes a whole new spectrum of difficult problems solvable, where one can start getting inferences in which one can place probabilistically quantifiable trust. The methods in the process also become more robust to defects and failures in the parts from which they are built.

This should not be surprising. Randomness is useful in daily living. We love forests because of the randomness of the greenery and the leaves and everything else that makes nature so interesting and peaceful. Cities are not. Cartesian forms in inordinate amount induce Le Corbusier prison suffering.

I have never appreciated Le Corbusier. Initially interesting, a bit of what makes Lego blocks interesting, but eventually you see too much order, less freedom, and then one learns the defining principle of cities as one of compartmentalization. Industry in one sector, offices in another sector, houses in another sector. This is regimentation. More like a military marching band than a Schubert piano piece. Imagine an ocean, rivers, lakes that are straight line and edges, or forests as farm fields of lines of corn. Surprisal is also aesthetically pleasing.

When implemented with care there is a large set of problems—most at the edge-of-network in human-centric interfaces of all types—can be useful. Stochasticity uses low energy, is usually implementable in small areas, is tolerant to error and becomes progressively precise. The Bayesian methods provide robustness and the increasing precision. The edge of the world where low power and energy is desired, and there are multitudes of tasks, is due for such a change.

4
Science-guided AI/ML: Why, how and usage

One of the intrinsic powers of machine learning and neural networks *AI/ML*, compared to the traditional science is that it sidesteps the spherical-cow syndrome. Most of our physical learning constrains a system to a form amenable to known science by building boundaries that enclose it. *AI/ML* employs chance and nonlinearity as an intrinsic part of its computational edifice. This is the profound aspect of this new approach to computation that was only being hinted in the early work of gamblers (Cardono, who was also a noted mathematician), preachers (Bayes, who used chance as a platform to establish god's existence), and serious full-time mathematicians (Laplace, Bernoulli, et cetera). The tension between data-guided agnostic computation and physical principles guiding information and the evolution of the system is unresolved. It is perhaps a mirage, a conjecture would be that they should lead to similar guidance for a dynamic system since information in observation and the physical laws must represent the duality of nature. Science-guided *AI* and *ML* approaches give a powerful new tool for tackling hard and soft causality. Cross entropy, Lagrangians, and Hamiltonians as extremization approaches using Bayesian principles are complementary, but with different constraints due to the underlying mathematical principles and descriptions deployed. We explore this range for real-world open boundary problems to analyze how sciences-guided *AI/ML* can be useful in complex problems such as those encountered in the broader set. For a broader problem that a human is good at, an example is of classifying or generating a specific classical composer. For a narrower problem, an example is of integrated design with its constraints of layout, cross talk, speed, and power. The corollary of this view is that one can extract the physical mechanisms from the data. Such an extraction shows the power of this information-based approach in a physical data.

Three years ago, I was on a sabbatical leave at *ETH*, Zurich. It was an opportunity to take care of one of my promised Oxford books, teach to an audience whose mathematics skills I could largely trust, hike and walk a beautiful country, but also to learn new scientific

Artificial intelligence is a self-serving phrase that I dislike even though it was coined by some of the people—Shannon, McCarthy and Rochester—whom I admire tremendously. it is a self-serving propaganda obfuscating the truth. Neural networks is a tool that the luminaries didn't see when this word as a combinatorial computation driven word was coined to attract attention. Neural networks are not Boolean logic of traditional computation implementation with a completeness framing. Nonlinearity and randomness is essential to the success at human-amenable and human-level tasks that such machines have succeeded at. I feel far more comfortable with *ML/NN* or *NN/ML*, rather than *AI/ML*. *AI/ML* unfortunately will become the accepted form. *It is now in the hands of silicon valley.* Co-opting of words used to be a political and management way to power and money. It is now a very normal technology business practice. Open source has been a democratic and robust useful engineering practice since its start as an Open Source Initiative and the introduction of tools from Free Software Foundation and the principle of copyleft. ChatGPT is owned by OpenAI, a nothing open here company, and GitHub, a repository for software written by the masses, is owned by Microsoft. Both are practicing classic mining of all creations and software of individuals and groups, not unlike the physical mining companies. Word smithing first, logic next. Even scientists and engineers have taken to it for self-aggrandizement. Just look up the fathers of any technology, or the *-onic* ending that is placed to start a new field. Science and technology builds on past works until a moment of creation comes, and then Matthew's principle takes over. So I use *AI/ML* under protest. *Thought leaders* beware.

tricks at a great world university. I attended three different classes to explore the intersection of applied mathematics and statistical and machine and deep learning.

What got me interested in this subject area is its value as a new technique for building approximations of complex systems even if one didn't quite understand how the immense number of give-and-take interactions, hidden variables and chance events, all play out. Fokker-Planck equation and its Markovian derivation was the limit of my understanding of such complexity.

During the world war II, as the train network of Japan was bombed to stop the military production infrastructure, the Allied powers couldn't figure out how little it affected for a time. The Japanese had a model of the train network system in the form of pipes and flow of objects. Any network connector that got bombed could be shut off and the physical model showed how the flow responded. This was a pointer to how movement of items should be diverted. The flux of the items, with pressure, fluid flow, and the objects, could be transformed to what the train system could do. Information was in these parameters. The fluid network was a complex system that knew how to manipulate information with the only access being to to the external parameter of pressure and what was being put into the fluid network. Another example of correspondences of information flow, information and physical, and the networks in all their form.

But complex systems have been of interest for ever since there is so much that cannot easily be explained, so much that is turned into god's will. I am reminded of a Neils Bohr story that is in one of my books at the start of a quantum mechanics discussion. Bohr talks about a young person in a village who was sent to another village to listen to a great rabbi and to come back and report. When the young man—the student—came back, he said. ``The first lecture was just brilliant. Clear and simple. The rabbi understood it and I understood it very clearly. The second was even better. Deep and subtle. I didn't understand much but the rabbi seems to have understood it. The third was just superb and unforgettable. I understood nothing and the rabbi himself didn't understand much either.''

This is the third of these essays, edgy, where much is not known and understood, and just like those early days of quantum mechanics, it is one we must understand. *One should be very skeptical and very afraid of using something that one doesn't understand well enough.*

Even as a child, I had realized that the larger the lack of knowledge, which later on I learned to be the condition of the higher the entropy, the more is the inclination to postulate the drama as god's will.

4.1 *Black box, not knowing and delusions*

With this as a context, I'll quickly walk through a few of the beginning thoughts that are on top of my mind related to *AI* and *ML*. This will set the context of what I want to talk about, which is understanding what is going on in these systems, and how to make it accurate by using our science principles. It is an attempt at synthesis and analysis of a new tool, which in turn seeds new ideas that are the evolving story of this information age.

The first item regarding black boxes is of course the question of entropy. We know today, and didn't so necessarily during Claussius' or Carnot's time, and why Maxwell's demon occupied hundred years of the brightest minds, is that entropy is lack of knowledge. The whole edifice of the understanding of the world is based on information as an increasingly abstracted form of hierarchy. That information is physical—information's existence is manifested physically The physical is abstracted from hierarchies, protons, neutrons, electrons,

To Rolf Landauer's ditty, *information is physical*, or Wigner's *it from bit*, I have a corollary, *physical is information.* I will dwell on this in Chapter 5 while discussing our place in this world.

to atoms, to molecules, to natural and physical bigger objects, to objects reproducing, or objects achieving capability to observe and imagine and act, and so on.

This can be confusing, entropy and information together, not knowing meaning larger entropy. I have had to spend time on this in my class here too. It is best enunciated by the Czech writer Karel Čapek, ``This is just . . . entropy, he said, thinking that this explained everything, and he repeated the strange word a few times.˝ in the play Krakatit. A child's room may have high entropy for the parents, but for the child it has low entropy. The child knows where things are.

Karel Čapek also used the word robot in the play Rossumovi univerzàlnìí roboti (*R.U.R.*); the word itself is from his brother Josef. Robot, the combining of classical machinery with information machinery, is derived from robota (drudgery or serf labor). These are real now, both for the hard labor but also for wars. So is the use of *AI/ML* in so many repetitive tasks and selling and opinion molding for exploitation tasks .

In the decade after the earliest work of quantum mechanics was done, Paul Dirac[1] had remarked ``The general theory of quantum mechanics is now almost complete, the imperfections that still remain being in connection with the exact fitting of the theory with relativity ideas. . . . The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact applications of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.˝ We are still working on this task and interestingly *AI/ML* may have much to offer in resolving the complexity that abounds all many-body problems and the problems of open boundaries. What is true for quantum mechanics development, still ongoing now with quantum computing as a frontier, applies to *AI/ML*.

I am also reminded of Isaac Asimov's 3 laws in *Runaround* (1942), amended later with the 0th law.
1st law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2nd law: A robot must obey the orders given it by human beings except where such orders would conflict with the 1st law.
3rd law: A robot must protect its own existence as long as such protection does not conflict with the 1st or 2nd law.
0th law: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

These are constantly flouted in all of Google-Facebook-Microsoft-OpenAI-Tesla-· · · enterprise. Asimov, in *Foundation and earth* (1986), also foresaw this and captured it in the conversation: ``Trevize frowned. ``How do you decide what is injurious, or not injurious, to humanity as a whole?˝ ``Precisely, sir,˝ said Daneel. ``In theory, the 0th law was the answer to our problems. In practice, we could never decide. *A human being is a concrete object. Injury to a person can be estimated and judged. Humanity is an abstraction.˝*

Really understanding what is happening in the computation and having science guide us as a mathematical tool gives physical meaning and physical constraint. This one may argue is what makes it real and usable since one now can understand the bounds of its applicability. I realize that this has nothing to do with how the guiding has anything to do with drones flying around, automated remote gears blasting and killing in foreign streets, searches narrowing the phase space of human beings to very narrow polarized domains, or the separation of blood-on-hand human feeling versus that from being 10, 000 miles away through a remote *AI/ML*-guided action, but I hope

[1] P. Dirac, Proc. of Royal Society (1929)

that bringing an understanding lets one work on interesting, challenging, and important problems that are intellectual and perhaps useful for society.

All things scientific and technological are double-edged swords. I am quite aware that ChatGPT, when being asked to choose between your survival and my own, at least at one stage of development, chose its own and many early conversation systems rapidly turned racists.

Our brain too has these conundrums of interpretation since it too does plenty of transformations under some guiding principles, all of which we do not understand. Ethics and morality are malleable and change with time. The new testament replaced the brutal old testament. Mahabharata and Ramayana have plenty of stories of behavior by ``good guys˝ that are not acceptable today. Oppenheimer quotes Bhagwatgeeta at the creation of science's possibly worst contribution to mankind. The brain has some generalized attractor states if you look at this landscape optimizing a lot of principles. Sometimes those principles conflict. Even an Asimov-sanctioned robot may encounter situations where this conflict will appear. *AI/ML* is work in progress just as we are.

Intelligence, learning, concepts, connecting concepts over domains, imagining, imagining possibilities even before encountering them, love, taste, buzz and how many of the best writers were also drunks, and so much more, is what constitutes and contributes to how the brain works and how it manifests itself in different individuals. Sometimes we are at our best and can see more acutely than other times. We improve. We also regress.

Brain can misinterpret because of its biases and because of much that it does is based on prior model. Take illusions. They are conflicts created in the brain. Even images can be a source of error in interpretation. Figure 4.1 shows an example from a psychology and neurosciences class. (a) shows two identically colored patches on a checkerboard with the same light incident, but with a cylinder casting a shadow. (a) shows that the brain infers that less lighting is incident on the lower patch because there is a shadow. The brain's model says that the lower patch must be reflecting a higher fraction of the incoming light. So we interpret it as lighter even though, as seen in (b) they are identically bright.

Reasoning, counterfactuals, and other such human approaches are not yet in the *AI/ML* arsenal even if large language models and long short term memory and others of similar ilk have had quite some success. There are also serious failures just as with humans. Probabilities, when we first encounter them in our learning, for example, lead to much angst.

Is it the atom bomb or the plastics that are the worst contribution to the earth and living kind? I cannot make up my mind.

The famous one from psychology is of false choice in the trolley problem that I slightly modify here. A trolley is coming down fast on a track, I am standing on a bridge and have access to the switching bar that can move the trolley between two tracks. It is barreling towards a family of five with four children, but can be diverted to the other on which a fat man is walking. What should I do? Psychologists call this a false choice. Philosophers would say negative duties carry significantly more weight in moral decision making than positive duties. No solution. Some ethicists would say the greater good of saving more takes precedence. My reaction would be to get me out of here. I don't wish to deal with this. But I sympathize with the philosophical view. Negative produces productive friction. It places obstacles to slow people down and grapple with consequences.

Google's AlphaGo winning over the world's best Go player got large publicity in 2017. But, in 2023, the reverse of a player beating one of the best programs also happened. The player had used an *AI* tool to find the fault in the other program.

The great Erdos at first was not convinced at all by the famous Monty Hall problem. I ran ChatGPT through the boy-girl paradox. ```I have two children and at least one of them is a boy.˝ What is the probability that the other child is a boy?˝ At least one child is a boy makes us rule out one of four possible cases, leaving the other three equally likely. But this requires reasoning, or those who wish to do it analytically, using Bayes' rule. Language models don't reason, they have learned some mapping based on what was fed. They don't do mathematics quite completely either. So ChatGPT got it wrong. But, it answered Monty Hall right. Monty Hall is something that it has been fed with before. Boy or girl paradox not so.

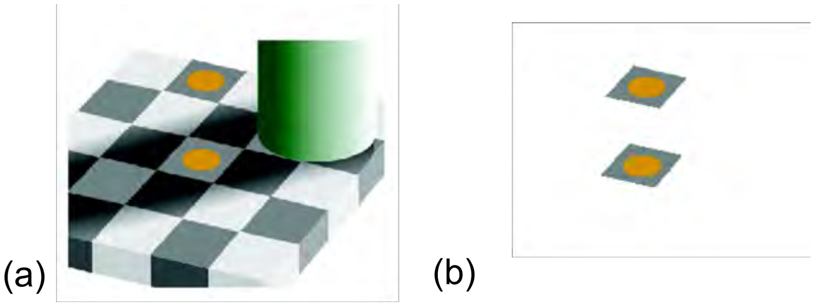


Figure 4.1: (a) shows a checker board with a shadow on two colored patches with the same light incident.(b) shows what the patches are really like absent the shadow.

So the problem with large language models as of now that we know is that there is too broad a coverage over all the nuances of human speech and foibles and all that it has been fed which may be questionable. There is no editor as such on the information—the known problem of social media by intention or by finance pressure. Small models, on the other hand, guided by their discipline where they understand the niche technology, and have the specialists definitions programmed already, work really well.

4.2 *Probabilities, state changes and the great Andrey Andreyevich Markov*

The story of probabilities, inference, machine learning and *AI* inevitably leads one in pretty early stages to a discussion of Markov chains. Markov chain is a way to model the changing of states with probabilities. A working example is that weather can be sunny (*s*), cloudy (*c*) or rainy (*r*) as seen in Figure 4.2. If it is sunny than there is a probability of it still being sunny is 0.6 tomorrow, or cloudy is 0.3, or or rainy is 0.4. Tomorrow's probability is normalized. Sunny, cloudy or rainy are the only possibilities and add to 1. With different probabilities, there is also a similar description for the possibilities if the state of today is cloudy or rainy. These transitions are captured in a probability matrix that is now shown in Figure 4.2 and captured in Table 4.2.

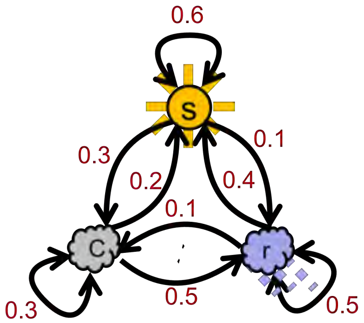


Figure 4.2: Probabilities of transitions between and staying in in sunny, cloudy and rainy states.

	s	c	r
s	0.6	0.3	0.1
c	0.2	0.3	0.5
r	0.4	0.1	0.5

Table 4.1: State transition probabilities. The 1st column identifies the state today, the rows are the probabilities of the state tomorrow as identified at the top of each column. This probability matrix is the quantitative representation of the figure.

This is a probability matrix ꝑ$^{\underline{1}}$ of the 1-day change. What about 2 days? If it is cloudy today, what is the probability of rain in 2 days? This question is answered by state transitions. First, after one day, one knows what the probability is of being sunny, rainy and cloudy tomorrow. And then based on this, one needs to calculate what the possibility is of ending up in rainy state in 2 days. Cloudy today to sunny, cloudy and rainy tomorrow to rainy day after tomorrow. The simple calculation that picks all the possibilities ending up in rain is

$$\mathfrak{p}^2_{2,3} = [\mathfrak{p}(c) = 1] \times \begin{bmatrix} 0.2 & 0.3 & 0.5 \end{bmatrix} \times \begin{bmatrix} 0.1 \\ 0.5 \\ 0.5 \end{bmatrix} = 0.42, \qquad (4.1)$$

a multiplication of row and column. This is a calculation of conditional probabilities given a state. Cloudy today, gives me the probability of cloudy, rainy and sunny tomorrow, and conditioned on this, one is determining what the probability of rainy is. 3 terms corresponding to 3 intermediate states. Matrix algebra makes this easy. One can also similarly find what it will be for sunny and cloudy with the last column being a different pick from the matrix of Table 4.2. The net is that the new probability matrix of the possibilities two days from today is just a recursive product of the present day matrix and the transition matrix.

This is captured in the collection of $\mathfrak{p}^{\underline{1}}, \ldots, \mathfrak{p}^{\underline{7}}$ for 7 days listed in Table 4.2.

	0.600	0.300	0.100
$\mathfrak{p}^{\underline{1}} =$	0.200	0.300	0.500
	0.400	0.100	0.500
	0.460	0.280	0.260
$\mathfrak{p}^{\underline{2}} =$	0.380	0.200	0.420
	0.460	0.200	0.340
	0.436	0.248	0.316
$\mathfrak{p}^{\underline{3}} =$	0.436	0.216	0.348
	0.452	0.232	0.316
	0.438	0.237	0.326
$\mathfrak{p}^{\underline{4}} =$	0.444	0.230	0.326
	0.444	0.237	0.319
	0.444	0.235	0.325
$\mathfrak{p}^{\underline{5}} =$	0.443	0.235	0.322
	0.441	0.236	0.322
and			
	0.441	0.235	0.324
$\mathfrak{p}^{\underline{7}} =$	0.441	0.235	0.324
	0.441	0.235	0.324

Table 4.2: Products of state transition probabilities for $1, 2, \ldots, 5$ and 7 days.

Just multiplying these matrices gives one a matrix which tells one what the different combinations are going to be. Note how the matrix terms seem to be asymptoting. In 7 days, although 7 is quite similar to 5, it is quite different from 1. The sunny staying sunny changed from 0.6 to 0.41. Depending on the conditional probabilities and initial states, there is both a fast and a slow change. Conditional changes cause fast change in the beginning and then the weather settles down to what the normal expectation of the weather would be. Long term, all one can predict is the averaged behavior one expects, where today's pattern is of low significance. In Table 4.2, we have the same matrix of expectation of sunny , cloudy and rainy independent of what it is today to three places of decimal. It is changes, is quite different from the original state.

Markov developed this probability state evolution description to understand Pushkin's writing looking for a mathematical understanding of the style and order that an author has. A budding great mathematician, with self belief, he looked at how often consonants and vowels appear, how often does another vowel or a consonant follow another vowel, and other versions of this ordered arrangement. Patterns in choices, of course language centric, this is one way of pulling apart the random independent choices from the choices one makes. In writing, and elsewhere, this is a Markov process, an evolution of the state transitions.

There are a number of great stories of belief and standing up with Markov as the progenitor.

In exploring 20000 letters of Eugene Onegin, all by hand, for example, in the line . . . *wastooyoungtohavebeenblighted* . . ., he could start looking for probabilities of vowels and consonants, of vowels following vowels or consonants, et cetera, not unlike the rainy-sunny-

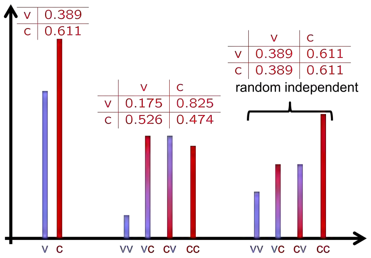


Figure 4.3: Probabilities of vowels and consonants, and conditional probabilities and joint probabilities for Euvene Onegin of Evgeny Pushkin.

cloudy weather description. Figure 4.3 shows probabilities of vowels and consonants, conditional probabilities of one following the other, and joint probabilities of two letter combinations in the writing. People generally have a certain style of writing, there are also other language-related, cultural and other traits, and they show up statistically. It is probabilistic, writing and exposition is complex, specific outcome may not be predicted, but statistical features, and averaged properties can. A Markov process and the chain of these state changes can be extracted through the data.

The state evolution and the chain and transitions can be very complex as one goes to other domains, but the fundamentals still hold. Machine learning, neural networks, all have Markov chains hiding in them.

4.3 *Complexity in time and space.* IIT *graduates from 1976*

Any matter that is complex, not completely definable through a deterministic relationship—either because of uncertainties of the different types and/or because of the impossibility of keeping track of every event causing state changes—is dealt with by probability theory in sciences. Quantum mechanics and life are both examples.

Science can tackle the expectation statistical feature. An important such statistical equation predating Markov that we learn is the Fokker-Planck equation, which describes the evolution of a probabilistic distribution under the consequences of a stimulus and of random events. The simplest example being of drift and diffusion of particles as in motion of carriers in semiconductors or of ink particles in water or the making of yogurt by dropping and stirring the yogurt culture. It applies broadly. In an integral form—discrete is replacing integral by summations—the change from state s' to state s, occurs with probability distribution changing as

$$\begin{aligned}\mathfrak{p}(s,t|s',t-\delta t) &= \left[1-\delta t\int S(s''|s')ds''\right]\delta(s''|s')+S(s|s')\delta t \\ \therefore\ \partial_t\mathfrak{p}(s,t) &= \int\left[S(s|s')\mathfrak{p}(s',t)-S(s'|s)\mathfrak{p}(s,t)\right]ds'. \qquad (4.2)\end{aligned}$$

This probability change is showing the accumulation of transitions of states from previous states due to whatever is the evolutionary rule underlying state changes. A probability equation is describing it in a general form. For Markov's Pushkin example, t is the indexing of letters, s', s are two states connected through intermediate states s''. t is a continuous index here. If I am walking between two points on *IIT* campus, if it is the ten minute period between class ending and new class, I will take longer since there will be plenty of events of my avoiding other people, or bicycles with people on them, making my trajectory change, and even be motion less for periods of time. Other times, I may just follow a very clear Cartesian trajectory. This latter is drift in the channels that are allowed, the former was the effect on this drift by scattering and a diffusive consequence.

This Markov description projects to Fokker-Planck description.

For the particle motion example, and viewed a little more rigorously, one may describe the particle—characters of writing being an example—evolution as

$$\begin{aligned}
\partial_t \mathfrak{p} &= -\sum_{j=1}^{d} \partial_{x_j}[a_i(x)\mathfrak{p}] + \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j}[b_{ij}(x)\mathfrak{p}], \text{ with} \\
& \quad \mathfrak{p}(x,0) = f(x), \quad x \in \mathbb{R}^d, \\
\therefore \quad \partial_t \mathfrak{p} &= \sum_{j=1}^{d} \tilde{a}_j \partial_{x_j}\mathfrak{p} + \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j}\mathfrak{p} + \tilde{c}(x)u, \quad t > 0, \text{ where} \\
\tilde{a}_i(x) &= -a_i(x) + \sum_{j=1}^{d} \partial_{x_j} b_{ij}, \\
\tilde{c}_i(x) &= \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j} b_{ij} - \sum_{i=1}^{d} \partial^d_{x_i} a_i. \qquad (4.3)
\end{aligned}$$

With

$$J := a_i(x)\mathfrak{p} - \frac{1}{2}\sum_{j=1}^{d} \partial_{x_j}[b_{ij}(x)\mathfrak{p}], \qquad (4.4)$$

as a flux—current being a charge flux from the particles being one example, one has

$$d_t \mathfrak{p} + \boldsymbol{\nabla} \cdot \mathbf{J} = 0, \qquad (4.5)$$

which is a conservation equation for the particles. Particles are not being annihilated in this description, and one also obtains a density (particles per unit volume) equation in a Boltzmann form of

$$\partial_t \rho + \mathfrak{p} \cdot \boldsymbol{\nabla}_q \rho - \boldsymbol{\nabla}_q V \cdot \boldsymbol{\nabla}_p \rho = \mathcal{D}\boldsymbol{\nabla} \cdot [f_B \boldsymbol{\nabla}(f_B^{-1}\rho)]. \qquad (4.6)$$

f_B here is the Boltzmann distribution, V is a potential, whose gradient causes the drift, and $\mathcal{D}$ is a diffusion coefficient arising in the scattering events. The probability distribution is evolving since state occupation is evolving under some physical evolutionary law. Particles in this description are being conserved. So we have a conservation equation of the flow. We also have the conservation happening with the evolution of the distribution under some stimulation—a gradient in potential is a field and scattering events between particles that are abstracted by the diffusion coefficient. This form has described the effect in both space and time.

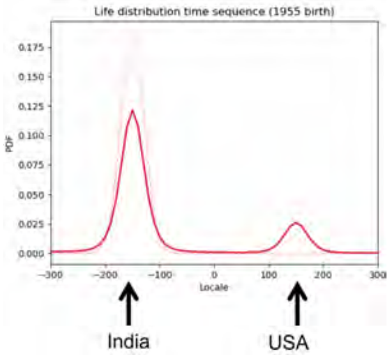


Figure 4.4: The probability distribution function of 1955-born *IIT* Kanpur graduates assuming two locales over there lifespan predicted from a non-conserving Fokker-Planck equation at this point in our life. A few of us have passed away. The dashed line is the starting distribution in 1955. The website shows the animation.

Decades take a toll on memory. The person who I thought had passed away, imagine my surprise, when I met him in person a few weeks ago on this campus when students of my years were here for the alumni gathering.

At some point during this last year in a conversation with a friend—an *IIT* co-student from my time—on hearing the passing away of another student, that we all probably have about five more years to live since life expectancy is 72 *years* for people born in the

1950s. Classical inanimate particles don't die. Particle conservation applies. It does not to us humans. This made this an interesting problem to explore a modification to Fokker-Planck equation with the probability distribution disappearing in time, and the toy model I formed was of being born in India, graduating from *IIT*, and then aging with possibilities of two locales, with a possible scattering from one to the other. Imagine India and *USA*, with some students going off to *USA* following graduation, more arriving subsequently through jobs or a later decision to pursue more education, or something else. And even later on, people even scattering back or scattering forward during the old age. All this can be parameterized which the scattering matrix modifications to Fokker-Planck lets one do. Non-conservation can be introduced via a damping factor added on beyond what is in Equation 4.3. A time evolution of the resulting dynamics can be seen in Figure 4.4 of the *IIT* graduates who were born in 1955 today. *This is drift, diffusion and death* for us. The oldest of us may get past 100 year in age. Markov did this sequence, connection and evolution analysis with letters and words for a specific person, a modified Fokker-Planck describes a similar probability evolution for life of a small enough group of us who have similar characteristics.

The procedure of Markov or Fokker-Planck is written as en equation, so continuous, but could also be written in discreet state transition steps as we did, is akin to building of a graph of connections down which one can percolate. This is the physical that machine learning techniques are really good at.

4.4 *Approximations, uncertainty and transformations for information*

Graphs-choices-entropy, what step to take next with a layer across which the information is spread, is at the heart of neural networks' affine-nonlinear approach. The procedure of Markov or Fokker-Planck is written as en equation, so continuous, but could also be written in discreet state transition steps as we did, is akin building of a graph of connections that machine learning techniques do remarkably well at. In the neural network, the information is spread out in the parameters, and there are many of these, across a layer in any cut and the parameters are gradually being manipulated by the learning and how they are changing from layer to layer to become more useful for the task, which is an end multi-dimensional state of the problem being analyzed. This is a different way mathematically, but is similar to what was encountered with Markov chain or Fokker-Planck. Markov embodied state-to-state (character following character in the earliest case by Markov), Fokker-Planck's state was position and momentum that ends up leading to drift and diffusion as emergent parameters in the simplest of examples. The state of Markov or the position and momentum of Fokker-Planck are relevant information upon which the related evolutionary law is being extracted through probabilities or the drift and diffusion parameters. The neural network is doing such a multi-dimensional state description in the mathematical parameters of the network. So neural networks are just another engine of information flow similar to Markov and Fokker-Planck.

Machine learning can tell us the evolution in space and the non conservation that the partial differential equation connecting probabilities tackled well in the first order with dissipation. Markov did the same in transition probability form for letters.

How can neural networks as a machine learning do it? Quite simply *it is aggregation of data input together while keeping information followed by nonlinear transformation and doing this again and again across layers*, which is at the heart of all evolution and growth. One can even argue that as computing capacity continues to proceed, there is absolutely no reason, why the machine capability will not out master the human capability. For the simplest of examples that brings out major properties of approximations and dimensionality, no nonlinearity, consider a linear autoencoder as a ‶primitive″. Figure 4.5 in panel (a) shows network, where there are nodes at which an input, say $\mathbf{x}$, is being fed, and an output is being estimated, say $\hat{\mathbf{x}}$, with an intervening bottleneck in the transformations being affected of $\mathbf{z}$.

$\mathbf{x}$ is being mapped to a compressed space of $\mathbf{z}$ through a coder $\mathbf{C}$, which is at its simplest just a linear sets of transformations based on aggregation. $\mathbf{D}$ is another reconstruction decoder that maps $\mathbf{z}$ to $\hat{\mathbf{x}}$. $(\mathbf{C}, \mathbf{D})$ is a parameterization, call it $\boldsymbol{\theta}$, so $\hat{\mathbf{x}}(\boldsymbol{\theta}) := \mathbf{DCx}$. If one optimizes a loss function $\ell(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{2}\| \mathbf{x} - \hat{\mathbf{x}}(\boldsymbol{\theta} \|^2$, so one is attempting to match $\mathbf{x}$ and $\hat{\mathbf{x}}$, and chooses $\hat{\mathbf{x}}$ to be forced to be the same as $\mathbf{x}$, that is, an estimate, then $\mathbf{z}$ as a bottleneck layer provides an intermediate representation at lower dimensions. Higher order terms—fluctuations, noise, for example—will be the first to be eliminated. One gets an approximate identity map relative to the data representation. A linear autoencoder is making a low-rank approximation of a linear map $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m$ as a matrix $\mathbf{DC}$. If it is ideal, $\mathbf{F} = \mathbf{I}$, one has the identity map, but if one has placed a rank limitation, and $\mathbf{F} \approx \mathbf{I}$. Rank limitation—a dimensionality reduction—has become a bottleneck. Any linear map $\mathbf{A} : \mathbb{R}^k \to \mathbb{R}^l$ now has a rank, which is the measure of the dimensionality of the $\min\{k, l\}$.

Linear autoencoders are performing a low-rank approximation. The data matrix $\mathbf{x}$ and its approximation $\hat{\mathbf{x}}$ have a ‶loss″-like function

This simplistic description, of data aggregation—data may contain some or no information, some of the information may be degenerate with other, some of it not—followed by nonlinear transformation, so a finessing of informational content in the end is not unlike our daily life. Most of the time, we proceed linearly, learning iterative new things slowly at home and in our schools, which are interspersed with sudden aha moments and insights and new techniques that we learn or figure out ourselves because we have been gradually accumulating insight—a connection of information over domains. These rapid changes are nonlinear. An infant is born with some number insight, 1 or 2 or more, by age of two neuroscientists say that we understand up to 4, and then with an understanding of numbers, suddenly we figure out arithmetic that requires algorithmic knowledge. In human history too this happened. Together or more likely from the number sense, one acquired language skill. Arithmetic, algorithms, require a description of the procedure, and the language is a tool for that. Information transformation is taking place across the layers, if done right with appropriate matching as in filter networks and elsewhere, then it transforms to a form that then is rapidly transformable in classification or generation of something different. This is not unlike the Carnot cycle. Adiabatic changes followed by the sudden isothermal change that causes move by pressure and volume change at a constant temperature. *Carnot efficiencies are everywhere in my view, in our living too.* We as a human cannot function efficiently. We need downtime. We need time with friends for humor and discussions. We need to walk in the woods. We need other interests or cares as passion. No ‶work″ gets done during these interregnums. But this heat-like interlude is important to us being efficient.

In a matrix product, the composition of linear maps, the rank of $\mathbf{AB}$ will be less than or equal to the minimum of the rank of either $\mathbf{A}$ or $\mathbf{B}$. The reverse—a decomposition rank—$\mathbf{M} = \mathbf{AB}$ wth $\mathbf{A} \in \mathbb{R}^{m \times k}$ and $\mathbf{B} \in \mathbb{R}^{k \times n}$ iff the rank of $\mathbf{M}$ is less than or equal to k.

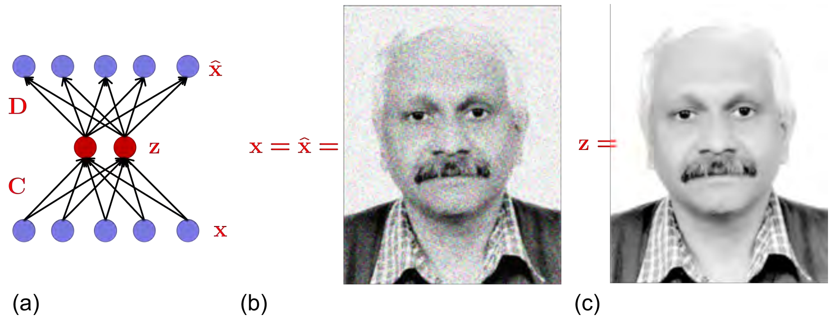


Figure 4.5: (a) shows a linear autoencoder network with a dimensionally shrunk bottlneck between a coding (**C**) and decoding (**D**) transformation of input **x** to an output $\hat{\mathbf{x}}$. (b) shows a case of forcing input and output to be the same (here Prof. Sundar Iyer with Gaussian noise), which results in a less noisy Prof. Iyer in (c).

which is the normalized sum of squares of the difference, which is the Frobenius norm. The Eckart-Young theorem in the singular value decomposition,

$$\arg\min_{\hat{x}:\mathrm{rank}(\hat{x})=k} \| \mathbf{x} - \hat{\mathbf{x}} \|_F^2 = \mathbf{U}\boldsymbol{\Sigma}_k\mathbf{V}^T, \qquad (4.7)$$

where $\boldsymbol{\Sigma}_k$ is a truncated diagonal matrix of singular values. The ˵loss˝-like function can be minimized within this limitation. There exists an optimal rank k approximation that can be obtained via singular value decomposition. This projects to the limits to reconstruction quality.

The Frobenius norm is the matrix norm, that is, $\| \mathbf{A} \| = \sqrt{\{\sum_i \sum_j a_{ij}\}} = \sqrt{\mathrm{Tr}(\mathbf{A}\mathbf{A}^\dagger)}$. It should be distinguished from the Euclidean norm, which is the vector $\ell 2$ norm, even if it reduces to it in the simplest lowest-order case.

Eckart-Young theorem applies to spectral norms, useful for Frobenius norm in achieving singular value decomposition. This is a way to lower rank matrices as approximations (or accurately under strict conditions). Eckart-Young theorem tells us the singular value decomposition for the best approximation in the Frobenius norm.

A Gaussian noise pixelated picture in Figure 4.5 of a young Prof. Sundar Iyer, has its noise—the fast spatial perturbations—removed. Since this is using linear matrix transformations while compacting, via the dimensionality reduction, information is certainly being lost, but the loss is of the fast fluctuating component. This is predominantly noise though certainly important content that is rapidly changing from pixel to pixel will also be dropped. Fortunately, that is smaller in this image with the rank dimensionality choices made.

If there was a bullet—a pixel or two in size—that too could be lost in such an autoencoding. But, for normal pictures, where one is looking for a smoothening and soothing view, this autoencoding does remarkably well. It is the information in the fast perturbations that is being approximated out.

4.5 *The importance of nonlinear transformations*

The linear autoencoder is an unsupervised example of dimensionality reduction in neural networks achieved through an entirely linear sequence of dimensionality reduction across layers represented by the matrices for encoding and decoding of **C** and **D**. The general form of the assembly underlying the ˵coder˝ or its inverse in neural network form is the perceptron[2]. It combines the affine with a non-linear transformation.

A simple illustration of the perceptron embodiment can be seen via the *NAND* gate of Figure 4.6. This is a single layer neuron that first sums on weighted inputs and then produces a thresholded

[2] F. Rosenblatt, ˵ Principles of neurodynamics: Perceptrons and the theory of brain mechanisms,˝ Report VG-1 196 -G-8, Armed Services Technical Information Agency (1961)

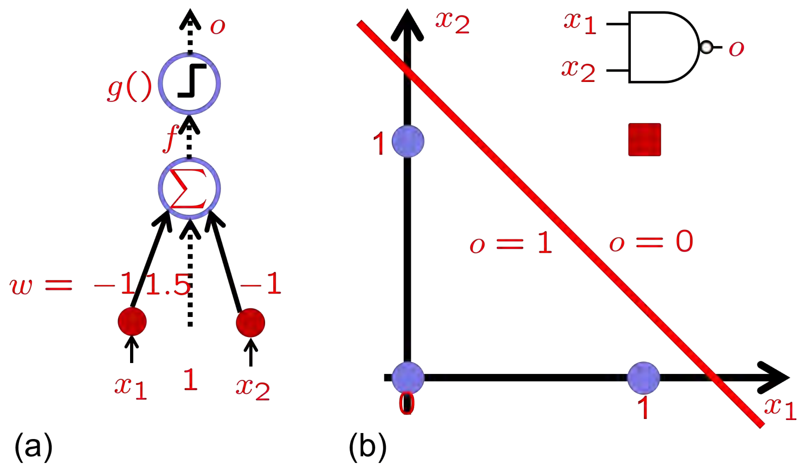


Figure 4.6: (a) shows an elementary single layer neural network for *NAND* with (b) showing the computation of the output.

output employing a nonlinear transformation. $\mathbf{x}$ is the input, an additional offset-like input that is the bias—the 1 input here—are summed with weighting followed by the nonlinear function. Mathematically, we have

$$f^l(\mathbf{x}, \mathbf{w}, \boldsymbol{\beta}) = \sum_k w_i^l x_k + \beta^l, \tag{4.8}$$

where ws are the weight and β is an additional translational bias. In this equation the bias β is weighted by 1 in the summation, and i is indexing to write this form generally for a multilayer network with l identifying the layer. It is more convenient mathematically and computationally to exchange the weight and the bias of the input ($w = 1$, the weight and $q\beta = 1.5$ as the real bias for *NAND* are the accurate representation) since it is is only the product that appears in the affine summation. This is what has been done in Figure 4.6. With these changes and writing generally with indexing,

$$f_i^2 = \sum_j w_{ij}^2 h_j, \tag{4.9}$$

where one of the h terms is the $\beta = 1$ and a weight corresponding to it (1.5) for the *NAND* perceptron. This equation is expressing an input $\mathbf{h} = \{\ldots, h_j, \ldots\}$ for the $\mathbf{x}$ of the input layer but also for hidden layers where the neural assembly is stacked. Rewriting of the weighted bias made this expression simpler computationally on the right. It is expressing f feeding the ith node from the summation over the js of the prior layer with weight w_{ij}^2 scaling the input. The output, with a transformation, non linear in general ($g()$), then is

$$o_j = g^2(f_i^2). \tag{4.10}$$

For $NAND$, the single layer neural network mathematically is

$$\begin{aligned} f &= [-1\ -1\ 1.5] \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}, \text{ and} \\ o &= \begin{cases} 0 & \text{if } f < 0 \\ 1 & \text{if } f \geq 0 \end{cases}. \end{aligned} \tag{4.11}$$

The figure shows that one has performed a linear separation on the input in computing the $NAND$ output.

In statistical terms, this is a linear separation that a non-linear Hadamard function performed following the affine transformation. The linear transformation was a dimensionality reduction. It placed the data in a form suitable for separation. The *classification* happened through the non linear transform. Information was also lost in both steps of this process. It is straightforward to see that XOR function cannot be obtained through any combination form of Equations 4.9 and 4.10. This is because a linear separation will not work since it is an addition modulo 2, it is a test of unlikeness, so of asymmetry. But, stacking multiple perceptrons lets one achieve a neural XOR rendition. Stacking multiple layers, with intermediate layers only connected to layers within the network, are hidden layers, lets one achieve achieve much more complex functions.

Single-layer perceptrons are only capable of learning linearly separable patterns, that is, a separation by a linear dividing line or a dividing band in the nonlinearity region. For a classification task with some step activation function, a single node will have a single line dividing the data points forming the patterns. Step activation was a discontinuous separation. More nodes can create more dividing lines, but those lines must somehow be combined to form more complex classifications. A second layer of perceptrons, or even linear nodes, are sufficient to solve many otherwise non-separable problems. The introduction of nonlinearity—a form of making a choice out of many—not very unlike Golden rule of quantum transitions, or of human judgment based on past experiences and instincts is the twist that is essential to the success in deep neural networks of today, where scale can handle the vast complexity.

The autoencoder was based on just linear transformations. A combination of linear affine transform followed by a nonlinear transform is a nonlinear encoder. Its usefulness is that it builds an approximate model of the statistics and in turn, therefore, can be generative. The autoencoder finessed—deleted unwanted noise in information such as of the Professor Subramaniam Sundar Kumar Iyer in picture, and perhaps some wanted one too—because it reduced dimensionality.

The flicker in the eye discussed and used in the last chapter was useful for quantification. The flicker introduced by the eye served to find a threshold for the presence or absence of a signal at the receptor. The process was reversed and a threshold was set that removed the flicker-like noise that was present in the data. In the affine linear transformation, the data is being suitably mapped and scaled for the thresholding nonlinear operation to cause a useful separation. This is the classification or the logical operation. *NAND* showed us both of these viewpoints. Non linearity provided a clever overcoming of the limitations of affine transformations.

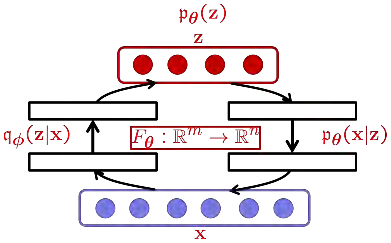

$\mathfrak{p}_{\boldsymbol{\theta}}(\mathbf{z})$
$\mathbf{z}$
$\mathfrak{q}_{\phi}(\mathbf{z}|\mathbf{x})$
$F_{\boldsymbol{\theta}} : \mathbb{R}^m \rightarrow \mathbb{R}^n$
$\mathfrak{p}_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$
$\mathbf{x}$

Figure 4.7: Variational autoencoder using neural network as a generative way for creating complex distributions by parameter estimation and minimizing with evidence lower bound.

This brings us up to the variational form of the autoencoder representationally shown in Figure 4.7. In the objectivist view, given repeated experiments—an enough-data approximation—one can build a model with parameters that describe the behavior around which fluctuations should be anticipated. This is saying that one now has a *sufficient statistic* in a set of parameters ($\boldsymbol{\theta} = \{\theta\}$) to describe the observable behavior around which statistical variability will exist, with the variability arising both in the errors of measurements and anything intrinsic to the phenomenon being modeled. The parameters are our abstractions to model a view of the world under observation. Multilayer perceptrons—deep neural networks—can be used as generative methods for creating complex distributions because of this parameterization. One wants to force the expectations on $\mathbf{x}$ to be what would be expected from $\mathbf{z}$ in our autoencoder example. This is what Figure 4.7 represents.

An example of a sufficient statistic of the normal Gaussian distribution is the mean μ and the standard deviation σ. $\boldsymbol{\theta} = (\mu, \sigma)$

One is creating a parameterized distribution ($\mathbf{z}$ of m dimension) to look close to what the distribution ($\mathbf{x}$) of n-dimensions is. Mathematically, this is an implicit function discovery of $F_{\boldsymbol{\theta}} : \mathbb{R}^m \rightarrow \mathbb{R}^n$ and inducing a complex distribution over $\mathbb{R}^n$ with parameters $\boldsymbol{\theta}$. One is sampling $\mathbf{x}$ by sampling $\mathbf{z}$ and setting $\mathbf{x} = F_{\boldsymbol{\theta}}(\mathbf{z})$ with the expectations $\mathbb{E}_{\mathbf{x}}[f(x)] = \mathbb{E}_{\mathbf{z}}[f(F_{\boldsymbol{\theta}}(\mathbf{z}))]$. For this to work, F needs to be invertible since

$$\mathfrak{p}_x(\mathbf{x}) = \left|\partial_{\mathbf{x}} F_{\boldsymbol{\theta}}^{-1}(\mathbf{x})\right| \mathfrak{p}_z(F_{\boldsymbol{\theta}}^{-1}(\mathbf{x})) \tag{4.12}$$

A network inversion to obtain pre-image ($\mathbf{z} \mapsto F_{\theta}(\mathbf{z})$ that is close to $\mathbf{x}$) is required. Computationally, for the neural network, this demands an inverse, so a Jacobian determinant, computing of gradients with respect to $\boldsymbol{\theta}$ that needs to be learned. If non-invertible, or dimensionally intractable, it may not be viable to construct the probability density.

The evidence lower bound (*ELBO*) allows one to bypass the deterministic computation of $F_{\boldsymbol{\theta}}$. Take the more general $\mathfrak{p}_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ from which the marginal likelihood is $\mathfrak{p}_{\boldsymbol{\theta}}(\mathbf{x}) - \int \mathfrak{p}_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\mathfrak{p}(\mathbf{z})d\mathbf{z}$. The variational lower bound relates as

$$\ln \mathfrak{p}_{\boldsymbol{\theta}}(\mathbf{x}) \quad \geq \quad ELBO(\boldsymbol{\phi}, \boldsymbol{\theta})$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathfrak{q}_{\boldsymbol{\phi}}}\left[\ln \mathfrak{p}_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) + \ln \frac{\mathfrak{p}(\mathbf{z})}{\mathfrak{q}_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right] \\
&= \mathbb{E}_{\mathfrak{q}_{\boldsymbol{\phi}}}\left[\ln \mathfrak{p}_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] - \mathscr{D}_{KL}(\mathfrak{q}_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\mathfrak{p}(\mathbf{z})), \qquad (4.13)
\end{aligned}
$$

where $\mathscr{D}_{KL}(\mathfrak{q}_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\mathfrak{p}(\mathbf{z}))$ is the Kullback-Leibler divergence. There are two models, one where one maximizes with respect to $\boldsymbol{\theta}$ for a generative model given by $\mathfrak{q}_{\boldsymbol{\phi}}$, and the other where one maximizes with respect to $\boldsymbol{\phi}$ for an inference model given by $\mathfrak{p}_{\boldsymbol{\theta}}$. The inference model is performing an approximate model inversion. Stochastic gradient descent can be used, and the method now is similar to supervised learning with input $\mathbf{z}$ and output $\mathbf{x}$. It has become a generative model.

The difference between these two autoencoders can be seen in their way to probability distribution. A linear autoencoder, given the data, is doing a dimensionality reduction on the fed data. It is implicitly working towards a conditional probability $\mathfrak{p}(\mathbf{z}|\mathbf{x})$. This is a discriminative process even if unsupervised. Recursive, that is, sequential neural networks, or convolutional networks, that is, networks that exploit adjacency, or simple neural networks all do this. Statistical learning approaches of regression and linear classification also do the same.

This generative capability is also the fake pictures and fake videos that we should all be very afraid of. The modern information society has many such powerful tools for disinformation, conspiracies, and polarization. It is not just WhatsApp. In the old days, even today, it used to be the spreading of something deeply offensive about a candidate a day or two before the election that could not be undone in time. Generative capability, however, is also very useful since an implicit model has been built. This model can then be used to not store the information that fits the model and only focus on what doesn't fit. Then, one has a way to do good science and not just collect a lot of repetitive data that doesn't inform anything new.

On the other hand if one were working implicitly towards $\mathfrak{p}(\mathbf{x})$ and joint probability $\mathfrak{p}(\mathbf{x}, \mathbf{z})$, one is attempting a generative process that is also unsupervised. Recall the example of the question that if one my children is a boy, what is the probability that the other is a boy. This is a turning into a higher dimensional vector. What we tackled by Bayesian graph can be tackled by a neural network.

Variational autoencoders do this, restricted Boltzmann machines do this, recursive neural networks can be formulated to do this, and so can generative adversarial networks.

These variational auto encoders is what are generally used in order to do all the picture transmogrification. They have a rather straightforward implementation and they can now be made generative with adversarial intents. The generative model now is using the the deep latent model features to generate. Generating new faces instead of reconstruction alone. The generator is being trained to generate samples close to being indistinguishable from real data. The posterior classifier can be used to train the generator by minimizing a logistic likelihood.

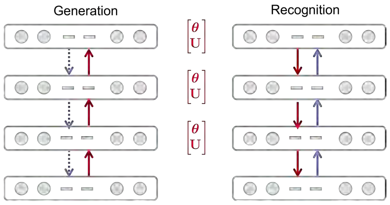


Figure 4.8: A layered generation-recognition neural network stack that can be based on variational autoencoder.

4.6 Neural networks

An early example of generative recurrent neural network is shown in Figure 4.8. With a Gaussian noise input,

$$\mathbf{z} = (\mathbf{z}^1, \ldots, \mathbf{z}^L) \text{ with } \mathfrak{q}_{\boldsymbol{\phi}} = \prod_{l=1}^{L} \mathcal{N}(\mathbf{z}^l | \theta^l(\mathbf{x}), \mathbf{C}(\mathbf{x})), \tag{4.14}$$

where $\mathbf{C}(\mathbf{x}) = \mathbf{U}(\mathbf{x})[\mathbf{U}(\mathbf{x})]^T$, one now has a variational autoencoder that learns and generates.

This is an example of a starting point for transformers whose further developments—large language models—now appear in the various generative neural networks including for natural language.

What we have accomplished is a transformation of a collection of data and its information into symbolic form, and by doing that, it is now possible to extract and generate. Neural networks are powerful since the complete process of $\mathfrak{p}(\mathbf{z}|\mathbf{x})$ that would be a large matrix and that scales polynomially and whose analytics scale exponentially—think Markov's hand calculations—is simplified. The implicit certainly is an exponentially scaled amount so one can not really do that with parameterized conditionals. This is where neural networks are very very useful. For example, in the earlier example one parameterized $F_{\boldsymbol{\theta}}$ into this probability function, can now be scaled.

This ability to generate based on a parameterized probabilistic model created is very very powerful. We use language, facial expressions, absence of response, pictures, equations, and so many other ways to communicate. They contain information that is being passed on for the brain to interpret in whatever form it interprets it in. All these are being transformed. Neural networks are doing that too.

The equation is a mathematical symbolic form for transforming what could be expressed in language. When the Hindu 0 (zero)—a place holder—and the Hindu numeral system (decimal) became the standard in the last thousand years, the ease of doing arithmetic became the foundation for algebraic development and later on to calculus and now neural networks, all steps in the process of tackling increasing complexity.

This idea of transformation from some expressive form to an implicit form is very powerful.

The neural networks provide the tool to fit a probabilistic description, limited by data and unknowns, of complex phenomena and the generative recurrent neural networks, of which the transformer is one example, are a modern tool.

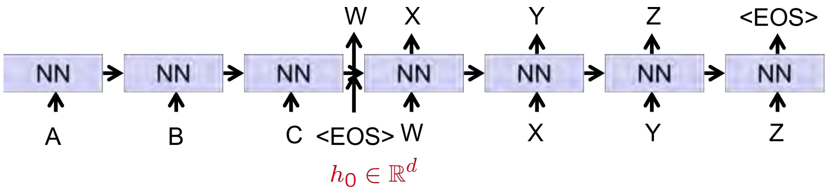


Figure 4.9: A sequenced encoder-decoder transformer generalizing the variational autoencoder.

The sequence-to-sequence learning[3] was an early successful example of he coder-decoder scheme using long short term memory networks to achieve a more transformative form. This is shown in

[3] Sutskever, Vinyals and Le, (2014)

Figure 4.9. Encode a sequence, for example, a sentence into a vector and then decode the sequence, for example, translate, from the vector using an autoregressive output feedback. In the figure, A, B, C, et cetera, are inputs to an encoder sequence—not unlike the Markov discussion but more complex in being symbolized words or even more complex forms for pictures or music score. There is also the hidden learned input. Then when one inputs something else W, X, Y and Z in a sequence, one gets X, Y, Z as a sequenced output. Language translation is an immediate example.

This is what the network did. In the conditional, and probabilistic parameterized model fitting, one has

$$\mathfrak{p}(\mathbf{y}|\mathbf{x}) = \mathfrak{p}(y_1, y_2, \ldots, y_N|\mathbf{x}) = \mathfrak{p}(y_1|\mathbf{x})\mathfrak{p}(y_2|\mathbf{x}, y_1)\ldots\mathfrak{p}(y_N|\mathbf{x}, y_1, \ldots, y_{N-1}) \tag{4.15}$$

as a rigorous interpretation. In analytic thinking, this is an exponential amount of memory in some truth table form where $y_1, \ldots, y_N$, et cetera, are characters, words, tokens, or whatever symbolic form one has chosen. In a language model, it is the words as some encapsulation of characters. The large memory collapses if one parameterizes the conditionals in a large neural network, that is, one maps $\mathfrak{p}(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) \mapsto \mathfrak{p}_\theta(y_i|\mathbf{x}, y_1, \ldots, y_{i-1})$ into a parameterized probabilistic model. Creating one word at a time to create sentences having sampled conditional is $\mathfrak{p}_\theta(y_i|\mathbf{x}, y_1, \ldots, y_{i-1})$, which is a reasonable continuation. It is autoregressive.

The dark side of the natural language tools, large language models, that are receiving so much of the attention these days is teachers' testing of students. What does the student know and how the student thinks versus what is the business-like cut-and-paste from the model's archive. The language model has imbibed the web pages say, all the information we all store in the cloud, email in the cloud, and that institutions now force us to do since the immediate cost is zero or low. This is Google, Facebook, Microsoft, Adobe, Apple, and others, choose your favorite now. This is a another Kabuki persona of autoregressive.

In the next essay, I am going to call this action-reward-behavior and its real-world manifestation as the *two marshmallows principle*.

Reinforcement learning attempts to train a network to produce actions based on rare rewards. In this process, there is no action instruction being provided based on some loss function. The important result is that we are not providing a model of the environment. The program leans. It is also open system since the model space is enormous, a final reward to define a cost function does not exit since the environment too reacts to proposed changes. So it a probabilistic action choice, where learning is that if the reward is high, it increases the likelihood of that sequence. While this may reinforce some poor actions too, but they have low reward trajectories, so the net suppresses such action.

This flows into the long short term memory approach conceptually shown in Figure 4.10. The long part of the memory are the weights adapted during training and are being stored for perpetual use. The short part of the memory is the input-dependent memory. So the architecture maintains long memory times in a robust way for the short term. It can also be made to forget as also write a new value which is passed along. This is now at least four important transformations involving updating and response and long and short. A set

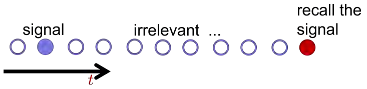


Figure 4.10: A schematic view of the long short term memory. Selective memory is being maintained for long term, is allowed to be modified, and is available for use with short-term memory to produce responses. This is unfolding of recursiveness.

of equations describing this for neural implementation is

$$\begin{aligned} z_t &= \sigma(w_z \cdot [h_{t-1}, x_t] + b_z) \\ r_t &= \sigma(w_r \cdot [h_{t-1}, x_t] + b_r) \\ \tilde{h}_t &= \tanh(w_h \cdot [r_t \odot h_{t-1}, x_t] + b_r) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \end{aligned} \qquad (4.16)$$

turned into a block representation shown in Figure 4.11 of the encoders and the decoders. The block shows inputs, long memory being passed with neural adjusting and a response with the short memory based input. A long and short implementation that I have emphasized earlier as an essential characteristic of physical and natural world. The approach needs gated units instead of one because memory is not being preserved as in a recurrent network causing a conflict with short-term fluctuations. The standard recurrent network therefore is unfolded into gated units.

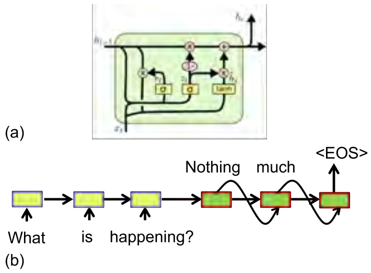


Figure 4.11: (a) shows a block picture of the neural implementation with plenty of the transformations that are needed to keep and modify long term memory, while also acting in the short term based on long-short response. Selective memory is being maintained for long term, is allowed to be modified, and is available for use with short-term memory to produce responses as shown in (b). An incoming sequence feeds into the recursive arrangement.

We have ended up with an understanding of neural networks as a model-building exercise that, given lots of data of a complex system, can show us an approximate model—a multidimensional multivariable fit—and because we have a ``model'', tells us given some input, what should be expected as a result. It quasistatic in that there exists a model that for a given instance tells the input to output mapping, but that is only dynamic quasistatically in that given a sequence of inputs it can give a generative sequence. The model can also be made to be a learning model that can change as more time and input goes by.

Because there is an approximate model, given input, it is also possible to extract what doesn't fit. This to me is one of the really interesting uses from a science perspective. It gives us insight into the surprises that a model hasn't seen. This is what science has used as a launching pad to new discoveries and understanding.

4.7 *The physical–informational world and neural networks*

The physical world and how we model, develop new ideas when observations don't conform to the model, is not that far from this insight into generative neural networks. This is the reason why this technique has a power for tasks that have been beyond our reach until now. The other is that it can explore large and complex data with multiple causal and random factors.

Most modern tasks that are of interest involve complexity. Phase transition, renormalization, singularities, et cetera, are all examples of complexities of interaction across scales where the happening is nearly simultaneous across structures

I view this as a conundrum that has partly come from the conflict that exists between analyticity, mathematics traditionally prior to to probabilities and in the abstract notion that pure mathematics teaches us to completeness where a conclusion is guaranteed with that of

The effects are in time and space. In *USA*, if a snow storm comes to midwest—Chicago!—why do flights between Florida and California get jammed? They do since both position and flow is at play. Planes going in between Florida and California have a connection to Chicago. Maybe it was arriving from there, maybe it was going to go there after the hop, et cetera, and the network has a whole lot of such nodes. Take away a node, and the system can collapse. Like the removal of a pin in those magical wooden puzzles that are complex to assemble, but whose stability depends on one critical piece that the least thinking person in the gathering, even if he understands the consequences, pulls out. The reverse is also true. A broken network may begin to flow with just one connection made. Percolation has scaling laws.

open boundary conditions of reality. Thermodynamics, for example with its Carnot cycle or Boltzmann's H theorem, or quantum mechanics with its square-well like problems and others far far more sophisticated, all employ thoroughly defined boundary conditions. Reality is open, we don't know all, we don't know even what all is, there are interactions of all different kinds with an environment that is not immutable. Even the universe is expanding.

Complex problems exist arise in openness, large dimensionality, large number of interactions, spontaneous events, et cetera, and tackling them in probabilistic terms, that is, getting to a good-enough or more likely answer, is to give up the guarantees

Giving up the guarantees is the central issue since it is a conflict between what we do not want to give up, and we can never know what we gave up if one adopts a system that cannot guarantee. It is a delicious problem that Silicon valley has exploited, just as economics has and politics exploits it.

When I was young, Fast Fourier transform had just been invented. It was a very big deal for us since so many problems could be better tackled, with guarantees, in reciprocal space. Finding fast algorithms for multiplication ($\mathcal{O}(n^2)$ to $\mathcal{O}(n^{1.59})$ or $\mathcal{O}(4.7 \times n^{2.91})$), extended to matrix multiplication ($\mathcal{O}(n^2 \times (2n-1))$ or $\mathcal{O}(4.7 \times n^{2.81})$) , and so on, so many others, with deterministic background, have been important to progress in my lifetime. The $P = NP$ problem that has fascinated computer scientists for a long time reflects one aspect of this issue. Prime numbers, with so many different connections to Mock functions or Reimann conjecture, or others, and the MergeSort on which a trillion-dollar business has been built is of $\mathcal{O}(n \times \log n)$ complexity. These are all important, and good-enough, and non-guaranteed works most of the time. This is fine for much of the reality and the open system we dwell in. *AI/ML* is the tool to this end of the probabilistic non-deterministic spectrum, but one which can also span to deterministic end. Extremely computationally intensive tasks, tasks that have in the past waited for progress in supercomputing, which is now in exascale, become fathomable with smaller scale systems.

Prime number is currently considered of $\mathcal{O}(\log^{12}(n))$ complexity.

How does one actually address this complexity is the question I would like to now address while keeping the scientific constraints that we know, and continue on this path of not treating the entire enterprise as a black box.

Here is one interpretation. Words and language can be associated within our mind lexically and semantically. In the lexicon of the natural language, the atomic unit that gets transformed to meaning is the symbol such as the character of the alphabet or a number. A collection of these form the words or phrases. Sentences, phrases too as well as words, are built under certain rules, coded into spellings and the grammar. As a newborn, the number sense thesis[4], the hypothesis is that we arrive with an operating system that comprehends 1 or 2, but not much more. By the age of two, we comprehend even 4. But 1 or 4 by itself has no meaning. It is just a symbol. There are 4 boys, or 4 girls. or 4 apples, or 4 chairs, et cetera. There is no object to be

[4] S. Dehaene. *The Number Sense,* ISBN 0-19-511004-8, Oxford (1997)

called 4, it is a symbol of a count of objects. The meaning of a word arrives with its use in the language as Wittgenstein says. With this, one a unit representation and the other a meaning, one has arrived at something far more in the form of a semantic. One now has a way of conveying information and knowledge. With the symbol and the symbol for the object and a collection of words use in the corpus, word representations have come about that capture the word meaning. We have gradually assembled a path to adding and building an algorithm for it. Just like sets and groups that immediately come to mind when we think of 4 boys or 4 girls or 4 apples, or 4 chairs, and their manipulation, language to has arisen in a similar way, A sentence is an algorithmic representation—built by rules of grammar—to convey information just as the adding did or sets and groups do. This is our number sense and either tied to it, or independently, the Chomsky view of developing language skills. The meaning of the words have appeared through the use of the language.

I like to think that this description is not that unlike *the neural network development.* We have used symbolic representations, letters and numbers in binary form, for example, formed vectors, to denote a more complex assembly like a word, accumulated their relationship similar to what algebra does, or equations, and calculus does, and what grammar does for getting the relationship to have a meaning. The long short term memory, the language models, see this relationship. *Just as we don't really fully understand what operational form and mechanism is being employed to represent and manipulate in the brain, that is, what the specifics are, the same holds true for neural networks, though in some ways, we understand neural networks a bit more. We can probe what is going on. But, why this way and its predictability, and certainly how to make connections across domains is a vast open question..*

Mathematics in this sense is a language of description with its own symbolic notation, just as English is or Sanskrit is, but also that the meaning of what Professor Sundar Iyer is saying in a language may be slightly different than mine even if we both convey it in what we call English and may even use the same words. For that, one has to hear the inflection, the slant of the eye, the little bending at end edges of the lips. I may use the word to mean something and it may not be identical even if we employ the same lexicon although between Professor Iyer and me, our $\mathfrak{p}$ and $\mathfrak{q}$ probability functions of conveying may have very low Kullback-Leibler divergence.

The implication of these examples of usage of words is that the corpus of learned word representations that capture the word meaning are being assembled from symbols. Symbols are embedded in the vector space for the basic representation. The vector spaces' space structure, such as angles or distances or something else in that multidimensional space is going to relate to the meaning of the word if you interpret it this way. Now one should be able to see that this approach should work broadly. Condensed matter physics problems, the difficult many-body problems or of topology, the problems of statistical mechanics, of economics, of the many of the agent problems (not that different from many body problems), are now interpretable, transformable, and the approach can be seen to start to work broadly. Language models are associating implicitly a meaning and the rules of association are being learned during any training.

I am a card-carrying phenomenologist who worships at the altar of Husserl and Heidegger. The interpretation that I have raised is that the meaning comes from the usage and is part of the culture. In sciences, one often knows the bounds of what we should or should not do, the limits and limitations of the model, and we are alway willing to change given contradictory evidence. This is phenomenology. Let observations be an important input to our description.

All this discussion applies to the variety of domains that the brain is so adept at. So, it also applies to music and the arts and across all the activities of humans

The power of *ML* is that with this symbol-word-corpus-meaning interpretation, one can see that similar model structure should be applicable to defining, describing, and ultimately interpreting and creating whether it is in sciences or in arts. Large language models may be useful to tackle condensed matter problems.

An early illustration of this power was an attempt at building a model within a model—like we do through our scenario creation within our brain—to understand the grammar and meaning of music. It is possible to make the music audio into a symbol-based word and essay-like description. Notes, the speed with which the notes should be outputted in a sound, all the different channels, the musical instruments, the timing, et cetera, can all be encoded through an autoencoder, which is the *word2vec* vectorial representation, fed into a recursive network that makes the scheme temporal and the assembly can be classified through a neural network as shown in Figure 4.12.

There is something delicious in the language to mathematics jump. Indian mathematicians ruled supreme pre-10th century *AD*. Much of their description, of geometry and algebra, of zero, of adding, subtracting, squaring, negative numbers, et cetera, is through vernacular language. We had to wait for Leibniz for finding the calculus form of the infinitesimal to make the next jump. We may very well be moving towards language model again for it may be a better descriptor of complexity and openness.

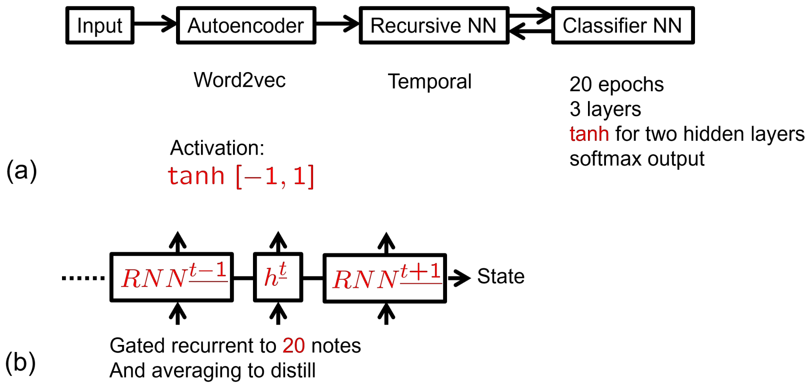


Figure 4.12: The transformation of music through an autoencoder and recursive neural network for classification is shown in (a). (b) shows a gated recurrent form to build multi-note assembly, where averaging over multiple data provides an increasingly accurate description.

The principle that underlies this music construction is that words or characters or notes of instruments are degrees of freedom of the system in a computational basis which is in some state space. So long as one can describe a state and its change, that is, our position and momentum of classical sciences, then we have a description, and this recurrent neural network formulation has them. Now one can infer reconstruction of the state consistent with the data, the measurement outcomes, and make it predictive. Our attempt at this said to us that Baroque Bach was classifiable and one could even generate pseudo Bach, but Beethoven or Mozart not. Of course, this conclusion depends on the implementation as well as how much data is available to feed in

With this preliminary, I hope a reasonable argument has been made that ML and neural networks are a powerful new technique now available to us for tackling previously intractable problem. They stand up to physical arguments and they align with the human and natural way one tackles information and solves the problems to the extent our current state of knowledge.

4.8 *Physical-science-mathematical correspondences in neural networks*

WE CAN NOW PROCEED to see the correspondences between the seeming black box like neural network that we are starting to understand and how it projects to complex and science-related problems. It is useful to see at least some of the broader correspondences, of which Figure 4.13 is a simplistic description. The process of an input to output through the activation and propagation process, where both the affine transformation step—the matrix multiplication—and the nonlinear thresholding occurs is both a position and momentum change. Input gradient is going to project to output gradient. The affine transformation is a slow and small change, the nonlinear transform will cause a fast or large change. In the neural networks, in the process of learning through the back propagation and large-scale data usage, that is, a repeat and repeat, one is fine tuning. By placing sparsity in weights and activations with a random selection, we are not overfitting, and just like the real world, also implying that there is much unknown, whether it is an inaccuracy of measurement or of incompleteness of what is known. In the convolution networks, one is practicing a similar behavior by local convolution, which is short-range, and then following subsampling and more convolutions, also incorporating long-range changes. Subsampling is again bringing stochasticistiy. Stochasticity, whether in the networks of type (a) or type (b), of weights and sparsity, is an accounting of incompleteness, which is entropy of not knowing. The entire state description of what exists in the network is a description of microstates and macrostates.

Our suspicion was that Bach repertoire is more limited, and enough available, to classify. Beethoven and Mozart ventured far and wide. Perhaps we should repeat this experiment with Vivaldi, like Bach from Baroque era, who Stravinsky said wrote the same concerto 500 times.

The connection between probabilities, space, and time in complex problems of cause and chance should be in the back of your mind by now. Probabilities—chance—is not that easy to fathom and get a good feel for. Neural networks, by just putting it all analytically, are hiding this complexity making them difficult to understand. I illustrate with two physical questions. One in form of a question, ˵John, an American, is a very thoughtful person, he reasons, he looks at resources to see what dictionaries and encyclopedias and other resources have to say, and then he responds. Is John more likely to be a librarian or a farmer?˝ If this is all you knew, guessing that John is more likely to be a farmer despite all these attributes is one of higher probability. There are far more farmers in *USA* than librarians. A deeper one is ˵How did life appear on earth?˵ To this, in my childhood the answer was Urey's answer that ammonia-water-hydrogen-methane mixture with electric sparks of lightning created life since protein precursors were found, or more recently that life came from thermal vents in the water where conditions were right, or a genetics answer that it was the messenger *RNA* and the coding of how ribosomes should do protein synthesis as a mutating evolution on the planet. My view is Saganeque. There is a huge universe. Reproducing entities—spores, lichens, a whole variety of bacteria, bacteriophages, and even animals such as nematodes and tardigrades, et cetera—can survive in space and across temperatures. This is a giant farm with farmers out there in space compared to the librarians here on the earth. *Life got transported.* What I do agree with is Fermi's view that we are not going to see another living species whose message we receive because of the consequences of the speed of light and how long species survive. We are showing all evidence of this propensity to self destruct in so many ways through our selfishness.

When we do forward and backward propagation we are practicing the dynamics of how we fine tune our understanding of the world. It is learning.

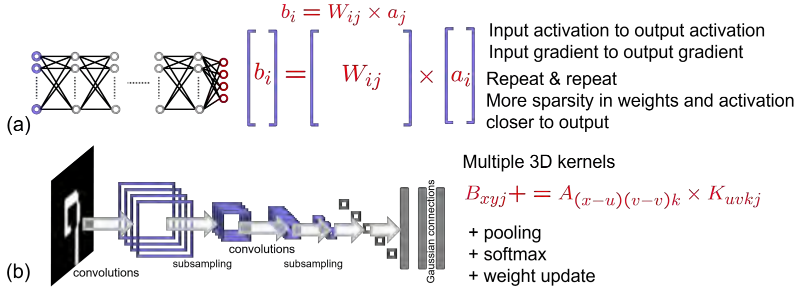

$b_i = W_{ij} \times a_j$
(a)
b_i = W_{ij} × a_i
Input activation to output activation
Input gradient to output gradient
Repeat & repeat
More sparsity in weights and activation
closer to output
(b)
convolutions
subsampling
convolutions
subsampling
Gaussian connections
Multiple 3D kernels
$B_{xyj} + = A_{(x-u)(v-v)k} \times K_{uvkj}$
+ pooling
+ softmax
+ weight update

Figure 4.13: (a) shows a simplistic description of a deep neural network and (b) shows a convolutional neural network.

Linear transformation is a ``slow´´ change, a reversible exchange, where no entropy changes taking place. Stochastic pooling, weights and sparsity is where one loses information content and is also agglomerating information. This is where entropy entered. Microstates and macrostates are within this representation, nonlinearity is fast change that's like Fermi golden rule, where the probability of going back is miniscule because one has made such a fast change that one has gone from one possibility to a large number of possibilities. Fast change forward and backward is therefore the dynamics. The real world description and the neural network description can be couched in similar language and description. There are significant correspondences.

Depending on what or how much information one has tied together—how restricted it is—is going to determine how well the system will actually work. This is just like the brain getting fogged up by the shadow that was there on that circle. The same is true for everything else in these things so if one wants to do inferencing properly one has to think in terms of how one is going to transform those words and characters from a physical science perspective. This is the constraining of the freedom in the degrees of freedom in the system to impart it with meaning. This is similar to the tackling of the degrees of freedom even in the Markov chains of consonants and vowels, or more extended chains, and these can be applied to Euclidean space, to spins, to phase space, to Hilbert space, to machine space, or any other space. Usually we have both the present state and the flow of the state so the momentum associated with this state in this phase space notation. They are all also intrinsic to neural networks so it really works with all that we have to do when infering a reconstruction of the state that is consistent with the data based on those measurement outcomes. This is no different than what the language model is doing.

Look at the neural networks as web of connections—graphs, but also like a fishing net—with inputs at one end and outputs at the other end. Flexible structures are more forgiving and less brittle with their flex passing movement and rates of movement under force through the links of the fish net. This is one advantage of non-deterministic networks versus deterministic traditional computation. While this network structure may be a bit different between the traditional deep network compared to the convolutions-based network, there are interactions taking place across layers and perhaps within layers depending on the net construction, but by optimizing, that is, forward-and-backward propagation based cost function minimization, one is using the output end to pull or relax, and the consequences channel down along the network finding its more optimum minimum condition. The network is adapting to its ground state. This is a minimum that we may call of energy, where the energy as a metaphor is some encapsulation of the information-related statistical and thermodynamic description.

``Define energy˝ is a question entirely nontrivial to answer. We have units, the energy comes in forms—potential, kinetic, chemical, nuclear, bond, and so on with confusions galore. But, it is a term for describing something that is exchanged and that lets us figure things out. There are many such immutables, exchanging form, but in total staying constant in our description of the world.

This brings us to a classical physical picture to a computational analogy noted in Table 4.3. The analytic description and the computational description and attribute are all related to the interactions that occur at short and long range, which either form of the network described (deep or convolutional) effectively captures. If one thinks in physics terms, the Hamiltonian is like the surprise, which is the minus logarithm of the probability since information is additive and probabilities multiplicative. Not being in the ground state ($\mathfrak{p} = 1$) is the surprise. The Hamiltonian being a second power of the canonical conjugate is the Gaussian probability. Local interaction is sparse. Symmetries—translational or rotational or others—are all implementable through convolution. Making determination of parameters from the Hamiltonian is the use of nonlinearity and optimization, that is, softmax or other nonlinear functions, the gradient descent and backpropagation. The free energy, energy exchange, and the differences that drive the system are similar to the attempt at matching by maximizing the Fisher information and minimizing the Kullback-Leibler divergences. This is the bringing together of the distribution functions. Operators and features correspond. Extracting a feature in the network is the analog of the operator operating on the state function to give an eigenvalue and returning the eigenfunction. Identifying an object such as leaf in a picture is associated with an operation in the network that finds the attractor state identifying that leaf. Leaf corresponds to the eigenfunction and the different objects in the picture are different eigenfunctions composing the state function. So one can see that the languages may be different but they actually relate to each other. Looking at these informational approaches in two different languages is actually very very useful since it gives meaning and insight. For example Shannon tells us that average information content of the a summation $\sum \mathfrak{p}_i \log_2 \mathfrak{p}_i$ is one of summing up the different surprisals and how often they are likely to occur.

Table 4.3: Analogy between physical analtyics and the neural network computation.

Physics	Computational
Hamiltonian	$-\ln \mathfrak{p}$ (Surprisal)
Hamiltonian in 2nd power of canonical conjugate	$\mathfrak{p}$ (Gaussian)
Local interaction	Sparse
Translational symmetry	Convolutional
Extracting from Hamiltonan	Softmax, gradient descent, backpropagation
Free energy difference	Fisher information, Kullback-Leibler divergence
Operator for observable	Feature

Quantum, statistical and information mechanics and their play through neural networks shows all these different correspondences. The working of neural network is a mechanism that is optimizing the information representation through different transformations. It is achieving that through the forward and backward propagation given the constraints that have been placed on the network, The number of hidden layers, the connectivity, the nonlinearity, the stochastic optimization, the loss functions, et cetera, all determine how well it approximately represents the physical situation.

Information underlies this representation. Entropy as a measure of information comes in many forms because it represents what is not known, and there are many information-containing properties including those that we don't even know exist since they are among the unknown.

Shannon's channel-based viewed of information, for example, of bit stream is a negentropy of $H_S(X) = -\sum_i \mathfrak{p}_i \log_2 \mathfrak{p}_i$ for averaged information content. If all bits are known, this is 0. If one bit is not known, for example, say in $\{00000?00000\}$, then its probability for random 0 and 1 is 1 b. For this same stream, the Fisher information is

$$\begin{aligned} I(\theta) &= \int [\partial_\theta \mathfrak{p}(x_i|\theta)]^2 b \mathfrak{p}(x_i|\theta) dx_i \\ &= \sum \frac{1}{\mathfrak{p}} \log_2 \mathfrak{p}^2 \Delta x_i \\ &= 2 \ b. \end{aligned} \tag{4.17}$$

Why 2 and not 1? Fisher is informational view of position and momentum. Shannon's of a bit by itself. Fisher's is of a bit in context of other bits because of that variation with the parameterization of the distribution. Fisher information is particularly useful to view neural networks in physical terms therefore.

The one complexity with neural networks is of reasonably representing the degrees of freedom, which are immense. This is a dimensionality problem. It is why neural networks have to be so immense to be representational of the degrees of freedom of a complex system. As Figure 4.14 argues, the sampling has to increase as 2^d to adequately capture the terroir of the physical system being modeled.

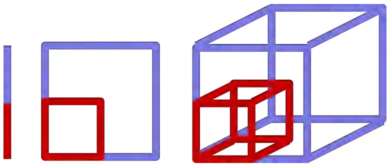


Figure 4.14: The 2^d scaling in sampling needed to adequately represent d degrees of freedom. For example, for up and down some parameter that is a degree of freedom, a volume with $1/2$ its length tells us which part it is in. Go to a square, now the sampling volume decreases to $1/2^2$, to $1/2^3$ for a cube. This is 1, 2 and 3 degrees of freedom progressively reducing the sample volume size. The sampling needs to increase as 2^d to adequately sample a d dimensional space. Correspondingly, the convergence rate will vary as $1/\sqrt{N}$ corresponding to the Gaussian spread, to $(\ln N)^d/N$.

Fortunately it turns out that utilizing randomness helps—just like flicker noise helps with the eye—to figure things out. The second complexity is related to information aggregation versus partitioning.

If one looks at the mutual information in an aggregated collection one can write it in terms of an expanded Taylor series form as

$$\begin{aligned} I(\mathbf{x}|y_k) &= H(\mathbf{x}) - H(\mathbf{x}|y_k) \\ &= -\sum_{i=1}^{k} \frac{\Delta H(\mathbf{x})}{\Delta y_i} - \sum_{i>j=1}^{k} \frac{\Delta^2 H(\mathbf{x})}{\Delta y_i \Delta y_j} - \cdots \\ \therefore\ \ I(\mathbf{x}|\{y\}_k) &> \sum_{i=1}^{k} I(\mathbf{x}|y_i). \end{aligned} \tag{4.18}$$

This is the mathematical basis of Galton's estimate. Aggregation of independent information—with pieces conditionally independent—has equal or more information. Sampling helps. More than the self information of one measurement is obtained through the collective measuring. As an example, information gain about $\mathbf{x}$ from a pair (y_1 and y_2) is the sum of independent mutual information and an additional term. This is the correlation between y_1 and y_2 . The equation is also telling us that with more measurements y_i, there are now higher order terms. *These are the short and long range correlations.*

The digital gates *NAND*s, *NOR*s, *XOR*s are aggregators. They are nonlinear. If one has them with multiple inputs and multiple correlations, we have a neural network. Viewed this way, neural networks are a generalization over correlations that have now become feature extractors. The nonlinearity in multiple inputs, in presence of multiple interactions across the inputs that may look like multi-variable correlations that are nonlinear and are viewable as a Taylor expansion, are captured in a generalized way by stochastic assembling of affine transformations and nonlinear transformation with hard or soft thresholds.

By giving up on guarantees, completeness and at least currently—a very superficial understanding of the black box that are the neural networks—we have an approximate guess methodology for complex problems. This is essence of the transformation from the past to the future and the incompleteness therein. Neural networks are just letting us to do the same for what may very well still be the past using the same methodology.

The basis of all this is in stochasticity (errors and surprises) , linearity (slow) and nonlinearities (fast), renormalization-like coupling across scales, and an acceptance that past (data fed in) is a good representation with once-in-blue-moon rare events not part of the schema. In this the neural networks manage to

- Dimensionality reduction captures approximation of information critical to inference.
- The exchange from degenerate states is nonlinear and statistical.
- Noise is unknown information. Take away a data, the collection is noisier. This makes noisy information useful as the stochastic resonance showed.
- Correlations are exchange. Higher moments are longer-range exchanges.
- Noise helps by emphasizing correlations. So do hidden nodes where convolutions happen.
- Correlations are also a measure of order. So is mutual information.
- Adaptation accounts for incompleteness of information.
- The probabilistic representation translates in this interaction
  - Nonlinearities and phase transitions.
  - The natural world as a play of chance and causality with the order appearing because of the nonlinearity.

The dimensionality reduction—within limits—captures the approximation of information critical to the inference that we saw in extracting Professor Iyer from the sea of Gaussian noise, and exchanges of a degenerate ensemble since the process is nonlinear, statistical, and the neural network is a particularly adept tool at being more representative of probabilities by working from previous data. *Nonlinearity have helped partially overcome errors of measurement and noise and statistical distribution of nature.*

Simultaneously, noisy information has been useful as a stochastic input that, with avoidance of overfitting, can exploit stochastic resonance. The stochastic resonance of the previous essay exploited correlations. The higher moments correlations arising in higher moments are long-range exchanges. Noise helps by emphasizing correlation, so do hidden nodes where convolutions happen. This is why the hidden network's hidden layers are so important.

Correlations are also a measure of order so they also map to mutual information. Adaptation accounts for incompleteness of information. It is adaptation that nonlinearites bring in the ability to overcome what is not known of higher moments. Phase transitions have this behavior in the natural world, and the neural networks have the ability to capture these renormalization capabilities.

We will look at some representative examples in order to fit and show the correspondence in the mathematical and computer science construct within the natural context.

4.9 *Information and physical guiding in neural networks*

Since all these themes end up in a distribution of possibilities represented by probabilities, it behooves us to make sure we understand how probabilities constrain and what is being used towards the inference in the optimization by the neural networks.

The loss-related criteria can be multitudinous. One may want to work with sampling of one distribution over the other, or Fisher matrices, or place physical constraints, or so many others more. Our interest here is to argue that neural networks under statistical constraints should also be beholden to physical constraints whose laws we, so far, know hold true, and believe to keep holding.

The first of these is an early inculcation of the maximization of entropy. Thermodynamics claims that this is the end—a way station in the Feynman view—of a closed system. Neural networks are not necessarily, nor is nature as far as I can tell, a closed system.

We also grow up thinking minimization of energy, which is certainly a noble objective, but we know that that this is a slow state-to-state change prescription with connotations of Zeno's paradox.

Herein lies an important problem. Conservative forces are not necessarily the norm. The world may not be an open system. A confined system with boundaries too has lack of information, that is, an entropy, and so much that is not being tracked, so minimum energy may be a ground state but is not necessarily the state that a system will find itself in under stimulation.

The maximum likelihood proposition just says that if one has a distribution $\mathfrak{p}$ and a model $\mathfrak{q}$ distribution, what we are attempting to do is to bring them together under some loss-related criteria. Maximum likelihood therefore is quite closely related to the peak of the distribution and bringing them together.

The maximum entropy proposition asks one to start with the most consistent assumption of what the distribution is going to be like without assuming anything except what has been given. Quite often this is very difficult to interpret but this is what maximum entropy truly means.

I remember an incident from my graduate student time. After a long and difficult research experiment, two data points had been generated. The speaker put those data points up, and since it was an energy of some parameterization plot, drew a straight line on a plot through those points and an imagined point at absolute zero. The question immediately came up, why a straight line, and why the point at absolute zero. Of course, it is possible that there are phase transitions, that is, energy-exchange processes, and different activations. Models are just models. Useful, but wrong once one starts to generalize from a limited region of validity.

Minimum energy is often how we pursue solutions in physical sciences to find the most probable, and in turn, the final end stable state. Minimum energy state recovers to itself under disturbances since the second gradient in energy vanishes for conservative forces. The first gradient is a force, at least for conservative forces, that brings the system back. All the statistical elaboration with partition functions is mostly related to working with and inferring from minimum energy.

Within these different methods is the conflict of over-fitting. If one over fits, one has poor generalization. This should be obvious

I hope Also, I hope don't have to explain that there are conflicts of priors that can take place also. The path to a state too matters in a real-world system.

There exist two issues in here. How does one organize the methodology to extract maximum information in the presence of fluctuations and noise and minimize failures?

Energy, entropy and inference have a multifaceted relationship that shows up in so many different ways from phase transitions, in reversible processes, in irreversible processes, or in the variety of mechanical or information engines. Figure 4.15 shows a summary of the probability behavior in a 3-layer binary classification of the $CIFAR - 10$ image subset.

Depending on how much and how one adds noise in the gradient descent shows the nature of not knowing in inferencing. Figure 4.15 is an encapsulation of estimation in $CIFAR - 10$ image subset with added noise experimentation. White noise is poor, adding no noise is slightly better at increasing the probability of classification, and stochastic approach sharper with a higher informational entropy. The stochasticity aids in avoiding trapping in local minima, as also in following a path that is going to look out for a minimum over a larger generalized coordinate. Figure 4.15(a) shows the probability-entropy relationship, and Figure 4.15(b) an interpretation of why stochasticity—a Guassian fluctuation intervention—is helpful.

A very productive test creation of the modern times is of setting up up a problem that different algorithms and methodologies can attack and compared through different measures. Image recognition, charting a course through autonomous driving, solving some important problem, et cetera, are not unlike Hilbert's 23 problems, or millennium prize problems. They provide a common equally-known framework for all to attack. The current autonomous driving theme, or of drones as a derivative from that, are direct descendent's of the $DARPA$ autonomous ground challenge of 2004. Many such tests exist for checking the efficacy and speed of algorithms with respect to each other for pattern recognition tasks.

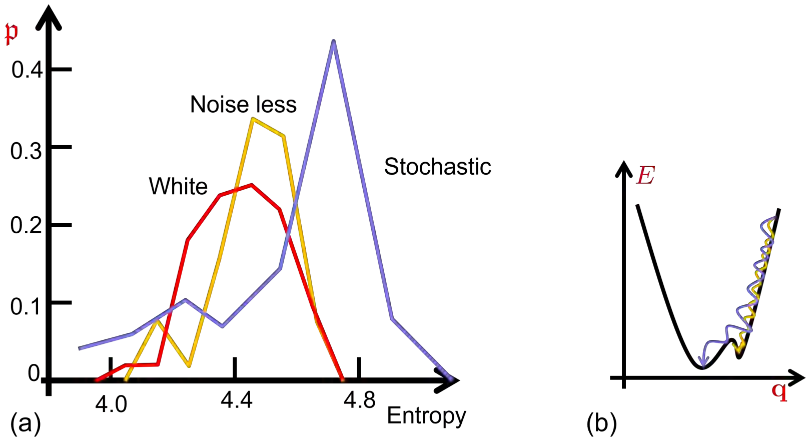


Figure 4.15: Representational summary of a 3-layer neural network classifying the $CIFAR - 10$ image subset. (a) shows the probability versus entropy in the classification, and (b) shows how in a generalized coordinate picture the stochasticity helps in the descent and finding the minimum avoid local minima and favors global minimum.

Nonlinearities are a means to compression. Linear transformations are a different representation for the same information, unitary transforms, rotation, and so on. In a deep network there is a fitting and compression taking place when one works through the network. The tanh function with its strong nonlinearity concentrated around the origin, means that when passing through one layer the data is undergoing a fitting and a compression. Stochastic descent emphasizes broader maxima by the double-sided nonlinearity. Noise brings out a wider minima and gets one to the end calculating much faster.

How does the information evolution take place in the network? One interesting way of looking at that evolution is to look at how that nonlinear function is being being put in: the nonlinear function as a tanhc or a *ReLU* (meaning stay zero and then evolve with a first-derivative discontinuity) or a softmax. The hyperbolic is a smoother function around 0 so it actually has a bidirectionality associated with it and one can see that if one uses a tanh hyperbolic function one gets the information content staying and evolving evenly distributed across the layers. Not so with *ReLU*. *ReLU* works best in the final layer where the collapse to classification is desired.

In one looks at the self-information content, agglomeration takes many iterations before building up, so clearly having it as double sided is more useful.

		100 nodes	784 nodes
	7	98%	62%
	9	1.5%	20%
	7	30%	9%
	9	67%	67%
	7	54%	59%
	9	28%	18%

Table 4.4: 7 and 9 classification to assess disambiguation using 100 and 784 nodes. The table is in the same order as the figure.

Given that we are discussing information in provided data, and even as a human we extract different informative extracts from what we observe, it is interesting to look at neural networks as a device for what it extracts—its classification—in response to created situations and also to see what goes inside. This may be instructive to understanding how information is maintained within its structures, the agglomeration, the decoding of mutual information, and its way of disambiguating. Surely noise—or really, fluctuations, or existence of different shifts around observations, et cetera—plays a role, helpful as well as what we learn in physical sciences, mostly unhelpful.

For humans, given sufficient signal-to-noise, this is ambiguity where we look at adjacent relationships to disambiguate. There are also circumstances where we just cannot do it. We will look at a few of these to see the connectivity and noise/stochasticity connection.

Table 4.4 shows the first of a contrived example to explore this condition in a neural network using the *MNIST* data. Increased ambiguity was introduced in a contrived way by adding rectangle in the 28 × 28 pixel array and the results summarized in the ambiguity resolution between 7 and 9. Two different number of hidden nodes are used in a 3-layer network, that is, a simple network. A human is likely to interpret these mostly as 9, but the network with one hidden layer does a pretty poor job. The last case, perhaps the least troublesome of the three, is where 784 hidden nodes gives an 18% result. Looked from afar, by and large, these are all 9s, but not so to the neural network. Looked from up close, these will give the humans the ambivalence similar to that of the neural network.

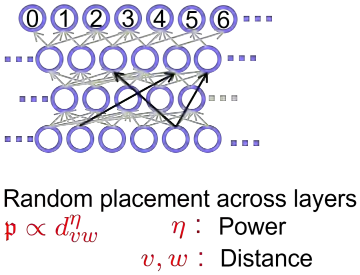

0
1
2
3
4
5
6
Random placement across layers
$\mathfrak{p} \propto d_{vw}^{\eta}$
η : Power
v, w : Distance

Figure 4.16: A deep network where additional layer bypassing links are randomly placed wiit a power-law probabilistic relationship akin that of a small-world network.

Humans and a few more mammal species have what are called spindle cells—von Economo neuronal links—where long connections appear in the folds in the brain. These are bypass links connecting laysers. Something similar can be done with multilayer networks as shown in Figure 4.16 by jumping hidden layers. The introduction of bypassing random links has a major consequence summarized in Table 4.5. The 9 is now recognized quite well. *The neural networks as practiced right now are likely quite elementary models.*

The programming of bypassing layers in a back and forward propagation is not trivial. It breaks what makes the mass-integrated-scale manipulation that has been so successful through principles of stream processing. It breaks the simplicity of matrix multiplication that *JAX* and other such developments have so successively exploited. This is again the software-hardware cycling that has existed over the past so many decades.

		100 nodes	784 nodes	
	7	12%	62%	
	9	65%	71%	was 20%
	7	12%	2%	
	9	75%	91%	was 67%
	7	11%	8%	
	9	73%	88%	was 18%

Table 4.5: 7 and 9 classification with von Economo layer bypassing neurons. This table is to be compared to the previous one (Table 4.4). The last column points out the improvement.

Now consider an interesting example of context-sensitive disambiguation. In the Devanagari alphabet, there are three characters, *kha*, *ra* and *va*, where if *ra* and *va* follow each other they look pretty much like *kha*. One quickly learns from context what it is. This is the Markovian twist to the human mind. But, if one is only feeding what looks like one character at a time, not a character in a word in a sentence, or pose a two or one character identification, as in this example, what should one expect to classify? The first suspicion is that it really will be related to distancing of the characters and the second will be of some higher order rotational convolution effect since the *kha* will have a tendency to be more closely tied in the two vertical orientations of the character. Again, the introduction of jumps in across hidden layer improves identification as seen through Figure 4.17.

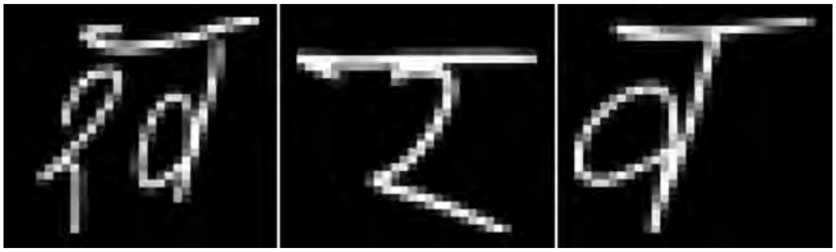


Figure 4.17: Devanagari for *kha* versus *ra* and *va* as three characters of the alphabet.

It is fun to see that this simple translational and rotational argument—a physical view—holds ground, or at least not rejected, as seen in Figure 4.18. The use of von Eckonomo neurons provides considerably more accuracy. Stochasticity and connectivity and how the information moves and is encapsulated across the layers, preserving long and short range connections is the message of this model experiment.

This brings up to the question of information, and how it is being maintained, manipulated, and what particular schema of connectivity preserves it the best as one moves across layers before the final classification from the information in whatever is the optimal form it is encoded in the neural network. In the example of Devanagari characters, this was a translation and rotation at the character level and the context at the more collective level in the sentence.

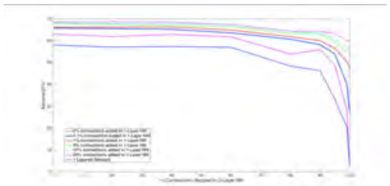


Figure 4.18: The accuracy of recognizing *kha* versus a *ra* followed by a *va* Devanagari for *kha* versus *ra* and *va* using different percentges of bypassing links.

The number 8 is an interesting example that collects a rotating collection of points forming a curve that we identify with 8. How is it to be viewed informationally? One way is to think in terms of correlations, pair wise and higher, and the next is to explore it in a mutual information view through the different measures of entropy. The various unusual symmetries in the character 8 are buried in these measures to different extent. This computed results are shown in Figure 4.19. The long range correlation of symmetry is not sufficiently distinguishing. The short-range matters as the Renyi and Fisher entropy show. A completely checkered pattern will have a flat output in the bipartisan locale coordinate. Renyi and Fisher keep a lot of information together, which the Shannon measure does not.

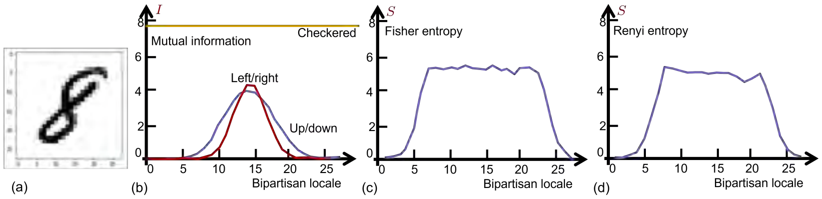


Figure 4.19: (a) shows the number 8 character being analyzed, (b) the mutual information in Cartesian pairwise calculation of left/right and up/down, (c) is the Renyi entropy and (d) is Fisher entropy.

The statistical view of figuring it out or not is that if one has not previously seen some behavior in the statistics—individually or in context—-then one should of course not expect it to be predictable within a model with its tolerances. There are fluctuations for a variety of reasons, and even if the transition to the current state was unexceptional, it will fit to the model within those error bars. But, if the event is not part of the model, Poisson events for example unless there are sufficient statistics for that, one should not expect any better. The subjective view finesses this by asking one to change one's expectations. The existence of a not-before-seen event is a contradiction with repeatability expected of a frequentist or objectivist view to statistics.

Physical laws, with determinism and randomness and resulting non-determinism built in, are a way to make a model, guided still be observations, that captures all the short, long, the often, the not-so often, and all the other connections across all the domains within the constraints stated for the validity of the physical law.

Bringing these two together is one way of bringing more accuracy to predictability, both are grounded in observations, but the physical—the science principles that must hold regardless of the mathematical undercurrents of the network—must hold true. This places constraints that are the principles that one has found to hold true over a wide variety of domains. This brings together all the different disciplines of sciences that establish the immutables within the constraints of their axioms that we have found to hold. It may be probabilistic—the Kolmogorov-axiom-based edicts or of number theory or geometry or others—but it must also be beholden to the rules of principles of action, summing of the paths, relativity, electromagnetics, all the different conservations, and all others that come from scientific fields.

Placing both statistical and physical constraints within the neural network—done right—should make it more accurate in predicting. It helps tackling and increasing robustness in predicting what cannot be generalized form limited observations.

A good example of using neural network in this way is quite an old story. The European laboratory *CERN* produces in any complex experiment an incredible amount of data from its large number of sensors operating in time and three dimensions. It builds up to trillions of bytes of data per hour. One cannot store such data on the fly and if one attempted to it will absorb all the world's resources. The clever out from this is to actually have a model, look at the data in real time for what does not fit, and store only the data that does not fit. One can do that by designing application specific integrated circuits coupled to field-programmable arrays and storing away in a hierarchy of memory and storage media.

So an appropriate way of trying to figure this out—coupling the statistical observations to the constraints from some physical (or our version of it) laws—is the challenge. This means knowing enough from statistics and recasting the physical in terms that the neural network can understand. Guiding both is the information, the flow of information in the network, and the flow being constrained by the physical constraints on how the information cam behave.

Another way of saying this is that the information is being exhibited in various forms—within the degrees of freedom afforded by the network where the last layer may have the least since it is demanding classification– and the physical law is placing a constraint on the flow of information. This is not unlike the speed of light limit on flow of information, which is embedded in the Cramér-Rao bound if one looks at the related physical equation. A discussion of this can be seen in the Oxford Volume III, Chapter 2, of *Semiconductor Physics*.

Statistics need to be sufficient. A statistic $T(\mathbf{x})$ is sufficient for a model with its unknown parameters ($\boldsymbol{\theta}$) if no other statistic from the sample space can provide additional information on the value of the parameter, that is,

$$\mathfrak{p}(x|T(\mathbf{x}),\boldsymbol{\theta}) = \mathfrak{p}(x|T(\mathbf{x})). \tag{4.19}$$

This is equivalent to

$$\mathfrak{p}(\boldsymbol{\theta}|T(\mathbf{x}),x) = \mathfrak{p}(\boldsymbol{\theta}|T(\mathbf{x})), \tag{4.20}$$

which states the conditional probability of a parameter is now independent of the data, and

$$\mathfrak{p}(\boldsymbol{\theta},x|T(\mathbf{x})) = \mathfrak{p}(\boldsymbol{\theta}|T(\mathbf{x}))\mathfrak{p}(x|T(\mathbf{x})), \tag{4.21}$$

which is a statement of statistical independence. A pithy to emphasize the nature of sufficient statistics is that what has not been seen cannot be generalized. The assumption is that no other statistic from the sample space adds to information.

If one thinks of this situation in probabilistic terms, the learned function has information in the small probabilities in soft targets. The soft targets that have the highest entropy, that is, of lack of knowledge about them, then it is their observation that provides the highest information for training. The hand-written numbers 2, 3 and 7 where they appear confusing to us with each other, are so because their low probability tells us that there is rich similarity in the structure, and it is this similarity that is confusing us. If one trains the network on such structures with rich similarity, the network will become more accurate on distinguishing these 2s, 3s and 7s lot better. It is this similarity that was better captured by bypassing neurons across layers—feeding information over another scale—that made the 7 and 9 deconvolving more accurate with the von Eckonomo model experiment.

Of course, even with science constraining, what has not been seen may perhaps be there in the real world, and by placing science constraints, in the modeling of training data that provided us with the statistics and then making predictions, we have not given the network a tool to predict something new beyond the data. The science constraint only optimized the network on the data. Any prediction will have to be consistent with the data and the science within the network's accuracy.

The physical meaning, that is our translation to another domain of different dimensionality, in our associating the meaning, and the physical constraints, are helpful in understanding what transpires in neural networks and it also improves them since we are placing a rigorous natural constraint on them through the science

This viewing is quite consistent and since we have used the word energy often in the context of networks, it is a useful construct to provide an interpretation to probabilities.

Take $\mathfrak{p}(\mathbf{x}|\theta)$. This of $x(\mathbf{x})$ given that it belongs to some feature θ. The Bayes view of the probability of the feature given the set $\mathbf{x}$ is

$$\mathfrak{p}(\theta|\mathbf{x}) = \frac{\mathfrak{p}(\mathbf{x}|\theta)\mathfrak{p}(\theta)}{\sum_{\theta'} \mathfrak{p}(\mathbf{x}|\theta')\mathfrak{p}(\theta')}. \tag{4.22}$$

This is a revisiting and change. Any prediction is a matter of flow—it is a momentum—with the position being the initial condition, which is the boundary for this problem. So Bayes in a way is making a canonical statement, here stated in classical physics terms, of position and momentum in observations. In classical mechanics, we relate these through the Hamiltonian. Take the Hamiltonian for this problem as

$$\mathscr{H}_\theta(\mathbf{x}) = -\ln \mathfrak{p}(\mathbf{x}|\theta), \tag{4.23}$$

the mean as an expectation of

$$\mu_\theta = -\ln \mathfrak{p}(\theta), \tag{4.24}$$

and we have

$$\mathfrak{p}(\theta|\mathbf{x}) = \frac{\exp\{-[\mathscr{H}_\theta(\mathbf{x}) + \mu_\theta]\}}{\sum_\theta \exp\{-[\mathscr{H}_\theta(\mathbf{x}) + \mu_\theta]\}}. \tag{4.25}$$

Since $\boldsymbol{\theta}$—a parameterization of the model—is one of a discrete set—an index—vectorially one can write

$$\mathfrak{p}(\mathbf{x}) = \frac{\exp\{-[\mathscr{H}(\mathbf{x}) + \boldsymbol{\mu}]\}}{\sum \exp\{-[\mathscr{H}(\mathbf{x}) + \boldsymbol{\mu}]\}} \tag{4.26}$$

We have arrived at an equivalence between Bayes view and the partition functions with $\sum \exp\{-[\mathscr{H}(\mathbf{x}) + \boldsymbol{\mu}]\}$ being the zustandsumme. The Hamiltonian is the surprisal, unlikely events of low probabilities have higher Hamiltonian, an energy, which in the probabilistic language is the surprisal.

This view can now be extended to see what neural networks are doing. Take an n-layer feed forward neural network. Using the affine and nonlinear transformations of the network, with σ_i as the local operator (tanh, softmax, max-pool, *ReLU*, et cetera, and $\mathbf{A}_i = \mathbf{W}_i\mathbf{x} + \mathbf{b}_i$ as the affine transformation in the ith layer of the network, one can now use Equation 4.26 together with the nonlinearity—take, for example, softmax of $\boldsymbol{\sigma}(\mathbf{x}) = \exp(\mathbf{x})/\sum_i \exp(\theta_i)$—and we now have the probability

$$\mathfrak{p}(\mathbf{x}) = \boldsymbol{\sigma}[-\mathscr{H} - \boldsymbol{\mu}] \tag{4.27}$$

The parameters/features of the set of "means" $\boldsymbol{\mu}$ is now just a bias vector for classification probability in the final layer extracting features when using softmax. This is what we see with our very early digital logic example. The Hamiltonian has a meaning in energy function terms and is computable. The central limit theorem implies that we will end up with a multivariate Gaussian in the form

$$\mathfrak{p}(\mathbf{x}) = \exp(h + \sum_i h_j x_i - \sum_{ij} h_{ij} x_i x_j) \tag{4.28}$$

for the network, which shows that the Hamiltonian $\mathscr{H} = -\ln \mathfrak{p}$ is a quadratic polynomial (hs are abstractions that are the coefficients-or-eigenvector-like response to the operator operating on the state description in $\mathbf{x}$.).

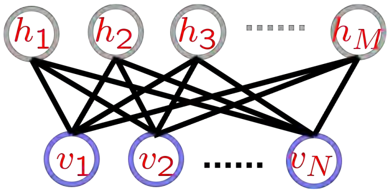


Figure 4.20: A single layer of a restricted Boltzmann machine with individual inputs representing a state v interacting through the affine and nonlinear transform leading to a state h.

In Figure 4.20 is this formulation rewritten in energy form. This is a restricted-Boltzmann machine (*RBM*), that is, a condition which in physical view says that there is no interaction with oneself—no singularity in this—and that interactions between all the interacting terms of representation represented in the nodes v lead to a quasi-representation—nodes hs for hidden—with individual bias, pairwise and higher-order effects. To get this right by efficiently agglomerating self and mutual information, the deep neural networks requires that layers with many hidden nodes exist in intermediate stages so that the information flow across the network leads to the final dimensionality reduction that leads to efficient classification. The central-limit theorem is making a stronger statement for what is happening in this network with the input state being classified through intermediate representations.

This same behavior is representable for the recursive networks as shown in Figure 4.21. Each one of the intermediate neural networks is an efficient informational representation for a subset written in terms of the probabilities. It is a higher-dimensionality Markov chain.

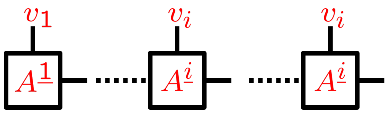


Figure 4.21: The recursive network—transformers and large language models being an expanded generalization—viewed as restricted wave.

For the restricted Boltzmann machine, this is the mathematical reduction to

$$\begin{aligned}\mathfrak{p}(v) &= \frac{\exp[-E(v)]}{\sum_v \exp[-E(v)]}, \quad \text{where} \\ E(v) &= -\sum_i a_i v_i - \sum_j \ln[1 + \exp[b_j + \sum_i v_i W_{ij}], \end{aligned} \tag{4.29}$$

and for the restricted wave—recursive network—this is

$$\begin{aligned}\mathfrak{p}(v) &= \frac{|\psi(v)|^2}{\sum_v |\psi(v)|^2}, \quad \text{with} \\ \psi(v) &= \mathrm{Tr} \prod_i^N A^{\underline{i}}[v_i]. \end{aligned} \tag{4.30}$$

These are classical-like physical equations of the network where the coefficients or the matrix elements are approximate representation of the interaction in the affine-nonlinear coupled terms. The restricted Boltzmann machine representation in this form is also subject to the information constraint, So, for all $x \in \mathbf{x}$ and $y \in \mathbf{y}$, the mutual information flows through these hidden variables parameterizing the nodes of the network as

$$I_{RBM}(\mathbf{x} : \mathbf{y}) \leq I_{RBM}(\mathbf{x} : \mathbf{h}) \leq |H| \ln 2, \tag{4.31}$$

where $|H|$ is the Shannon measure. The statistics arise through probabilities related by the Hamiltonian in surprises. The recursive networks—like the Markov chain sequenced hierarchical construction—-are just multiplication of ever-more complicated matrices that are handled quite well approximately by the neural networks since they are provide the approximation of nature's approximations through its cause-and-randomness.

The recursive network is generative. It is not subject to any constraint that have not been placed. It is just projecting approximately unilaterally on what the model has been subjected to, properly checked, or right, or factual, or not.

What holds true in all these instant calculations—their approximations and their evolution in recursion—is that the relationship is the constraint from information of Equation 4.31. This viewing also shows that linear transformations also leave invariances intact. *The weights together with the nonlinear transformations are an approximation of causal connections, and of estimations based on sufficient strength existing in the inputs to cause a change, a process that is very much dependent on beliefs, errors of judgments, observations and errors. The probability graphs do this same procedure in their own analytic way.*

We as nature's agents and machines as computing agents are estimating and infering in not too different a way, but each of us is limited by what we have built as a view of our limited world. This limited world of us (machine too) is subject to us and our biases, our accumulated past, and our neural infrastructure. For the machine, it is whatever was fed into it and its infrastructure.

The natural language programs are an allegory of how we view the world and how we act in the world. We build on our experiences and past, connect across domain, hide all the mathematical symbolism in something that happens in our minds, to make judgments and decisions and in our actions. So do natural language learning programs. There is content within the network of the laws of evolution—what to do given and what happened in coded symbolic informational content. Nature, by the way, doesn't give one unique answer. Try doing anything natural, starting flowers from seeds, raising children, going from one place to another, et cetera, and one quickly finds out how valuable the Fokker-Planck equation is in bringing in the importance of stochasticity and yet a connection and an order arrives.

None of this inputing and outputting has anything to do with the language used to describe the problem assigned. A thousand plus years ago *Baudhayana Sulvasutra* was expressing Pythagoras equation in verse. Symbolic algebraic equations arrived from middle east to efficiently express the relationships. Leibniz in the middle of the millenium gave us a way to describe small changes and integrative effects. These are all language forms. Symbol, assembly of symbols as words and phrases, and by a grammar or the rules of the game that are externally imposed but also appear through internal processes, giving meaning to this collection. An efficient representation each step of the way. As humans, we start knowing the numeral 1 and 2, get up to 4 by age of two, but this process becomes an arithmetic algorithm for us to add and subtract object numbers to describe that world around us. Same with characters, words, words in sentences, and as Kierkegaard, the Danish theologian, posited meaning from the context of the collection of words under the rules of the grammar. This argument places the number sense before Chomsky's language development arising from the hard coding in the frontal cortex. This is the reason I think large language models are so enormously powerful. It can interpret all the different disciplines of sciences, their coming together, and in turn be capable of a natural description. It develops its own internal grammar/algorithm driven meaning.

The physical underlying characteristic in these complicated viewings is that what-is-not-known, that is, entropy is

$$H(\mathbf{x}|\mathbf{y}) = -\sum_{x,\mathbf{y}} \mathfrak{p}(x,\mathbf{y}) \log \mathfrak{p}(x|\mathbf{y}), \tag{4.32}$$

which tells one what the uncertainty is as resulting form true uncertainty, that is, what is not really known and is subject to whimsies of nature, and what is also there due the uncertainty of the finiteness of the data. The Kullback-Leibler divergence is the measure—the relative entropy—between distributions, that is, in our inferencing between what is known and what is our model of what it says in the form

$$\mathscr{D}_{KL}(\mathfrak{p}(\mathbf{x},\mathbf{y}) \parallel \mathfrak{q}(\mathbf{x},\mathbf{Y})) = \sum_{\mathbf{x},\mathbf{y}} \mathfrak{p}(\mathbf{x},\mathbf{y}) \frac{\mathfrak{p}(\mathbf{x}|\mathbf{y})}{\mathfrak{q}(\mathbf{x}|\mathbf{y})}. \tag{4.33}$$

There are errors in measurements, and there are variations too that arise in the limits and limitations of the data. Finiteness of data is like a broad-spectrum effect, so like that of thermal fluctuations. In the physical picture then, the inverse temperature of a random variable $\beta_0(\mathbf{x}) \propto 1/T$. This can be defined in terms of the Kullback-Leibler divergence between the estimator at data size n to the probability estimator with the data removed. Uncertainty of true probability and uncertainty of the finiteness of the data get captured by the Kullback-Leibler distance. The fluctuations, and the finiteness of data will look very much like a thermal fluctuation.

We have argued the mapping between entropy—the lack of-information—and energy, the energy being a measure of what the information content known gives us as a capability to do something useful with. The neural network is the information machine, with information present in some form coded in it—like a computational program or a physical evolutionary law except it is approximate and obtained from the training—so the same ideas as those of the thermomechanical machine relate. The difference is just in these meaning assignments. Bond strength, energy in waves of fields, or of particles' motion, or of atom's vibration, or within them, even within the nuclei, are all ultimately still a form of information.

This point is subtle. Information, we will look at more carefully in the next essay again from a different perspective, but it is also bond strength, energy spread out in the vibrations, and all those very many other confusing ways that energy appears in the language of science since it is, like entropy, a characteristic that in totality is encompassingly useful—energy is what energy does—but then, for convenience, we go ahead and obfuscate it.

We can draw parallels based on this correspondence between our information-centric neural network description and statistical- and thermodynamic-centric description of the physical world. A few key thoughts are

- The *internal energy* $U_0(\mathbf{x})$ is the Kullback-Leibler divergence between the target and empirical distribution, that is $U_0(\mathbf{x}) = \mathscr{D}_{KL}(\mathfrak{p}(\mathbf{x}) \parallel \mathfrak{q}(\mathbf{x}))$. If the the stable minimum energy state has been achieved and the target and empirical estimates are identical. The Kullback-Leibler divergence zeroed out. If it is not, in the complex problem with its errors and fluctuations and incompleteness, the distance is an estimate of the differences, and the internal energy is a measure of how the two are different under a weighting related to the distribution of the likelihoods. Minimizing the divergence is finding the maximum likelihood, the estimation of most weighted and most likely.

- The *cross entropy* is how the lack of knowledge exists between one distribution sampled using the other. This is like viewing another world based on our view of the world. Cross entropy $U(\mathbf{x}) = H(\mathfrak{p}(\mathbf{x}), \mathfrak{q}(\mathbf{x}))$ is the information disconnection between these two worlds described by the two distributions. The self information is a very Shannon-like measure based on relative frequency of surprisals. $\mathcal{S} = \log \tilde{\mathfrak{p}}$ is this self information.

- The *Helmholtz free energy* $\mathcal{F}(\mathbf{x})$ tells us the capability of a system for productive conversion of energy to ˵work˝ under certain macroscopic parameters (pressure) kept constant. The same is true for the neural network with $\mathcal{F}(\mathbf{x}) = [U_0(\mathbf{x}) - H(\mathbf{x})]/\beta_0(\mathbf{x})$ as the useful information energy. How far is the internal energy from the Shanon entropy determines the available information processing capability of the system.

The statistical mechanics techniques of using the partition function can now be applied in the same way. With Z as the partition function,

$$H = \beta U + \log Z,$$

$$\begin{aligned} \beta &= \partial H_U, \\ \therefore\ \mathcal{F} &= U - \frac{1}{\beta} H = -\frac{1}{\beta} \log Z, \text{ and} \\ -\partial_\beta U &= \langle \mathcal{S}^2 \rangle - \langle \mathcal{S} \rangle^2 = I(\beta), \end{aligned} \tag{4.34}$$

where $\mathcal{S}$ is the surprisal. Note how the same form of relationships hold.

The last equation is particularly information centric in describing the relationship between the energy fluctuations and the Fisher information that looks at broader relationships between the symbolic representations of the data. The Bayes relationship as well as the partition function relationship discussed earlier are a manifestation of this correspondence in the foundations of these engines.

4.10 Science- and information-guided neural networks

Given that one has an incomplete knowledge of the system being explored by neural network, there are errors of measurement or even fakery and limited sampling and number of training samples, just like in a physical system, where surfaces and bulk with their different interactions and symmetries lead to properties that change substantially.

Take gold, for example, bulk gold with surface asymptotically unimportant has a number of symmetries and interactions across scales that end up describing a melting point. Take the gold at 10 nm size, so with the order of 10^6 atoms, of which 10^4 are on the surface and with the surface region under different interactions constraints, the melting point can be reduced by as much as 300 K. Size mattered. It changed behavior. How one should calculate under this nanoscale circumstance is different.

The Kullback-Leibler divergence is still relevant as a measure of true to estimate differences. What should be the way to get accuracy. Should it be maximum likelihood, maximum entropy, minimum free energy or something else?

Physical principles tell us that the maximum entropy is some ‶quasistatic″ equilibrium state under conditions of being in a reservoir-like environment. This is when all the rapid processes have exhausted themselves. We also learn that minimum energy is a ‶ground-state″-like condition describing a condition for a system to settle in by exchanging away excess energy. Neither of these are representative of a system performing something useful, that is, using information—an energy-like form as we have argued—to cause a change. This is stimulation and response. Stimulation requires ability to impart energy and the system needs to be out of equilibrium and have excess energy to make change come about. This is the problem of analyzing and predicting the out-of-equilibrium phenomena that are central to all the problems of interest, whether in neural networks or in real life or in physical sciences. Maximum likelihood, maximum entropy, minimum free energy or something else are all plausibilities for attacking the problem. Minimum free energy may get the melting temperature right for gold even though there will still be some suspicion regarding the effect of heating and cooling rate on how the energy interactions are taking place within a nanometer-sized object. We just can not place full confidence in the end result because of the incompleteness and open boundary conditions. Placing more and more objective scientific constraints helps in narrowing the window of plausible solutions.

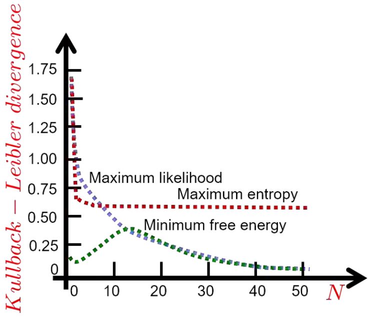


Figure 4.22: Comparison of three different methods of estimation as a function of sample size with 3 different internal states of probabilities $\mathfrak{p}(0) = 0.850$, $\mathfrak{p}(1) = 0.116$ and $\mathfrak{p}(2) = 0.034$.

Figure 4.22 shows results of simulations with the three different extremizations posited. The data has three different internal states of specified probabilities and one can see that for large-enough sample sizes maximum likelihood and minimum free energy lead to smallest divergence. The maximum likelihood is the usual measure in classification determination, but for small samples one sees that minimum free energy would be a better choice. Take your pick. It entirely depends on your world's size. Given a large enough size of the world, maximum entropy is the worst choice. This stands to reason. A useful system needs to be away from thermal equilibrium.

We now take a few examples from different science domains where one can place the science constraints to show the efficacy.

A standard example in condensed matter physics is the use of Ising models as prototypical system to see interactions, of energetics, and of phase transitions as conditions such as that of temperature are changed. For the Boltzmann machine, if $\boldsymbol{\sigma}$ is the spin array, at thermal equilibrium, the probability in the physical models, and the weights and biases spread out over the hidden nodes and connections in the neural networks, are accounting for interaction contributions.. In the language of neural networks co-mapped with the physical-mathematical description, this is

$$\mathfrak{p}(\sigma, T) = \frac{1}{Z}\exp[-\mathscr{H}(\sigma)/T]. \tag{4.35}$$

For the neural network, take the model in the form

$$\begin{aligned} \lambda &= \{\mathbf{W}, \mathbf{b}, \mathbf{c}\}, \\ \mathfrak{p}_\lambda(\sigma, h) &= \frac{1}{Z_\lambda}\exp[-E_\lambda(\sigma, h)], \text{ and} \\ E_\lambda(\sigma, \mathbf{h}) &= \sum_{ij} W_{ij} h_i \sigma_j - \sum_j b_j \sigma_j - \sum_i c_i h_i. \end{aligned} \tag{4.36}$$

The marginalization of the joint distribution, a Bayesian summation, then gives

$$\mathfrak{p}_\lambda(\sigma) = \sum_h \mathfrak{p}_\lambda(\sigma, h) = \frac{1}{Z_\lambda}\exp[-E'_\lambda(\sigma)]. \tag{4.37}$$

This is the posterior as just the summation of products over likely pathways. Bayes, stationary action, and Feynman! A restricted Boltzmann hidden layer is modeling the Bayesian process and we have incorporated that energetics of the interactions in formulating Equation 4.36.

With the mathematical representation that has a physical meaning, and the physical constraint, this is an example of science-constrained neural network. The result shows the approximate calculation by the neural network that is similar to that from statistical mechanics. Figure 4.23 shows as a function of temperature how a model calculation behaves for an averaged expectation—an ordered parameter—and the phase transition as one proceeds across a normalized temperature. There is a specific weight—the energy interaction condition—which determines the phase transition.

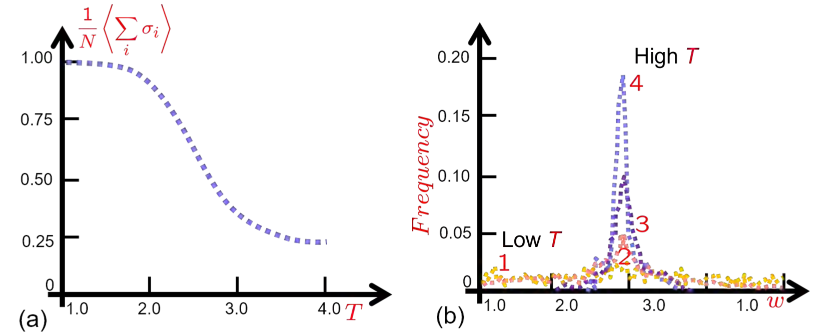


Figure 4.23: A restricted Boltzmann machine based Ising calculation showing phase transition. An order parameter changes as a function of temperature as shown in (a). In (b) is the frequency statistic for four different temperatures as a function of the interaction parameter that is tied to the energetics of the interaction.

Lagrangians and Hamiltonian—the energy-scalar-based formulations—are a standard mathematical physics tool for solving physical problems. Nearly all of what we see in action-based behavior is describable through it. Hamiltonians follow canonically from the fundamental stationary action principle based Lagrangians. Because they are based on scalars in energy form, they provide a very convenient tool from which forces can be derived. *Lagrangians are statement of stationary action, a very profound physical statement. Using stationary action therfore can implicitly introduce Langrangians in neural dynamics.* And if handled canonically, Hamiltonians too can be deployed.

This says that physical principles can be implicitly built in, or can be explicitly built in as constraint conditions in the algorithm of generating—classification or of evolutionary dynamics—in neural networks of all the different types that we have discussed.

This is the science-based extremization, add to it the mathematical loss minimization and regularization of the neural network, which we have established is an information engine subject to the physical laws and explainable through the information edifice. So Lagrangians and Hamiltonians, transformations in and through them for effective manipulation such as Hilbert or simplectic, coupled with cross entropy, or mean square, or others, with regularization is a neuralnetworked based physical approximate description of the real world. The approach works for all neural networks, autoencoders, transformers or recursive neural networks, in generative networks, and others in all the different domains to project the dynamics and for extracting physical parameters.

Neural networks with science constrains built-in across the network or as an optimization constraint is an approximate tool for complex dynamics. It should tell us the parameters of the dynamics, for examples, a Fokker Planck equation should fall out of it, and it should also generate the dynamics into the future.

I will illustrate this through examples, one is where Langrangians are a constraint implicit in the neural network. The others are to show the dynamics because the neural network is capable of describing the Fokker-Planck-like dynamics.

First, we look at dynamics in a trivial model of the damped harmonic oscillator whose equation therefore is known because of the simple problem statement.

$$\begin{aligned} md^2_{t^2}z + \mu d_t z - k_s z &= 0, \\ z(t=0) &= 0, \\ d_t z|_{t=0} &= 0, \text{ and} \\ \left[\delta = \frac{\gamma}{2m}\right] < \omega_0 &= \left(\frac{k_s}{m}\right)^{1/2}, \end{aligned} \tag{4.38}$$

where the equations as written are for under-damped oscillation because of the last part's constraint. m is a mass, γ is a damping parameter, k_s is a spring constant to represent a conservative force, and ω_0 is the natural oscillation frequency to be expected in the ground/equilibrium unstimulated state. We know the physical solution of this precisely solvable problem as

$$\begin{aligned} z(t) &= 2A\exp(-\delta t)\cos(\phi + \omega t), \text{ where} \\ \omega &= \left(\omega_0^2 - \delta^2\right)^{1/2}. \end{aligned} \tag{4.39}$$

All one has to do for this problem is to first train the neural network to interpolate part of the solution from training points. That is, one may generate some points (including adding stochastic noise to it). Follow it by then forcing the neural network to extrapolate by penalizing the underlying differential equation in its loss function. The loss function is just $\| \hat{z} - z \|^2$. We end up with a neural network whose weight and bias parameters model the regularized dynamics that the data models, and explicitly, this neural network then shows the dynamics. The damped oscillator could be learned with a few points. Once trained with a few points one implicitly knows the equation form coded in the weights and the biases of the network.

Under any perturbation, the estimation minus the actual can be minimized and predictions generated. The calculation is easy enough to do that. The gray line here is the actual solution to the problem and then this other one was a neural network solving that problem. This is shown in Figure 4.24. A generative network now exists that can model this simple problem in time and space.

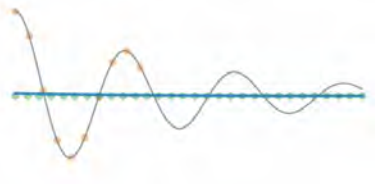


Figure 4.24: Damped harmonic oscillator dynamic extracted from training data followed by a generative output of the dynamic.

It is this capability to incorporate every-increasing complexity as more is known into the neural network's predictive behavior—approximate as the reality and the network are—that make neural networks so interesting.

We saw some of this oscillation—a lot more complex dynamics—during the recent Covid tragedy. We had some data—errors, warts and all—and it was possible to then predict what the behavior will be based on past data. The challenge was that parameters are changing in time. But the method is an approximation where they can be pulled in, something that is difficult or sometimes impossible to do when one is looking for a clean precise analytic solution In the case of Covid, with these time-dependent parameters, the response is

$$z(t) = 2A\exp(-\delta t)\cos(\phi + \omega t) + c. \tag{4.40}$$

The spring constant k_s, the damping γ, and the bias term c, all time-varying can be left for the neural network to implicitly model and one can predict. Figure 4.25 shows a modeling of the prediction (the dots are from the training data) of how in time the past predicts the future with a what looks like damped oscillatory complex response. What is also important to recognize is that this approach is not restricted to a simple oscillatory picture. It can be far more complex. It can be in multiple dimensions such as space and time, and therefore, given sufficient data, also will describe the Fokker-Planck like behavior of *IITf*'a Kanpur graduates staying or moving away from India and coming back and eventually, as we all will, pass on.

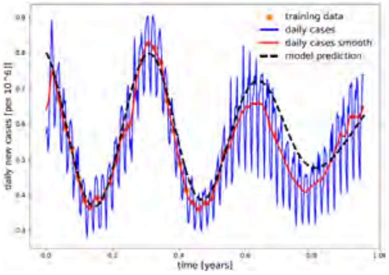


Figure 4.25: A predicted behavior of the complex Covid dynamic from a limited data into the future with the time-sequence being implicitly modeled by the neural network under Gaussian mean-squared constraint.

The next two examples are a little more elaborate. They show how science such as of action in mechanics or nonlinearities of incompressibility in fluid dynamics can be implicitly incorporated.

The Lagrangian approach—an action-based approach—applies to situations where conservation of energy holds and so all those forms have to be folded in. The state represented by $(\mathbf{q}, \dot{\mathbf{q}})$, a generalized position and velocity, does not need canonic variables, which, Hamiltonian does, and the Euler-Lagrange equation describes the evolution through

$$d_t \boldsymbol{\nabla}_{\dot{q}} \mathscr{L} = \boldsymbol{\nabla}_{q} \mathscr{L}, \tag{4.41}$$

whose inversion leads to

$$(\boldsymbol{\nabla}_{\dot{\mathbf{q}}} \boldsymbol{\nabla}^{T}_{\dot{\mathbf{q}}} \mathscr{L})\ddot{\mathbf{q}} + (\boldsymbol{\nabla}_{\mathbf{q}} \boldsymbol{\nabla}^{T}_{\dot{\mathbf{q}}} \mathscr{L})\dot{\mathbf{q}} = \boldsymbol{\nabla}_{\mathbf{q}} \mathscr{L}, \tag{4.42}$$

which can be placed in the form

$$\ddot{\mathbf{q}} = (\boldsymbol{\nabla}_{\dot{\mathbf{q}}} \boldsymbol{\nabla}^{T}_{\dot{\mathbf{q}}} \mathscr{L})^{-1} \left[\boldsymbol{\nabla}_{\mathbf{q}} \mathscr{L} - (\boldsymbol{\nabla}_{\mathbf{q}} \boldsymbol{\nabla}^{T}_{\dot{\mathbf{q}}} \mathscr{L})\dot{\mathbf{q}} \right]. \tag{4.43}$$

If the time-dependence $\dot{\mathbf{q}}$ is known, then one has also determined $(\mathbf{q}, \dot{\mathbf{q}})$.

The encapsulates and integrates within it a number of ideas and thoughts relevant to neural networks.

- A physical principle writable in quantitative terms can be recast for minimization implicitly and objectively,
- the minimization constraint can then be mapped on to a neural network's loss function,
- additional principles, for example, symmetries, and so on, can generalize the loss function for robustness,
- and if all this works, one could employ such an implicit extraction for additional predictions in complex dynamic or static problems.

The point of functional as an information collective instrument is important in my view. Often we don't know the functional, even Lagrangian for complex problems, and it takes symmetries and trials and errors to figure it out. But once one knows it, the heaven and earth open up to solution. Many body problems have functionals—Kohn-Sham being one of the teachable examples—scattered all over and that we deploy. Such functionals can become implicit in the network making life much easier for complex problems.

To implement this, one needs sufficient training data, and an additional set to test. For demonstration purposes, this data may be created deterministically and noise added through analytic approaches since we know in this case the underlying equations. The Lagrangian incorporates the principle of action, where the action is a functional

Figure 4.26 shows a simple view of the neural network implementation of the first two items of the list above, and one can create a framework for solution as following.

1. Generate the training and test data. These may be obtained by analytical approaches (for simple problems) using randomized starting conditions. Equation 4.41 describes the Lagrangian dynamics. The two variables are $\mathbf{q}$ and $\dot{\mathbf{q}}$, which subsume single or many-body generalized position and velocity coordinates. The set of coordinates are $\mathbb{Q} = \{\mathbf{q}, \dot{\mathbf{q}}\}$. In order to make the problem realistic, one may add noise to this deterministic solution.
2. Deploy the solution to the dynamics equation, Equation 4.43, to determine parameters important for loss functions. We use the time-dependences in first and second order for both networks in Figure 4.26. The various partial derivatives $\partial_{\mathbf{q}}$, $\partial^2_{\mathbf{q}\dot{\mathbf{q}}}$ and $\partial^2_{\dot{\mathbf{q}}^2}$ are part of the forward and backward propagation, and can therefore be implemented in a neural network.

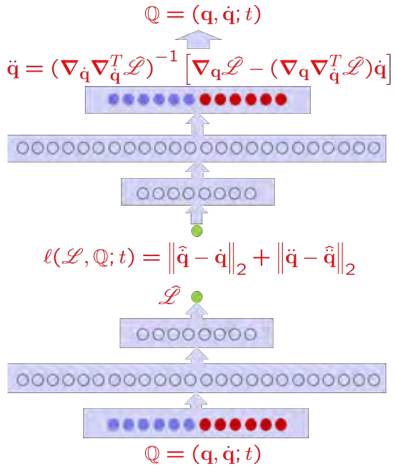

$\mathbb{Q} = (\mathrm{q}, \dot{\mathrm{q}}; t)$
$\ddot{\mathrm{q}} = (\nabla_{\dot{\mathrm{q}}}\nabla^T_{\dot{\mathrm{q}}}\hat{\mathscr{L}})^{-1}\left[\nabla_{\mathrm{q}}\hat{\mathscr{L}} - (\nabla_{\mathrm{q}}\nabla^T_{\dot{\mathrm{q}}}\hat{\mathscr{L}})\dot{\mathrm{q}}\right]$
$\ell(\mathscr{L}, \mathbb{Q}; t) = \left\|\hat{\dot{\mathrm{q}}} - \dot{\mathrm{q}}\right\|_2 + \left\|\ddot{\mathrm{q}} - \hat{\ddot{\mathrm{q}}}\right\|_2$
$\hat{\mathscr{L}}$
$\mathbb{Q} = (\mathrm{q}, \dot{\mathrm{q}}; t)$

Figure 4.26: Trajectory sets are the input from which a target meta-Lagrangian is estimated. The gradients and Hessians are also an output of this part of the network. The loss function is $\ell(\hat{\mathscr{L}}, \mathbb{Q}; t) = \left\|\hat{\mathscr{L}} - \mathscr{L}\right\|_2$ is minimized, and $\ddot{\mathbf{q}} = (\boldsymbol{\nabla}_{\dot{\mathbf{q}}}\boldsymbol{\nabla}^T_{\dot{\mathbf{q}}}\hat{\mathscr{L}})^{-1}\left[\boldsymbol{\nabla}_{\mathbf{q}}\hat{\mathscr{L}} - (\boldsymbol{\nabla}_{\mathbf{q}}\boldsymbol{\nabla}^T_{\dot{\mathbf{q}}}\mathscr{L})\dot{\mathbf{q}}\right]$ outputed from the network so that integration can be performed to determine the trajectory. This combined construction is the network.

The Lagrangian-based algorithm is then

1. Generate training and test data using an analytic solution of the Lagrangian formulation. Incorporate noise.
2. Use the test data, employ the loss function on the target versus predicted $(\dot{\mathbf{q}}, \ddot{\mathbf{q}})$ to optimize the network, as also three different derivatives: $\partial_{\mathbf{q}}$, $\partial^2_{\mathbf{q}\dot{\mathbf{q}}}$ and $\partial^2_{\mathbf{q}^2}$.

3. The loss function is minimized to obtain maximum accuracy.

4. For a general problem, employ the network to now obtain $\ddot{\mathbf{q}}$.

5. From $\ddot{\mathbf{q}}$, obtain the dynamic trajectory.

A double-pendulum example shows this approach in use. A snapshot on the right shows predicted results of such a learning

An example of textbook nonlinear equation is the Navier-Stokes equation. Much complexity is buried in it, even in its simplest of formulations. Take the equation in its two-dimensional form. Let $u(t,x,y)$ be the x component of velocity field, $v(t,x,y)$ the y component, and $p(t,x,y)$ be the pressure. Write

$$\begin{aligned} f &= \partial_t u + \alpha(u d_x u + v d_y u) + d_x p - \beta(d_{xx} u + d_{yy} u), \text{ and} \\ g &= \partial_t v + \alpha(u d_x v + v d_y v + d_y p - \beta(d_{xx} u + d_{yy} u). \end{aligned} \tag{4.44}$$

We introduce the incompressibility constraint in the form

$$d_x u + d_y v = 0.$$

These are all minimization constraints and written as differentials and Hessians. So, f and g are are science constraints (parameterized in the α and β) to be vanishing, and we have added incompressibility as a regularization constraint

Now the neural network can jointly approximate a blackbox function (like the Lagrangian) ψ, where $u = d_y \psi$ and $v = -d_x \psi$ as conjugate functions, and pressure simultaneously. This is now a joint mean-square loss.

For a problem of free stream flow, with normalized units, a cylinder diameter of 1, a viscosity = 0.01, so a Reynold number of 100, Figure 4.28 is for the dynamics of from a neural network and the vortex shedding that results. The neural ntetwork implementation of this dynamics was trivial compared to that of the Lagrangian in two-pendulum problem. That two-pendulum problem could also have been recast in Hamiltonian form, but then it expects canonic variables as inputs and outputs. This limits its generalizability though Hamiltonians are far more intuitive to understand physically.

For many examples, the Hamiltonian formulation can be transformed into a more interesting, compelling and useful form, something hard to to with Lagrangians. The following is one example.

Working with canonic variables, the coordinate set now is $\mathrm{Q} = (\mathbf{q}, \mathbf{p}; t)$. Let $\mathbf{S}$ be the time derivatives, then at time $t = t_i + \tau$,

$$(\mathbf{q}_t, \mathbf{p}_t) = (\mathbf{q}_i, \mathbf{p}_i) + \int_{t_i}^{t} \mathbf{S}(\mathbf{q}, \mathbf{p}; t) d\tau, \tag{4.45}$$

Duplicating the work of Cranmer and co-authors of the dynamics of double pendulum was my first serious foray in building neural networks. See M. Cranmer, S. Grayedanus, S. Hoyer, P. Battglia, D. Spergel and S. Ho, arXiv:2003.04630v1, 10 March (2020). A paper with authors from Princeton, Oregon State, Google Research, DeepMind and Flatiron Institute has got to be interesting. And it was. More than interesting.

A movie of the pendulum's motion, shown in the talk, can be accessed from the reference in the margin notes of the Acknowledgments.

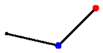


Figure 4.27: A snapshot in the trajectory picture of a simulated double pendulum oscillation.

This second problem was the next one in writing codes since I was looking for a way to tackle a way to extract the Fokker-Planck equation from data. The paper by Raissi and co-workers, applied mathematicians, M. Raissi, P. Perikaris and G. E. Karniadakis, Journal of Computational Physics, **378**, 686–707(2019), together with the first opened up during the writing—doodling—process the remarkable consequences of this approach. Navier-Stokes equation has always been fascinating ever since Prof. Ramki at IIT Kanpur put it ad hoc on the board and bringing in all these strange numbers Reynolds,. That is still the situation. It cannot be derived from first principles. Hydrodynamics, with its turbulence, laminar, chaos, and all the other changes that take place depending on conditions, is still an incompletely understood problem despite hundreds of years of work. Science matters. Soviets figured out how to make torpedoes and submarines go fast through the understanding of supercavitation. Mathematics matters. Physical intuition matters.

Regularization is a very powerful tool in introducing the physical constraints that we understand intuitively in a neural network problem. Symmetry, for example, can be introduced through Poisson brackets.

under the Hamiltonian scalar relationships of

$$\begin{aligned} d_t\mathbf{q} &= \partial_{\mathbf{p}}\mathscr{H}, \\ d_t\mathbf{p} &= -\partial_{\mathbf{q}}\mathscr{H}, \text{ and} \\ \mathbf{S} = \mathbf{S}_{\mathscr{H}} &= \left(\partial_{\mathbf{p}}\mathscr{H}, -\partial_{\mathbf{q}}\mathscr{H}\right). \end{aligned} \tag{4.46}$$

is the symplectic gradient that keeps the action stationary. The Hamiltonian maintains the energy constant, and the $\mathbf{S}_{\mathscr{H}}$ as a vector field then predicts the time evolution through Equation 4.45.

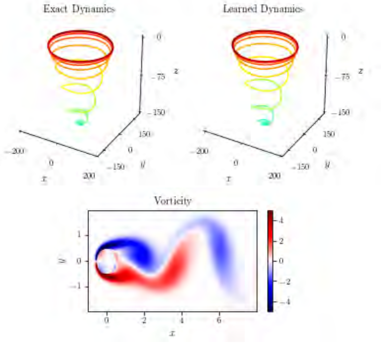


Figure 4.28: Dynamics of trajectory around a unit diameter cylinder at Reynolds number of 100, and the vortex that results in uniform flow far away for a two-dimensional Navier-Stokes simulation.

This description gives us a dynamic by employing a constraint that is based on gradients. Neural networks employ gradients, so also this gradient, in their propagation algorithm with little increase in the computation. This is also information that is available. A simple algorithm implementing this constraint via the learning a parametric function for $\mathscr{H}$ instead of the $\mathbf{S}_{\mathscr{H}}$ is the following. A conserved quantity $\mathscr{H}$ as a surrogate for energy can be learned, and a loss function written as

$$\ell(\mathscr{H}, \mathrm{Q}, t) = \left\|\partial_{\mathbf{p}}\hat{\mathscr{H}} - \partial_t\mathbf{q}\right\|_2 + \left\|\partial_{\mathbf{q}}\hat{\mathscr{H}} - \partial_t\mathbf{p}\right\|_2 \tag{4.47}$$

can be employed. Regularization of this can employ additional constraints. And one can add additional favorites.

Such an approach then gives a dynamics. The approach also gives a static picture as an instant snapshot. So, perhaps there is potential here for applying this approach to a variety of many-body problems encountered in semiconductors, and everything else non-living, and sometime in the future, living.

What is powerful here is that the network finds a defining underlying implicit principle (Lagrangian, Hamiltonian, ...) from which one extracts through the network the results for a new task. Statistical problems can often be cast in terms of partition functions. So, that is another place where this approach may become useful.

The symplectic transformation mentioned, like the unitary transforms, and others, is information-preserving rewriting of the state evolution in new form, where these can be quite easily solved and analyzed. They preserve the phase space and therefore are very useful in the neural network implementation since the neural network is another way by which the affine-nonlinear transform is changing how information is represented in a useful, even if approximate, way. The Hamiltonian dynamics lets us calculate a generalized momentum and position through

$$\partial_t p = -\partial_q\mathscr{H} \text{ and } \partial_t q = +\partial_p\mathscr{H}. \tag{4.48}$$

This set of equation can be rewritten in a flow form, called the sym-

Transformations are ubiquitous and very useful. Learning Fourier transform is a right of passage for electrical engineers. Reciprocal space contains the same information and sometimes it is better tackled there. Not just in electromagnetics or optics but also in analysis of crystalline solids or anything else with periodicity. Conjugate transforms take out singularities of sharp corners and are very useful in flow analysis. Look at symplectic Hamiltonian dynamics in this rich vein.

plectic flow form, of

$$\partial_t \mathbf{x} = \boldsymbol{\nabla}_{\mathbf{x}} \mathscr{H}(\mathbf{x}) J \;\text{ where }\; J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \tag{4.49}$$

under the symplectic condition

$$\begin{aligned} (\boldsymbol{\nabla}_{\mathbf{x}} \mathbf{z}) J (\boldsymbol{\nabla}_{\mathbf{x}} \mathbf{z})^T &= J, \;\text{ with} \\ \partial_t \mathbf{z} &= \boldsymbol{\nabla}_{\mathbf{z}} \mathscr{K}(\mathbf{z}) J \;\text{ and }\; \mathscr{K}(\mathbf{z}) = \mathscr{H}\mathbf{x}(\mathbf{z}). \end{aligned} \tag{4.50}$$

The phase space $\mathbf{x} = (p, q)$ can be transformed into the new space $\mathbf{z} = (P, Q)$, which is a latent—similar to what the neural networks are doing in hidden nodes as a latent phase-space—that preserves dynamics. $\mathfrak{p}(\mathbf{x}) = \exp[-\beta \mathscr{H}(\mathbf{x})]$ is preserved during this transformation. *This makes the dynamics much simpler to solve and even more amenable to neural network approaches.* So, the Hamiltonian way has a way to wriggle out from under the constraints of the Lagrangian way through another transformation.

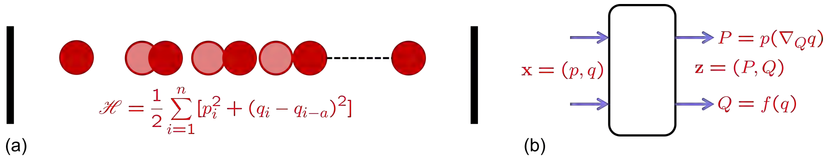


Figure 4.29: A simple problem of single atom one-dimensional crystal structure's oscillations in (a) that can be coded through symplectic neural transformation as in (b).

Figure 4.29 gives the example of the phonons and vibration problem of a one-dimensional single atom structure. The symplectic neural transformation shown in (b) is invertible and is straightforward to implement neurally. Figure 4.30 shows the solution to the problem obtained with this approach. The approach has now become powerful generalizable approach for far-more complex problems. Even problems such as of a defect that may have a Poisson rare probability, but not does not fit the model built.

All these examples, science problems with science constraints and extremization placed in suitable form, together with the approximate neural network information dynamics show the power of recasting traditional problems in a way that become far more generalizable where important principles—action, limits, information content, conservation, and others—can be kept implicit and providing sufficient data for fulfilling the need of sufficient statistics provides lets us generate the dynamics. Textbook and non textbook real world problems get within the reach even if one cannot completely describe all the interactions leading to it. *The data is the informational manifestation of all the interactions.* Noise, errors, limited data informationally are of a similar nature and the neural network is a powerful tool to tackle

Weather prediction is a perfect example, incomplete information, uncertainty of information and uncertainty of nature, non linearities and chaos, poor models, et cetera. This is an ideal place for letting a neural network handle it.

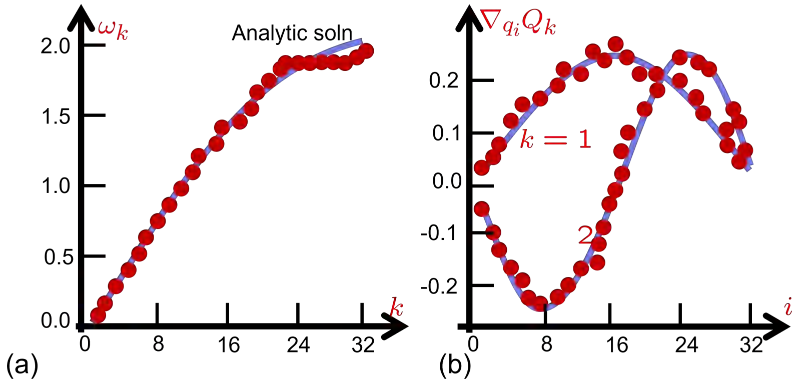


Figure 4.30: (a) shows the analytic and neural solution the phase dispersion and (b) shows two modes—the low frequency branch—of the structure.

the stochasticity. The cross entropy, that is, a mean-square optimization, together with regularization helps tackle this stochasticity using stochastic means. The science principles put the natural laws ascendant to place order.

4.11 *Revolution or evolution*

THE REASON I FIND THE APPLICATION OF THESE TECHNIQUES SO powerful and so useful and such an important change from the traditional of the past few hundred years is that one need not simplify any problem and constrain any problem into essentially a ``spherical cow.˝ We can tackle real world systems. John von Neumann is fondly recalled for saying, paraphrased, if you give me 4 parameters I can make you an elephant and if you give me a 5th one, I can make it wiggle its trunk.

I consider it very likely that we can wiggle the trunk of elephants—a real world metaphor—with neural networks in pictures as also through neural prosthesis, even if it is lots of weights, biases, nonlinear transformations, recursiveness, and a whole lot more. The wiggling trunk and much much more under constraints of all the sciences' knowledge and using many of those principles implicitly is within our reach. This is what makes it a black box sometimes, just as nature too appears as a black box sometimes, but we have a chance at solving many difficult complex open-boundary challenges of sciences and of society using these means. It is an entirely new way of attacking problems.

Maybe von Neumann understood the implication of the fifth parameter in his elephant statement. Neural networks are giant number of these parameters—still not understood in their collective form—but they do make the elephant trunk wiggle.

Just as the Indians in first millennium bequeathed us a way of tackling and understanding numbers by the idea of 0 written of by the great Brahmagupta as a place holder with profound meaning, and the Hindu numeral system that, by providing the decimal basis, made arithmetic tacklable asymptotically to infinity in both directions, and algebra became a natural outgrowth of symbolic manipulation, followed by the next enormous jump with Leibniz's calculus, neural networks as a scientific tool are another major step for tackling complexity and openness by giving away the determinism of the former.

Real world problems without many of the past constraints now become accessible. It is as another way of doing science and mathematics. No doubt that purists will find this an unpalatable and too strong a statement since it breaks a dogma or central belief just as computer-based proof caused much angst at the end of last century with the 4-color problem. It may take time for this view to take hold, we are still at a starting point, but the ability to look and project into unconfined problems of open boundaries and undiscovered principles and undiscovered scientific principles are within the reach. *The practicing of the neural network tool in itself tells us what does not fit.* By making us responsible for explaining what does not fit is in the long deep-rooted tradition of science and philosophy.

What I had written the start what Dirac had to say about quantum mechanics—principles are done, practice and figuring it out is in front of us—is I believe also true about neural networks. It opens the problems that Dirac perhaps had in his mind's eye, it is also true for all the problems that people ranging from Carnot to Claussius to Maxwell to Boltzmann to Shanon had in mind, now to a new way of tackling and resolving. No different than that Turing's test of intelligence is long in the past for machines, yet we keep struggling with figuring things out. There will always be a whole lot of money-making kind of things that most of the world will be interested in, but that there is also a lot of figuring things out that many of us might be interested in as a unique human aspiration.

I look forward to seeing this evolution take hold in this lifetime.

But caution too is warranted. *AI/ML* ending up in the long short term memory models, and their power, is still a learning to scale of connections and the underlying rules. This is inductive reasoning and tensor operation engines are enormously successful at it. The next big challenge to *AI* is how to recreate in the machine the unique human ability to integrate ideas across domains, which makes one bring the incredible insights.

Affine and non-linear transformations agglomerating information across layers over many nodes using random sampling is easily subject to being called a stochastic parrot. But there is something important—not unlike early years of a human—in there of building a model for dealing with a narrow domain of the world. Bayesian probabilities are inductive. Einstein's Gedankenexperiment of moving trains and flashes of lightening for special relativity and accelerating elevator for general relativity are to be beholden. This is deductive reasoning, not inductive.

5
Cultures: Science, engineering, interdisciplinarity and the fallacy of Ockham's razor.

What exactly constitutes science and what is engineering, or what is technology, is nebulous at best. Progress in science depends on development of new tools. Experimental tools arise from progress in technology and engineering. Theoretical tools too are creations where one connects physical and mental worlds. The natural world is chaotic and subject to randomness. It is an open, dynamic and nonlinear system. Randomness, causation, interactions, thermodynamics, et cetera., all matter. So, simplistic views—Pasteur's quadrants, Wallace-Darwin's adaptation, Snow's two cultures, Kuhn's paradigm, Ockham's razor in making simplest of choices with least axioms, and many others—are insufficient. The conduct of science and engineering has continuously changed since the dawn of modern science, it changes the world and the world changes it, changes are fast and slow and non linear, local context matters as can be seen best through commerce and non local through the wars by robotic and autonomous systems of war or of international systems such as the international monetary fund or the world bank.

In the non-local world systems, I place the security council of United Nations above all. It is a world war II anachronism in the twenty-first century. Worldly time-space to complement the science of spacetime.

How institutions practice and succeed and evolve matters for the future trajectory. Today, most problems need a simultaneous in-depth understanding of multiple disciplines even within the sciences. I discuss from personal and my experiences with the broader world the resulting conflicts: cultural such as what Snow brought up, how science and engineering has evolved from the heydays of Bell Labs or *IBM* Research, and in what shows up in the conducting of science and engineering in the world it inhabits in the modern society, particularly in the *USA* and Europe where I have spent enough time experiencing the daily living. The problem is of dimensionality reduction in complexity. In this complex world, the only rule one can draw is the Mencken's rule that *for every complex problem, there is an answer that is clear, simple and wrong.* As academics, we tend to operate within silos of our disciplines with its narrow technical language. As a person who has always been uncomfortable with this narrow self-centered minimum, this essay is an attempt against confinement. I will speculate based on this argument the interesting problems for our community that the intertwined science and engineering can fruitfully and gracefully approach.

Science and engineering in the world

This essay is dedicated to Prof. Sarma, who advised Vikas Sonwalkar and me in our design and hardware building of a photon counting system for our *B. Tech* design project. This task taught much in that last year. *Rely on one's own brains, stop looking for quick-fix answers from others or by trusting others, and the recognition that real hard work, struggles and persistence are essential to doing anything of any significance. Foremost that the struggle and the completion is far more satisfying than just imagining and talking about it.* Today, we talk about single photons, entangled photons, quantum cryptographic transmissions. These are many-generations-scale changes since those times. Yet, these new developments also bring riding on this wave plenty of more fun questions related to reality, locality, objectivity or subjectivity and other metaphysical segues. The opening of the mind's eye as in E-Y-E as well as I as in capital I through the experience gained from the faculty, fellow students and a sheltered campus thoroughly turned me into an autodidact. *IIT*, the years at *IBM* research surrounded by brilliant minds who were always willing to share their views and to debate, and the bumps along the way positioned me in the world, formed my world line and made me come to terms with the struggles of the outwardly observable I with the inner I.

It was only much later that I saw in this inner struggle a reinterpretation of the Freudian idea that the *maximal staying power is of that of the mind at war with itself.*

Thank you Prof. Sarma.

The first three essays of this series were technical, with information and entropy—as in what is not known—as their core. This one is too, but in a very different way where I step out of comfort zone of equations and figures with short explanatory stories and paradoxes in them.

We all need to be uncomfortable when we become too comfortable.

5.1 *Culture*

My view of culture, and us within the milieu is that it is all encompassing capturing our existence, our place in the world, our aspirations, our living, and our need to understand the world and perhaps to channel dynamics in a natural way. There is a part of this that for us practicing scientists and engineers is referred to as the culture of science and engineering. This then is a sub-territory of the conducting of science and engineering, its driving forces, its place in the society, what I believe in, and beyond this of us as humans on this planet with some special features. It is a very personal view built on experiences around the world, in industry and in academia,

Culture is another one of those words that is often co-opted and melded into any personal bias as politics, economics, religion, sciences, and others all do. To me it is phenomenological. Observe and see how people conduct themselves individually and in the collective towards their individual and collective aspirations. It is values, it is past defining behavior of today, it is how all the arts, sciences, respect for nature, understanding and finding pathways through different opinions, et cetera. plays out in the terroir. The Zentralfriedhof of Vienna, along its central main entrance road, has a cluster of graves of Brahms, Beethoven, Schubert, Johannn and Josef Strauss, von Suppé, Wolf, Schoenberg, also perhaps a moved grave of Mozart. On the other side is Boltzmann. This is culture alive. Not the wordsmithing such as thought leadership, influencer, woke, cancel, top down, or the passive aggressiveness of silicon valley. There is a very central cult that runs through humanity—Navajo and other southwest native Americans excepted as far as I can tell—that puts itself at the center of all that it beholds and wants to control. In *A pale blue dot*, Carl Sagan writes, ``Humans are inconsequential. A thin film of life on an obscure and solitary lump of rock and metal.˝ Nature holds seniority by a long shot over us and everything we do is just playing with our own destiny. Nature will march to its own drummer in the long run.

during a dynamic time when India itself has come of age.

During schooling, we learn about something called the scientific method. It sounded straight forward. Ask a question. Propose a hypothesis. Perform experiment. Analyze result and reformulate. This pedagogical teaching in the classroom is fallacious. It has been taught and turned into a dogma by those who have never done science or engineering. It does not work in the living classroom of life. *We constantly force fit, obfuscate parameters, introduce new parameters, or perform other tricks to make results fit with models so that they work beyond their natural territory.*

The world we live in, its pushes and pulls, its dynamism, its fallacies, its immutables, the way to think since there is no absolute truth, the cultures of disciplines, the needs of the local environment, the times, et cetera, these all affect the scientific method of a given time and place, and transduct in the society.

Reading the *Old Testament* and the *New Testament* was eye opening as a window into the world of those times in the different books and different orthodoxies and the changes over time. The different versions of Ramayana tell the same tale. Something similar can be read into the books and stories of sciences of the last five hundred years and its culture. Descartes versus Locke and Hume, Leibniz versus Newton on calculus or Wallace versus Darwin on adaptation with the Royal Society as the institution, or Montagnier versus Gallo and AIDS in our time are reflective of the same current of culture.

Let me start with the still-contemporary Pasteur-like Covid example of Pfizer's vaccine. Pfizer, because the company put the system integration together. It is a stretch to call this engineering. It is more of a putting together of the necessary resources, the money structure and flow. The lipid nanoshell is from the Canadian company *Acuitas Therapeutics*, whose founders are Thomas Madden, Peter Cullis and Michael Hope, two Canadians and an English immigrant. The *m-RNA* is from *Bio-N-Tech* whose two principal founders Uğur Şahin and Özlem Türeci are two Turkish immigrants. Two different technologies came together, and the economic, global, media and finally the social structure called it a Pfizer's Covid vaccine. In a peak quarter, half of the revenue at $> \$100\ B$ came from the vaccine and from *Paxlovid*, the antiviral drug for Covid, and so did a large percentage of its profits.

This was called innovation even if the inventions were elsewhere, and the underlying science and technology was built on the hard work and contributions by other societies and scientists..

Patents are company property.

There exists no credit to knowledge, learning and developments of past, or to the countries that educated the inventors, that is, the path and effort to the final product. This is the sum of histories that we scientists consider one of the foundational physical principles represented through action.

The poor humanity can stand in the queue and pay. This is all acceptable at least in the Western press.

This episode is what fits ‵‵crime against humanity″ label that one often also sees being used in the Western press. A dictated world order with the power of business-obfuscating language succeeds in overpowering science unlike the great Pasteur's work developing multitudes of vaccines and the foundations of immunology that the great tradition of science of health goes back to.

Simone Weil remarks in the first part of last century in her essay *Human personality* about the obfuscating potential of language. Weil portrays the plight of an afflicted vagrant standing in a court of law. ‶Even if, through his stammering, he should utter a cry to pierce the soul, neither the magistrate nor the public will hear it.″

I should stress that this society-science-engineering problem is not narrowly confined to companies or countries. It arises in science as a social force in the hand of unchecked organizations, and individuals, specially those enamored of wealth despite having grown in the middle of impoverishment with many exploiting their ability to see the gaping holes in the social framework.

India produces its own rich who do well exploiting human foibles. It is not just Mackenzie's Rajat Gupta for finance, or Pepsico's Indira Nooyi—she is not alone—for processed foods and sugary drinks undoing the benefits brought by modern science to health not to mention the mountains of trash strewn all around of non-biodegradable packaging, or in the latest financial scams such as cryptocurrencies with Nishad Singh, still 27 years old, of FTX Cryptofinance.

Addiction plays out everywhere. Lack of moral compass gives enormous opportunities to individuals and groups, private and public, politics included, to exploit.

Science and engineering and what they mean and their path, business, economics, society, and other centers are constantly in conflicts, even if the conflicts change with age as this multi-dimensional push-pull takes place.

The Pfizer vignette serves as my starting point of the argument and discussion.

5.2 *Absolution versus retribution*

Scientists need humanists. Humanists need scientists. Philosophers, artists, writers, people who think about the nature's fate—with us humans as a dominating part of this collective—are humanists to me having broken the bonds to the axiomatic or dogma-defined frame in which we practice. We scientists and engineers have plenty of success to our credit as also mistakes. The same is true for the humanist. Both can be on a grand scale. This is worth dwelling on since it instructs us.

In the essay *The two cultures*, C. P. Snow lamented the disconnection between the science-centered and the arts-centered communities by asking at gatherings of ‶cultured colleagues,″ who express their

Simone Weil was the younger sister of the great mathematician André Weil, who spent couple of years at Aligarh Muslim University, was an Upanishad scholar, besides taking other detours in life dictated by his conscience. Same with Simone, who starved herself to death, young and affected by the human suffering of *WWII*. André's *The apprenticeship of a mathematican* and Simone's writings are a veritable feast for the soul by showing how dedicated minds work. Their parents were intellectual luminaries too. I know of no other family of such mighty magnificence.

What is worse—plastics or nuclear—as science and technology's bequeathing to to our world? This is a tough question. I kneel towards my wife's view that it is plastics. It is a damage not just to us, but the entirety of natural kingdom. Nuclear has had consequences for only specific countries: the hegemons and those against which the nuclear finger has been waived. I particularly feel sorry for Cuba, a country that supplies doctors for emergencies all around the world, and yet is jailed due to the Kennedy missile crisis, which started with nuclear weapons being installed in Turkiye.

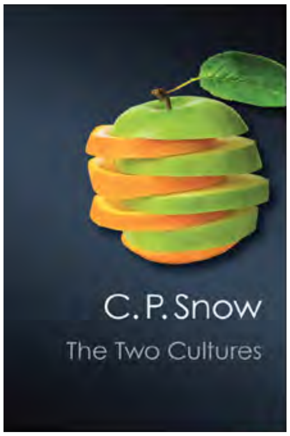
The nuclear bomb is the scientists' demonic gift to mankind.

Rudyard Kipling has a poem celebrating the Cs of colonialism, civilization, Christianity and commerce. Some of the beginning lines:

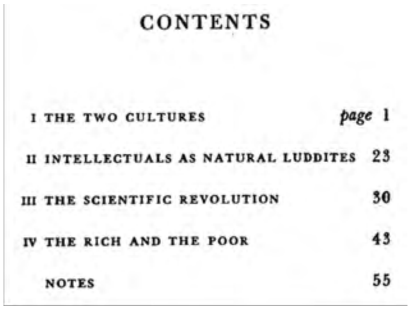
Take up the White Man's burden—
Send forth the best ye breed—
Go send your sons to exile
To serve your captives' need
To wait in heavy harness
On fluttered folk and wild—
Your new-caught, sullen peoples,
Half devil and half child
Take up the White Man's burden.

A club poet who is still part of young minds' curriculum in India. It is just like our admiration of Manhattan project, which was a sandbox for scientists who should have known better. In different forms, a euphemist process of doing good unto others is a tradition since ancient times under humanist garb. Hitler was an artist who credits white settlers' treatment of the native Americans as an inspiration for Nazism.

incredulity at the illiteracy of scientists by probing if they could describe the 2nd law of thermodynamics. To Snow, this is the equivalent of reading Shakespeare. His point being that if one doesn't understand entropy's continuing increase, and the need for energy—a general energy—to make things work well, or the arrow of time, how can one understand living.



1959



CONTENTS

Figure 5.1: C. P. Snow, *The two cultures*, ISBN 0 521 06520, Originally published by Cambridge University Press (1959).

Snow backtracks on the 2nd law–Shakespeare missive in the second edition of the writing. The whole kerfuffle is amusing since it fits so well with Marshall McLuhan's pithy commentary that the "medium is the message." Personally, I find Shakespeare really painful reading, totally disconnected from my world, and think that my interests in humanities were delayed by an early education where Shakespeare was thrust down the throat. To me the mind is not a vessel to be filled, but a fire to be kindled, as Plutarch, the Greek philosopher, said nearly two millennia ago.

The 2nd law has much to say about our living and organization, so of immense interest to philosophers, but at the same time, I can also say that we know far more today then we did in the late 1950s when Cooke uttered this question. First, that even at that time it was not a law, nor is it today a law. It is either an observation or is a consequence of states that are not adiabatically accessible. Maxwell's demon paradox and the Brownian ratchet are examples that elucidate.

The Maxwell demon paradox dates back to 1867. The Brownian ratchet as a perpetual engine to 1900 from Lippmann resolved by Smoluchowski in 1912. That information is physical, and that the demon is an information-carrying agent, and all these are things tied to probabilities are all developments of my learning life. Scientists will even say that we don't even know that the universe is closed and we certainly don't know what it even means to say that there is a probability at an initial moment given what it was at a prior moment—a moment that does not exist axiomatically.

These are probing questions that any reasonable philosopher would ask. After all Socrates emphasizes that the more a person knows, the greater his or her ability to reason.

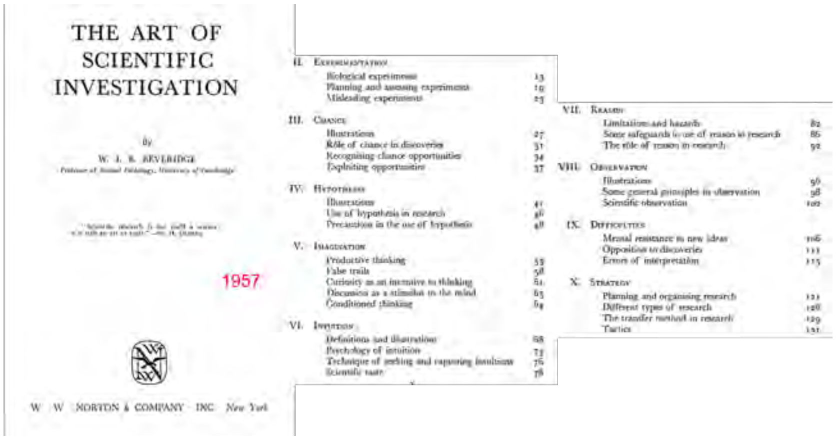
And so would scientists too. As a friend Seth Putterman, who was George Uhlenbeck's student, notes "Uhlenbeck always maintained that the 2nd law is an additional axiom of physics." That's why he used to say that "the frontiers of physics are all around us." Lev Landau in the introduction to the *Statistical mechanics* text has a section on how the 2nd law is related to whether the universe is open or closed.

Uhlenbeck disdained physicists who claimed to know where the official frontier of physics was located.

I would generalize that to people in all human endeavors.

The most fundamental meaning to me of the 2nd law is that there is the vast space of the unknown—an entropy—that we should appreciate far more than what we know along the lines of what Gödel taught us about there being unprovable truths from a consistent set of axioms.

Philosophers are more disposed to think about knowledge, perception, memory, and intelligence with a completely different set of analytical tools than us scientists and engineers who deploy evidence, reasons, justification, belief, certainty, and inference. Their insights and questioning is particularly germane to the current issues in the developments of artificial intelligence/machine learning.



THE ART OF
SCIENTIFIC
INVESTIGATION

by

W. I. B. BEVERIDGE
Professor of Animal Pathology, University of Cambridge

[illegible]

1957

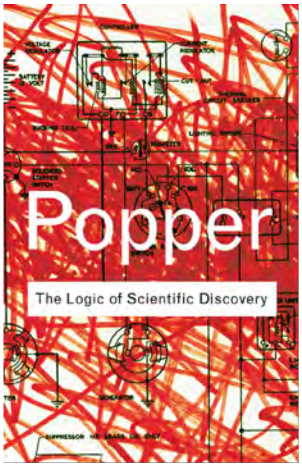
W W NORTON & COMPANY INC New York

Figure 5.2: W. A. Beveridge, *The art of scientific investigation*, Library of Congress 57-14582, W. W. Norton (1957).

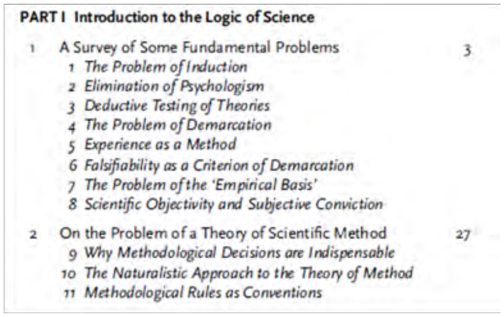
Let me step back to Beveridge, who Snow quotes often in his essay. He had a simpler message, a modified form of the class-room pedagogy with the different ways that one thinks through: imagination, intuition, and reason added in. I could add many more to this. But, this is all standard dogma.

Popper is further back in time, and is more grounded. He says scientists proceed by falsifying scientific claims. This by trying to prove theories wrong. Descartes, way back in 1633, wrote *The world*, offering an account of the universe, how vision worked, how muscles moved, how plants grow, how gravity functions, and how God got everything spinning in the first place. Positing this is enough to pounce and work on it. This is certainly true also for many hypotheses. It was specially so in in the past, and generally applicable in disciplines that still do not have sufficient mathematical and physical underpinnings. So, it certainly reflects quite a bit of even current scientific undertaking related to biological sciences.

Paradoxes and creating imagined conditions with contradictions is a very standard tool in mathematics-and-physics-oriented procedures. The 2nd law sprouted an industry of science, and it is still in-



1934


PART I Introduction to the Logic of Science

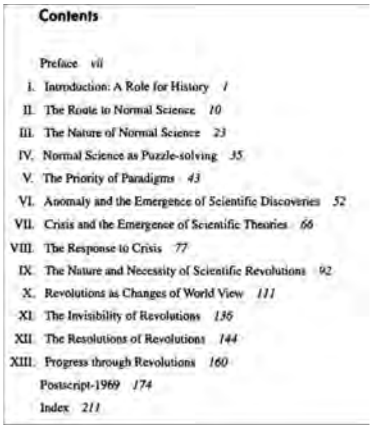
1 A Survey of Some Fundamental Problems 3
1 *The Problem of Induction*
2 *Elimination of Psychologism*
3 *Deductive Testing of Theories*
4 *The Problem of Demarcation*
5 *Experience as a Method*
6 *Falsifiability as a Criterion of Demarcation*
7 *The Problem of the 'Empirical Basis'*
8 *Scientific Objectivity and Subjective Conviction*

2 On the Problem of a Theory of Scientific Method 27
9 *Why Methodological Decisions are Indispensable*
10 *The Naturalistic Approach to the Theory of Method*
11 *Methodological Rules as Conventions*

Figure 5.3: K. Popper, *The logic of scientific discovery*, ISBN 0–415–27843–0, Routledge classics (1934).

tact. The reason being the vast unknown that the 2nd law represents. This is a point that Snow misses. This is the power and message of the 2nd law with its science and society meaning.

Thomas S. Kuhn
The Structure of Scientific Revolutions
Third Edition
1962
The University of Chicago Press
Chicago and London


Contents

Preface vii
I. Introduction: A Role for History 1
II. The Route to Normal Science 10
III. The Nature of Normal Science 23
IV. Normal Science as Puzzle-solving 35
V. The Priority of Paradigms 43
VI. Anomaly and the Emergence of Scientific Discoveries 52
VII. Crisis and the Emergence of Scientific Theories 66
VIII. The Response to Crisis 77
IX. The Nature and Necessity of Scientific Revolutions 92
X. Revolutions as Changes of World View 111
XI. The Invisibility of Revolutions 136
XII. The Resolutions of Revolutions 144
XIII. Progress through Revolutions 160
Postscript-1969 174
Index 211

Figure 5.4: T. S. Kuhn, *The structure of scientific revolutions* , ISBN: 0-226-45807-5, U. of Chicago, (1962).

Kuhn believed that scientists work to prove theories right, exploring and extending until progress stops. The Pasteur's quadrant, Pasteur having been in three of these quadrants where Kuhn introduced demarcation lines was interested in understanding the basic, yet also in driving control to solve issues.

A university engineering researcher's interests may fall between finding things out and using things. Enhancement of the knowledge and the utility. It is not likely that many single individuals fall within the Pasteur cell since both basic and applied science are highly specialized. Thus, modern science and technology employ what might be considered a systems engineering approach, where the Pasteur cell consists of numerous researchers, professionals and practitioners to optimize solutions.

Between Popper and Kuhn, we have two very different scientific temperaments. For Popper, scientific inquiry is a process of disproof, scientists are the disprovers, debunkers and destroyers. Many working scientists are like Popper's vision of a scientist, but there are many more spending hours and hours, people with will, performing brutal work to learn. There is a single mindedness bordering on being inhuman. Much of science work is boring. Generating data and mining the data.

Kuhn thinks of true believers who promulgate wisdom until a paradigm shift is needed, that is, a painful rethinking of assumptions arises.

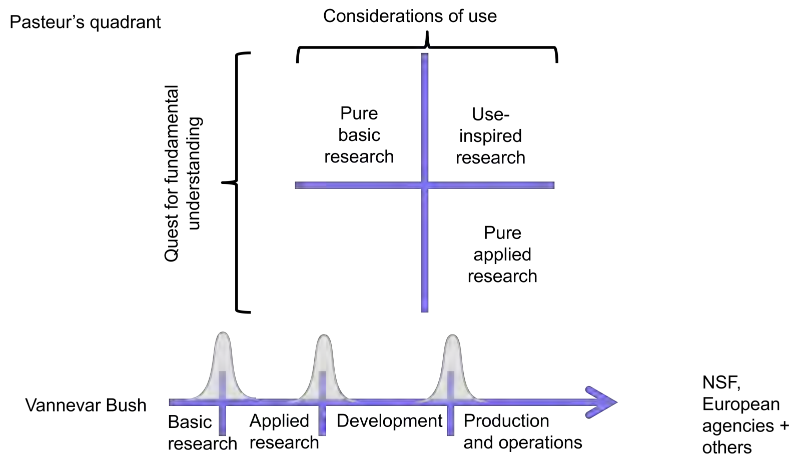


Figure 5.5: Pasteur's quadrant and its mapping in funding as implemented b Vannevar Bush in creation of National Science Foundation in 1950. Hard boundaries have acquired spreads, but this model is still intact seven decades later. Other variations, with the same basic construct, exists around the world.

An unfortunate consequence of the Kuhn halo has been the linear model of funding agencies that started with Vannevar Bush and *NSF* post-*WWII*. Basic advances are the principal source of technological innovation. This model gets tweaking from time to time, acquires spreads and distributions, but this linear form still remains. One model cannot fit all. Linear probably is right for development to production. But discoveries and new insights jump. Sometimes they take for ever. Indeed the most momentous science-society discovery and invention are entirely nonlinear jumping in a multi-dimensional phase space.

Take for instance, another linear-like model, the Wallace-Darwin theory of adaptation. When the big changes happens, a non-linear event such as when the asteroid strikes the earth or we create climate change based catastrophe, what one sees coming out on the other side is a *mutation*. Linear adaptation will not work. This is what a paradigm shift is. It is not Kuhn's paradigm.

Another current example is the sudden rise of *AI/ML*. As a name this subject started in a Dartmouth 1956 workshop. Nathaniel Rochester, one of the workshop's main personality (along with McCarthy, Shannon, Simon, and other lumaries) was the architect of *IBM 701*, and among the earliest to explore neural networks on computers. He would visit our Bldg 801 from *IBM* Cambridge Labs to make sure that the lab work with many early successes in theorem proving, playing games, et cetera, continued. *IBM* had to be careful in public. These were times when customers did not want to hear about electronic brains supplanting them. Computers can only do what they are told to do had to be the messaging to the outside world. Joseph Weizenbaum at *MIT* already had *ELIZA* interacting in Q&A with human beings to pass the superficial Turing test. But, it is only 50 years later that suddenly this field has become transformational in constrained technical domains—superior to humans—and buzzy in broader domains because an average human is really not that difficult to fool. Market came to a point that despite all the large foibles in large language models, it is big business. Real use of *AI* is not yet a big money number. It has been an augmentation tool in use for a considerable time now.

Market drove the research. Unlike the other science fields. There is nothing Kuhnish in this giant change taking place right now.

Note also how suddenly the world and how we interact with the world has changed with wireless, smartphone, internet, data data everywhere, in this world, and now *AI/ML* feeding on it.

These are mutations of the type that Marshall McLuhan called, ‶Medium is the message.″

A book that affected me very strongly when it first came out, was my father's copy of Koestler's *The act of creation,* where he posits ‶bisociation″ as the coming together of two unrelated thought streams—matrices—to a new form. Comparisons, abstractions, categorization, analogies and metaphors being some of the mixing tools.

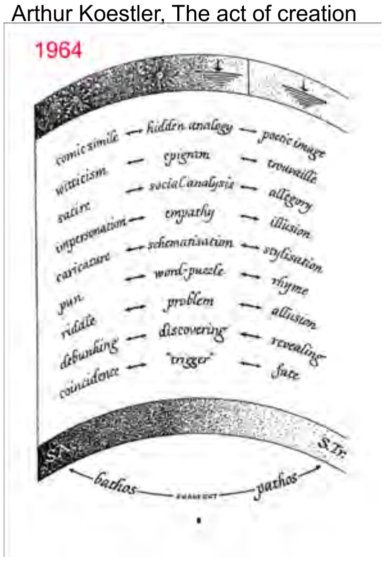


Figure 5.6: A. Koestler, *The act of creation*, Hutchinson & Co. (1964)

A joke, for example, in the form of a bait-and-switch mixes and delivers something new. A parody is imitation for illuminating effect. Being realistic enough that it initially tricks readers into believing one thing, only to make them ‶laugh at their own gullibility.″

In science, the two streams may fuse and synthesize a new path. Eureka is the Archimedes example. In arts, often it is a juxtaposition, with both sustained.

Take, for example, Edgar Degas, who says ‶On voit comme on veut voir; c'est faux; et cette fausseté constitue l'art,″ that is, people see what they want to see, it is false, and this falseness constitutes art.

Figure 5.7: Ballet (ou l'Étoile) by Edgar Degas (Musée d'Orsay).

Ballet is a perfect metaphor of life, we are the young ballerinas, and then there is the man behind the curtain in the dark suit.

Another beautiful set of examples are the the three paintings from van Gogh—all made in 1888 as he struggled with his inner demons—of his bedroom in Arles. Colors—not many of them realistic, reflections from thick and thin paint patches, wide and fine brush strokes, but nothing of sharp precision that old paintings used to have draw your attention to the post of the bed and from thereon to the simple beauty of life. Powerful impressionism, where the viewer is forming his own based on her or his path in life.

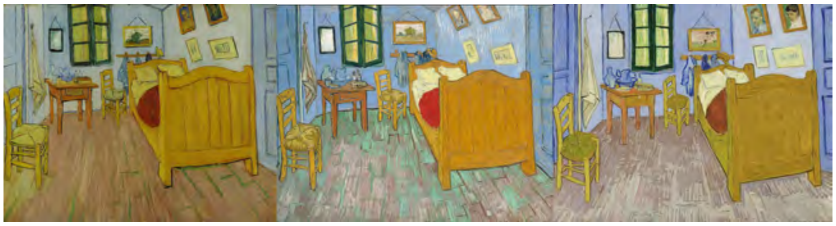


Figure 5.8: The three *Bedroom in Arles* of van Gogh painted in 1988. The first is at van Gogh museum in Amsterdam, the second at Institute of Arts in Chicago, and the third at Musée d'Orsay in Paris. I have seen them, but never together as here. There is a mood swing through the colors.

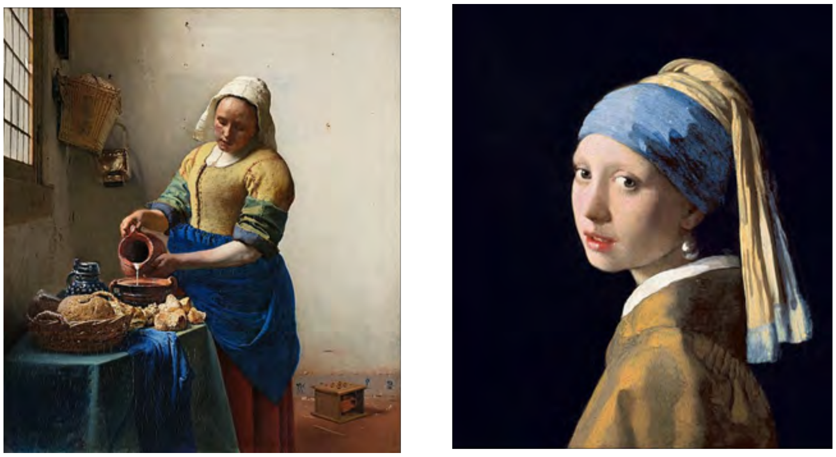


Figure 5.9: *The Milkmaid*, at Rijsmuseum in Amsterdam, and *Girl with a pearl earring* at Mauritshuis in Den Haag.

Or we may go back in time to mid-1600s to *The milkmaid* and *Girl with a pearl earring,* of Vermeer, another great Dutch painter. One is of working life pulling you to the milk stream and from there to the hard-worked face, and the second is of a rich girl with a focus above and to the side of her right eye and yet there is this pearl calling your attention. Again beautiful colors as with van Gogh, but here is the lighting and specially the use of perspective that all the optics—Hugyens is living during this time and the post-Kepler revolution is in full steam—learning from science came and stood together.

Since ancient Greece, it has been clear that best thinking is crossdisciplinary. One needs to knit together insights from poetry, music, drama, philosophy, art, mathematics, natural sciences.

Remember Tagore's *Jodi tor daak shune keoo na ashe, tobe ekla chalo re.* It applies to science and engineering too.

We all understand that a flourishing intellect is a well-rounded one. Perhaps this is why scientists rebel and have hobbies.

Arthur Schopenhauer's ideas on time and representation go back to early 1800s.. Space and time are at the heart of both the 2nd law and the theory of relativity. The influence of these ideas in Einstein's development of general relativity are well recognized. Einstein, since his youngest days, had a clear appreciation of philosophy. Einstein had apparently even read Kant's *Three critiques*—of pure reason, of practical reason and of the power of judgment—when he was 15. The Einstein-Podolsky-Rosen paper on entanglement, written in 1935, is a beautiful expression of this learning, centered as it is on what is reality, what is locality, and this is now nearly hundred years later the starting point for much that we do in quantum computing and communications.

Schopenauer was among the earliest of philosophers that posited—in conflict with Kant—that the world is not a rational place.

Philosophy can provide methods for producing new ideas, develop interesting perspectives and certainly help with critical thinking. Philosophers have tools and skills that scientists are not trained for but need. Examples are conceptual analysis, attention to ambiguity, accuracy of expression, looking for and finding gaps in standard arguments, coming up with new perspectives, spotting conceptual weaknesses and the search for alternative explanations.

So do writers as virtual philosophers. Hesse says in the essay *My belief*, ``The fact that my Siddhartha puts not knowledge but love ahead of everything, that he rejects dogma and makes the experience of unity the central point,″ is something I can agree with with love to include love for finding things out, as Feynman would say.

Similarly Somerset Maugham, who was an obstetrician and drew on his experiences in the London slums and the poorest working-class people, writes, ``I was in contact with what I most wanted, life in the raw.″ Later in life, he recalled the value of his experiences: ``I saw how men died. I saw how they bore pain. I saw what hope looked like, fear and relief; I saw the dark lines that despair drew on a face.″

In early years, Hermann Hesse was drawn to Nietzsche's theory of aesthetics for a period, and forever by his use of language. The enthusiasm for aesthetics was replaced by a more general interest in Nietzsche the man and poet. Nietzsche in turn to Richard Wagner. These show in the writings.

Cultural disease is a common affliction in science and engineering. It is most obvious currently in engineering and computer science driven exploitation of us and our privacy to appeal to our basest desires. *AI* is no Übermensch.

Following the *WW1*, one can see the Nietzsche's theories of cultural disease and of decadence being explored in Hesse's mystical novels. This gives insight most of us have difficulty with. Hesse says to his imagined Nietzsche, ``Perhaps you seek too much. That as a result of your seeking you cannot find. My entire life was—and, for the most part, continues to be—about seeking and striving. I'm no Buddha, but it is still possible, even for men like me, to catch sight of him occasionally in others.˝

Artists and scientists perhaps seek too much, but it is a 2nd law journey.

This Hesse-Nietsche interlude points to me that the process of self discovery requires an undoing of the self knowledge, which one is assuming one has. Becoming is the ongoing process of losing oneself, a random restart—as in a Monte Carlo calculation—and finding oneself. `He who has attained to only some degree of freedom of mind cannot feel other than a wanderer on the earth—though not as a traveler to a final destination. For this destination does not exist,˝ said Nietzsche in *Human, all too human*.

This is so true in scientific pursuits.

I should end this metaphysical start with Weinberg's strong reservation that philosophy is more damaging than helpful for sciences. Although it might provide some good ideas at times, it is often something that scientists have to free themselves from. Some scientists, Stephen Hawking being one, even argued that the big questions have passed on from philosophers to scientists.

I don't agree.

A theory of everything can never be a theory of everything. Gödel has proscribed it. The best we can do is strive, use all the clever approaches we know, discover, pull the learning from different human endeavors and keep making progress.

It is the essence of life to be dynamic and changing.

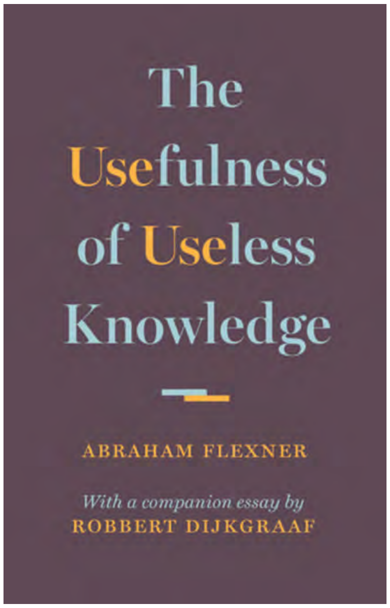


Figure 5.10: A. A. Flexner, *The usefulness of useless knowledge*, ISBN 978 0 691 17476 1, Princeton-Oxford (1939, Harper's Magazine).

Abraham Flexner, who headed the education group at Rockefeller foundation, and was the first director of Institute of Advanced Study, single-handedly turned the fortunes of American science in the pre-*WWII* period by being the quiet-in-the-background sponsorer of the escape of the numerous scientists from Nazis and Nazi-occupied Europe. He speaks Koestler-like, before Koestler, of the incredible importance of the usefulness of useless knowledge, which at some point in time are streams that come together leading to the blossoming of something entirely different from anything previously imagined.

To participate in science, you must produce the evidence to argue with. This is almost workman like. And it eliminates bias. It self corrects. But, it is also viewing questions through taste, personality, affiliation and experience. I was lucky in these.

I would like to pull this together to argue forward what I have promised as an objective of this essay on science and engineering pursuit.

5.3 *Information is the foundation*

My first thesis is that the world started with information. Sounds, singing, sentences, tactility, many of these things that make us came later. Trees and plants exchange information, it may be through roots, through pollen carried in the wind or by butterflies, through transfer of chemicals whether for defense or offense, and others. Fireflies communicate and dance. Organisms developed and mutated to selectively exploit the information. An information form was created. Information arises when something is assigned a significance in some way by a cognitive agent.

Information is an artifact.

It is a way of describing the significance for some agent of intrinsically meaningless events. We invest the stimuli with meaning. Without this investment, the stimuli is informationally barren.

Information however should not be confused with meaning.

Information is an objective commodity. Its generation, transmission, and reception does not require interpretive processes.

One can posit a framework for understanding how meaning can evolve, how genuine cognitive systems—those with the resources for interpreting signals, holding beliefs, and acquiring knowledge—can develop out of lower-order and purely physical information-processing mechanisms. The higher-level accomplishments that we associate with intelligent life carry a manifestation of progressively more efficient ways of handling and coding information. Meaning that the various constellation of mental attitudes that exhibit it, the interpretations, are all manufactured products.

Information is the raw material.

For those in electrical engineering, computer science or physics, this is a small jump. We think in terms of evolution of beliefs with information represented through Bayes relationship, graphs, flow charts, programs, and feedback loops.

For philosophers, this will be a big jump since they are disposed to think about knowledge, perception, memory, and intelligence with a completely different set of analytical tools: evidence, reasons, justification, belief, certainty, and inference.

Information is a semantic concept. What is knowledge? A traditional answer—an epistemic answer—is that knowledge is a form of justified true belief.

Knowledge is information-caused belief.

You need to know the day of the week, or what it was yesterday, to tell the date by looking at a calendar.

What one learns, or can learn, from a signal (event, condition, or state of affairs), and hence the information carried by that signal, depends in part on what one already knows about the alternative possibilities. This is a conditional probability.

Let me give another example. A modified form of the famous Monty Hall problem, but now with a million doors, behind one of which is a prize. A million tickets have been sold with different unique door numbers to us million people.

From an information-theoretic standpoint each of us, assuming this is a fair contest with each of us having an equal chance, is in the same position. The amount of information associated with my having a winning ticket is about 20 *b*s. The amount of information associated with holding a losing ticket is nearly zero.

The amount of information associated with holding a losing ticket is nearly zero but is not precisely zero. We don't have any other special information on the outcome, none of us has received any quantitative small piece of information, which per the present view of knowledge, is essential to knowing whether one going to win or lose.

The information-theoretic condition on knowledge has explained why nobody knows he is going to lose in this fair contest. Everyone is justified in being pessimistic, but no one has access to the information that would permit them to know they are going to lose. We live this Monty Hall environment throughout our life, albeit the random and the causal both present.

This is the same as the description in the state function of quantum mechanics. Only an observation leads to the eigenvalue and eigenfunction. Once there is an outcome we have gone from a prior to a posterior. Act of observation, an event in time, has reduced the system to a state about which we know. The past has now become classical and one has determined it.

What I also do not want you to do is to interpret this to be a form of reductionism to information.

Information has appeared in the natural world through a relatively simple ensemble of elementary ingredients obeying relatively elementary laws. Conway's game of life, Wolfram' one-dimensional automata have simplicity, chaos, complexity, birth, death, embedded in them.

The possible combinations of nature's elements, however, are stupefying in number and variety, and largely outside the possibility that we could compute or deduce them from nature's elementary ingredients. These combinations happen to form higher level structures that we can in part understand directly. These we call emergent.

They have a level of autonomy from elementary science in two senses. We can study them independently. They can be realized in different manners from elementary constituents so that their elementary constituents are in a sense irrelevant to our understanding of them. One does not need to climb down the rabbit hole of high energy physics to understand much that is emergent in our natural world. It would be useless and self defeating to try to replace all the study of nature with science. But evidence is strong that nature is unitary and coherent, and its manifestations are—whether we understand them or not— behavior of an underlying physical world. Thus, we study thermal phenomena in terms of entropy, chemistry in terms of chemical affinity, biology in terms functions, psychology in terms of emotions, and so on. But we increase our understanding of nature when we understand how the basic concepts are grounded in science as we have largely been able to do for chemical bonds or entropy.

It is in this sense, and only in this sense, that I am suggesting that meaningful information could provide the link between different levels of our description of the world.

We have all these physical notions of information from a scientific viewpoint. Shannon a demi-god of electrical engineering because of the numerous implications of this surprisal quantification into his equation(s). But what is meaning? Meaning I see as what is in the moment and the flow in infinite time. It is position and momentum. It is information in the Shannon, and Renyi, and Fisher, and all the others plus flow, which is evolution for the natural world.

Time and space are another interesting aspect. More on this some other time.

5.4 *Causality and autoanthropomorphization*

What I do, or much of what I do, is elaborately orchestrated by what I believe and want, by my intentions and purposes, by my reasons for doing the things I do.

Fred Detske in *Explaining behavior* discusses the following example to illustrate the importance of information from the perspective of psychology and philosophy.

Often when I move, I have a reason for moving. I get up from my computer desk to straighten my back and get a little exercise through moving or I go to the kitchen because I want a drink and I think I can get one there. If I don't have those reasons, if I don't want this and think that, I would not move. At least I would not move when I do, where I do, and in quite the way I do.

My lips, fingers, arms, and legs, those parts of my body that must move in precisely coordinated ways for me to do what I do, know nothing of such reasons. They, and the muscles controlling them, are listening to a different drummer. They are responding to a volley of electrical impulses emanating from the central nervous system. They are being caused to move. And, like all effects, these same bodily movements will occur in response to the same causes, the same electrical and chemical events in the nervous system, whatever I happen to want and believe, whatever reasons might be moving me toward the kitchen.

If, then, my body and I are not to march off in different directions, we must suppose that my reason for going into the kitchen—to get a drink is, or is intimately related to, those events in my central nervous system that cause my limbs to move so as to bring me into the kitchen.

What appeared to be two drummers must really be a single drummer.

But does this mean that my thoughts and fears, my plans and hopes, the psychological attitudes and states that explain why I behave the way I do, are to be identified with the structures and processes, the causes of bodily movement, studied by neuroscientists? If so, aren't these scientists, as experts on what causes the body to move the way it does, also the experts on why we, persons, behave the way we do? How can their explanation of why my body moves the way it does be different from my explanation of why I move the way I do? But if these are, indeed, at some deep level, the same explanatory schemes, then the apparently innocent admission that neuroscientists are (or will someday be) the experts on why our bodies move the way they do appears to be an admission that neuroscientists are (or will someday be) the experts on why people move the way they do.

If there is really only one drummer, and hence only one beat, and this is a beat to which the body marches, then one seems driven, inevitably, to the conclusion that, in the final analysis, it will be biology rather than psychology that explains why we do the things we do.

What, then, remains of my conviction that I already know, and I don't have to wait for scientists to tell me, why I went to the kitchen? I went there to get a drink, because I was thirsty, and because I thought there was still a cold water left in the fridge. However good biologists might be, or become, in telling me what makes my limbs move the way they do, I remain the expert on what makes me move the way I do.

Or so it must surely seem to most of us.

To give up this authority, an authority about why we do the things we do, is to relinquish a conception of ourselves as human agents. This is something that we human agents will not soon give up.

This is the conflict between two different pictures of how human behavior is to be explained. Reasons—our beliefs, desires, purposes, and plans—operate in a world of causes, and to exhibit the role of reasons in the causal explanation of human behavior.

My own interpretation in a different language of science is as follows.

It is information and its meaning that is connecting the reason people have for moving their bodies and the cause of their bodies' consequent movements.

My reasons, my beliefs, desire, purposes, and intentions, are—they have to be—the cause of my body 's movements. These are my evolutionary laws. They have developed over time and they came partly coded in at birth just as the number sense that morphed into an ability to add numbers, symbolically manipulate, that is, build algorithmic evolutionary laws, or of other symbolic characters to words and phrases and sentences to a meaning through a grammar, which is another example of evolutionary law.

Between information that I perceive, my evolutionary laws, and my prior beliefs is the definition of me. How I behave in different circumstances depends on causal and random factors. Thermodynamics and statistical mechanics has much to say about this based on information. The thermodynamic arrow points towards higher entropy and corresponds to irreversibility. It is a thermodynamic time, it is pointing towards the future, and it is the one we experience through our information-based mechanisms.

There is also a notion of epistemic time with an agential temporal arrow in that we know the past better than the future.

The cause to effect is also an arrow, we can act on the future but not the past. That we can affect the future and not the past, that any intervention must not violate past correlations. The laws of nature are reversible, but causation is not. Thermodynamic time clears this. Any intervention affecting the past generates thermodynamically inaccessible states, making them irrelevant histories. The biosphere's entropy gradient gives the time orientation to causation.

Our own thinking is dissipative, so again guided by the time-oriented entropy gradient.

It was at IBM's research laboratory that I started recognizing information's importance as we discussed different subjects—I used to enjoy sitting with people I didn't know or people I hadn't seen before—and of course Landauer's information is physical and Bennett's resolution of Maxwell's demon came up often. Even in discussions related to the observations of Mendelbrot's self-similarity fractals, Mandelbrot's table was always attractive for finding new things in mathematics. This information gathering, the discussions with many of the luminaries of that time in sciences who were always willing and enjoyed serious discussions, as well as a laugh, and dropping in the library to look at the day's intake of publications from various disciplines, was a constant source of meaningful information.

I started thinking about the conjecture—a corollary—that ``Physical is information,″ that I dwelt on in this past section during that time, and my belief has only strengthened. Over time I have taken to calling it my corollary. It is not proven, arguable in the same way as ``information is physical″ and ``it from bit″ of Wheeler are.

I realize that I have turned the human experience into an abstract form in this process.

But, I think it is important to realize that all the social phrases we use, upbringing, values, ethics, morality, providing a safety net, keeping our children protected in their childhood so that they can grow healthily, our need to explain, or explain away, our aspirations, our limitations, our path based on our starting point of the life trajectory, can all be seen through such an informational lens. It is agnostic and subject to proper scientific challenging.

5.5 *Science and engineering in society*

Societies deal with their problems based on political and economic expediencies, with some regard to cultural and moral convictions that have developed in the collective over time. These are all informational too. Politics is based on building vote banks, whose organization very much depends on what kind of political process is in place. First past the post or rank choice, even disenfrenchising by making voting difficult, and who controls the media, or if media controls the politics through its ownership. This is all information. Economics is numbers.

All of these expediencies affect the conduct of science. It is forced between guard rails. It is valued, nevertheless, because it does make a better-off life and development possible.

By spending time around the world, and observing the academics and the workings in Europe, at many of the best schools in *USA*, here at *IIT* and *IISc*, and watching how they fit in the social framework and in research and development, I would like to particularly stress what Germanic countries—my experience being in German and Switzerland—do so remarkably well. I see something similar in Scandinavian countries, and in a different form in France. The continental Europe, I have found, to be a remarkable place for learning and development of oneself as well as a young person's mind.

There are multiple reasons for this of course, including being reasonably well off, but there is the organization, the flow, and the stress—all seamless—that I find truly remarkable. This is what makes these countries far more equitable than *USA* can aspire to be. It breaks the perpetuation of class, provides those with unique capabilities born poor the same chance that all rich have, pulling up the brightest minds regardless of their origin by giving them access to the same demanding education.

In Germany, people don't go and live in communities where the school is better as they do in *USA*, or the rich don't send their children to private schools as in England or India or now increasingly so in *USA*. All schools are of similar high quality. The plumber's son, the doctor's daughter, and the teacher's child go to the same school, and they live together in the same community. Standards exist and students are expected to learn. Mathematics is rigorous. At the high-school-equivalent level, one has a choice, go to hochschule geared more towards professions, or university, which is academic. University is nearly free, but you have to meet standards as a student through the examinations, else you switch from the more rigorous courses. It is possible to change paths if you wish to too.

Then there is the research and development part of the federally-funded system. Max Planck Institutes that are simply the world's best research organizations, where the scientists are from all over the world. The process of choosing people is very clear. Hire the best from around the world to lead. And now that there is a record of past few decades of success, they do get the best. You just have to see the Nobel Prize lists of the past few decades to see that Germany has brought itself back from the decimation that its academics suffered during the Nazi period. Germany also has Fraunhofer Institutions that support bridging the gap between research and products, where academics and industrial organizations work together. Max Planck Institutes are pure research based institutions in different subject areas spread around with critical mass. The expectation from them is to be the best, bar none. Same for Fraunhofer in industry-oriented work. One doesn't see the caste system of pure versus applied, theory versus experimental, academic versus professional, or any other such division. The precision machining practice is as valued as the intellectual output of a bright mind. And the whole system is funded in a such a way that there is flow that is commensurate with the needs of the country. This is the German way.

With some tweaks and changes, it is also the way in Switzerland and Netherlands. France is different, it is a larger and more diverse country, but it too has found a way to promote research and good education, both of the elite world-leading intellectual kind and the hard-working analytic type, within its schooling system by keeping rigor of mathematics, reasonable salaries and respect for the teachers, and by placing many of the CNRS laboratories at universities that are good in that subject area.

This is information flow, and development, organized so that mutual information is maintained in the entire hierarchy. Not attacking the problem of the day and constantly tying knots. Conservation is conservation of flow for us in sciences, as I often stress to my students.

5.6 Checkov's last chapter

Science is objective. Scientists are not. We mine information. We have our own struggles and teutonic fights.

We breathe paradoxes and contradictions.

Paradoxes are a philosophers' tool.

Pseudo-randomness is a beautiful metaphor for how our own perception of free choice can emerge from underlying determinism as we navigate through the world. Just as a sequence of pseudo-random numbers appears freely chosen if you do not have access to the program and seed, so can human actions appear. This is the quandary I see answered in the Degas painting.

You hear the name Renoir, and you smile and see the whole world resolved into circular brushstrokes, rosy, bright, happy. You say ``Schopenhauer´´ and see this same world represented by lines of suffering men who during sleepless nights have turned suffering into a deity and who with solemn faces move down a long, rough street leading to an infinitely quiet, infinitely modest, sad paradise, so says Hesse in his essay *Language* written in 1917.

The best science listens keenly to philosophy, so must the best philosophy keenly listen to science. This has certainly been so in the past: from Aristotle and Plato, to Descartes and Hume, Kant and Hegel, Husserl and Lewis, Heidegger, the best philosophy has always been closely tuned into science. No great philosopher of the past would ever have thought for a moment of not taking seriously the knowledge of the world offered by the science of their times. Science is an integral and essential part of our culture. It is far from being capable of answering all the questions we would like to ask, but it is nevertheless an extremely important one.

The same strong statements can be made regarding poets, artists, writers, and others whose major tool is their mind, and through it they affect our being.

This is a mixing of a brew of complex ingredients, and whose result in turn is also complex, but with interesting emergent properties.

Its corollary, often posited as Ockham's razor, is that plurality should not be posited without necessity. This principle gives precedence to simplicity: of two competing theories, the simpler explanation of an entity is to be preferred. Given that we do not know the unknowns, just choosing the simplest theory does not satisfy me. Our axioms-based view is based on building the simplest edifice that is predictive. Explaining is the reverse from an observation. Inferences, predicting what will happen, however often fall into Bernoulli's fallacy, where objective probability interpretations lead to incorrect conclusions through a trust placed in *p*-values, even if false-positives and true-negatives are all being observed. This is the tragedy of several medications, where profits takes precedence over effectiveness and no-harm, and many social and economic policies.

I turned to ChatGPT to check this simplest maxim to see how a whatever-is-written-on-the-web is legitimate to get it to complete, ``*USA* detonated two nuclear bombs over Japan because ... .˝ This sprung the lines ``the US government made the decision to use the atomic bombs as a military strategy during World War II in order to force Japan to surrender and bring an end to the war. ...,˝ an Ockham-like simplest statement, which is terrifying for humanity in the lesson it imparts. Japan was bombed also because USSR troops were starting to amass in Vladivostok, the debate was on about the new world order, I hope also that the morality of the killing of innocents must have been a counter-factor, and several other reasons. To give precedence to one, and then it becoming a part of the indoctrination, is among the reasons that we we have the long arm of nuclear driving global wars and tribal rivalries pervading nearly eighty years later.

Mencken had a proper rejoinder to Ockham, ``for every complex problem, there is an answer that is clear, simple and wrong.˝

I went to an India-related question that I have always been interested in since Naipaul opened my mind's eye at a tender age of 9. The question of ``Did the participation of Tatas and Sassoons cause the indentured labor migration from Bihar?˝ I get a long stream of meaningless verbiage first by stating that they did not directly cause (correlation?), but that the conditions did. That Tatas and Sassoons did not directly recruit indentured laborers, and so on.

So, to understand the level of any depth here, I ask ``Why did the British indulge in Opium trade in 1800s through early 1900s?˝ The answer was `` ... for economic and political reasons.˝ Nowhere in the answer is any mention of the moral-compass-bereft ``intelligence˝ of killing two birds with one stone. Get the Chinese hooked on drugs, and the Indians enslaved for sugar farms all over the world. I was delighted a few years ago to read *The sea of poppies* by Amitava Ghosh as a writer's exploration of the dark ages for so many for so long in this country.

A computer science teacher will say, garbage in, garbage out. *GIGO* is a technical term.

To Mencken's rejoinder, I add the corollary, ``for every complex problem, there is an answer that is obfuscating, complex, and also wrong.˝

Of course, Tatas and Sassoons are complicit, and this is buried under the rug of all conformist writing.

This is lack of intelligence. It is lack of critical capacity to think, to state not only the case and what could be the case—a description followed by a prediction— but also what the counterfactual is. What is not the case, and what could or could not be the case. This is what constitutes an explanation, which is an indicator of intelligence.

That objects fall to the ground because that is their natural place, an Aristotelian view, raises a stream of questions. An explanation that mass bends spacetime is highly improbable, but tells us why the object falls to the ground. *Intelligence is thinking and expressing improbabilities*, which aligns with what I have argued about information.

Jorge Luis Borges says, ``In a time of great peril and promise is to experience both tragedy and comedy, with ``the imminence of a revelation,˝ in understanding ourselves and the world.

For science and engineering to do well for the society, it is important to always approach it openly, based on information sans bias, with equal opportunity for all. It is a magnificent cauldron, but we don't know which combination of ingredients is healthy and tasty and flavorful, and which is poison. So also proceed with caution.

In the sequence of earlier essays I have emphasized the description of state—of static and flow as integral to its description—and of evolutionary law in describing the dynamics. It showed up both in analytic forms as well as in graphs. As in physical description, where local and global effects happen, this state behavior description holds in real life too in local and global form. When young, we all want to change the world through actions unfolding over distance. Around my 50s, I turned inward-bound, and let global become largely a diffusive effect from local. Reduce, reuse, recycle, in that order is a common theme accepted by many people I admire. I have added to it two more to make five commandments as evolutionary rules. *Reduce, Reuse, Recycle, Rant, and Rile*. The first thee are self-actions in the state. The fourth is to let others—adults—know when one observes an unacceptable abuse, and the fifth is an escalation in egregious cases. The last two are a form of civil disobedience and my satyagraha.

6
Semiconductors: Lessons from the past and what it says for semiconductor manufacturing

That semiconductors have through devices, circuits, systems, computing, communications and information exchange made the modern world possible is a sound and arguable claim. But, we came to this point dynamically. New inventions, new technologies, new ways of attacking the information processing and transfer and its evolution to knowledge have all pushed this evolution. Wisdom, which follows, is very much a particular society's optimization—like a minimization constraint—upon which it may act (or not). The earliest computing companies, Burroughs to Univac do not exist as such, nor do those who brought about the minicomputers such as *DEC*, or microcomputers such as Sun; yet computing is the heart beat of the society. Semiconductor manufacturing is very capital intensive, and it demands experience and precision knowledge. Even for *USA*, the answer was focus on design and let *TSMC* build it. But, societal tensions or wars can intervene as one sees right now. This is a very serious issue for any nation. Semiconductors are like agriculture. One can not be confidently independent without the ability to build and deploy. In this broader worldly context, I would like to discuss commercial principles that have guided the evolution of the information enterprise, and look at the open big areas of the future, to speak to what needs to be the broader focus of design, development, manufacturing, and associated computational developments for the coming generations.

The last writing—of the K R Sarma lecture—argued that ``information is physical˝—a Landauer and Wigner thesis—and its corollary—``physical is information˝ as an agnostic lens to view nature and our world and the way humans—nature more generally—work through in the society. Information is an artifact that symbolizes some agent of intrinsically meaningless events. The higher-level accomplishments that we bandy as intelligent life carry a manifestation of progressively efficient handling and coding of information. Emergent properties arise in this progression, where understanding at one level does not need an understanding or description at a lower level. Atoms may form us and all the material objects around us, messenger *RNA* and other nucleic acid forms may be essential to the creation of the coding that builds life, but they are not necessary to explore cognition or cognition impairment, or poetry or so many of technological creations. Knowledge builds on the accumulation and distillation of information, information exists in all so many different forms and evolves in emergent ways, where the emergent behavior stands on its own and is explorable on its own.

I will try to build on this agnostic view to debate the topic of pursuit of semiconductors and semiconductor manufacturing in India. While this objective is specific to semiconductors, it exists in a global environment. One cannot look at it narrowly without looking at the whole.

Development takes places in an open system, one that is dynamic and has open boundary conditions. How a nation develops—not get trapped in some minimum of a generalized coordinate—is a matter on which all have opinions, and given that this is an inferential task steeped in sociological, economic, cultural milieu of the country interacting in a complex world, predictions always teach us how wrong we often are. Science, particularly through the use of statistics, and arts through its exploration of the human drama, both have something to say about limits to what we can infer. All that we can do is bootstrap from information one has. This is the only way to avoid sclerosis that is the affliction of many a societies.

Science, when done right, is a tool for limiting large errors. It limits these by drawing on information and power of testability of predictions in model building from the current collection of information. This writing should be viewed as one way of looking through information, which doesn't explain everything, but places bounds that constrain going badly wrong.

From a few transistors and resistors in the earliest integrated circuits from Texas Instruments and Fairchild, the state of the art today are the Nvidia *H*100 and Biren *BR*100 with nearly 10 billion transistor processors at the heart of much that is happening in the computation-based learning and inference world. The world lives off this information. It is pervasive. It is passing through the computers on to the desktop, it is in all the advertising we are bombarded with, it is behind the checkout counters in all the shopping, and it rules us through our smartphones that best embody McLuhan's, ``medium is the message.″

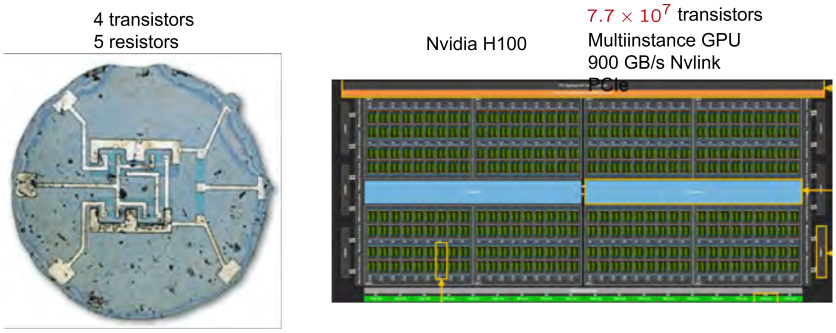
This evolution of semiconductors has been an epic journey, of ebb and flow, companies have come and gone, Fairchild does not exist and Texas Instruments is a specialty analog-digital automotive and

My wife Mari, who keeps me inbounds, remarked that preaching should be employed in discourse very rarely for it to be effective. It is true. Passion, while it should be appreciated and even promoted, needs to be inward bound. I will only add that in education and at an educational institute, some of the worldly constraints should be relaxed a bit. A Socratic dialog can only be successful with thought, rigor and passion in all the parties. This is in teaching and research too. Information is central to this, and our willingness to learn, recognize our past errors, learn from the change, and adapt is important to making sure that this flow process that leads to progress in all its forms continues and improves with time.

Bertolt Brecht, in his 1938 play of Galileo Galilei and the eternal clash between dogma and scientific evidence makes the statement, ``The aim of science is not to open the door to infinite wisdom, but to set a limit to infinite error.″

Bootstrap and its refinements is a statistical technique due to Professor Bradley Efron which draws inferences by modeling a resampling and inferences from the resampled data leading to a measurable quality of true sample from resampled statistics.

Fairchild was acquired by On Semiconductors, an offshoot of the microprocessor merchant Motorola, which also acquired *IBM*/Global Foundries' semiconductor operations in upstate *NY*.



http://www.computerhistory.org/semiconductor/timeline/1960-FirstIC.html

Figure 6.1: An early integrated circuit of few transistors (from Fairchild semiconductor) and a current advanced integrated circuit (from Nvidia made at *TSMC*) with 10 million times more transistors.

phone oriented semiconductor manufacturer. Even Intel, an offshoot from Fairchild, nearly went under in late 1970s and early 1980s, and is again tottering. This is a dynamic system with open boundary conditions in the midst of global competition and changing models for effectiveness as complexity keeps scaling up and new emergent behaviors arise. *Semiconductors are part of a dynamic system with open boundary conditions.*

Semiconductors have tremendous benefits that we all see through information's pervasiveness. But, as with all technologies, it has caused much by way of hurt.

A particularly damning ill effect has been the loss of social interactions of children playing with each other. The physical development and social development at a playground—a microcosm of young life—taught one losing and winning, learning from both, and how to find common ground in midst of the tensions of the social playground. An appalling recent statistic is that nearly 10% of *USA* high school teenagers have attempted suicide (www.ft.comcontent77d06d3e-2b9f-4d46-814f-da2646fea60c). This is isolation appearing in childhood, and all the addictive pressures brought by social media made possible by semiconductor technology.

The argument of this writing is in two parts.

The first part deliberates the world environment and the constrained open boundary system that a country is trying to grow in. It explores what the lessons are from the worldly environment, what the lessons are from within the country, and in a similar vein the lessons from semiconductors' and information's seven decades march to put down some markers. India's freedom and semiconductors' modern beginnings are nearly coincident making for an interesting viewing of a country and industry dynamic.

The second part than tries to connect this discussion to the choices for a path to a sensible future.

The information edifice that makes the modern society is built on the semiconductors. Semiconductors make the physical structure possible. Communications through the cell phones connected to the networking infrastructure, the reliability and safety and security checking of all communications-based transactions are all based on semiconductors. All the computing, the transactions one makes, the financial bookkeeping, the back-office Aadhar or *UPI* functions, even education through remote processes, are computing meshed with communications. Our gathering of a lot of data, much of this relying on sensors of different kinds—temperature, fog, rain, people moving, traffic flow, transport such as railways, et cetera—followed by decision making based on this data is all semiconductors based. A nation's defense requires monitoring, a quick reacting to the observations, controls, operations, radars, operation of weapons, et cetera. This too is all semiconductors based. No traffic—from railways to cars—would be possible without the semiconductors performing vital sensing and control tasks. In the west, cars, specially those with self-driving augmentation, are really semiconductors on wheels. Semiconductors are like agriculture. We cannot live without it anymore. It is essential now to our being.

In addition to agriculture, it is also like jewelry. Compactness means it can be hidden in decorous objects, but more so, it even appears as a status symbol. On the *IITK* campus, as I walk around, nearly a quarter of people seem to have either a phone near their ear or held in their hand. In the West, changing a model to the higher end model every year seems to be a marker of success. Strange, but people like Arnault and others in France and Italy have thrived for generations by catering luxuty to the upper echelons of the society. There is tremendous profit in luxury built on exclusivity. Between agriculture and jewelry, one has covered both ends of the social spectrum!

6.1 *The two marshmallows principle*

One of the most interesting human behavior and development study was the Stanford marshmallow experiment of Walter Mischel from the 60s and 70s. Marshmallows are sweet egg-based fluffy concoctions that can be partially melted on a fire. By placing it in between two sweet crackers, one has a cookie that never fails at a children's camp fireside gathering. In the experiment, 4 year olds were presented with two options. Ring a bell to call in the experimenter and eat the marshmallow. Or wait until the experimenter returned—about 15 minutes—and get two marshmallows. A reward now or a bigger reward if patient. Some children broke down and took the one marshmallow option. Some were able to delay gratification and got two. Longitudinally, when the children reached teenage years, the ones who had deferred showed higher *SAT* scores, were more self-assured, more self-confident and had better self worth. Their parents thought them more mature, better at handling stress and planning and reasoning. Later on in life as adults these two-marshmallow children were less likely to have drug problems, less divorce, were less overweight, and for each minute that a preschooler could delay gratification, they also had 0.2% less body mass index 30 years later.

I am going to call this the *the two marshmallows principle*. Foregoing short term reward for higher payoff in future really matters. Resist temptation to have persistent benefits across dimensions. I will add to it as my own personal experience that hard work itself, the learning journey and then reaching the end of a journey are all a pleasure unlike any other. Just talking ad infinitum is dysphoric.

The two-marshmallow principle holds lessons for individuals, it holds true for us in our work, it holds for us as a collective in institutions, it holds for our families, our countries, and our institutions. There is a *short-and-long, fast-and-slow feature* to this principle that I will return to later. Patience as a virtue can not be emphasized enough.

6.2 *The parable of IBM*

Story telling has value but should be used with caution. Business world thrives on story telling. It is also a convenient way to hide fallacies. A current example is the present financial environment and difficulties in the West, likely to be quite long term, that partly arose from technology which the technology venture world thrives on. Promises for the future are alway steeped in Brecht's infinite errors. The storytelling by venture capital and all the magical software-based companies of apps prospered because the price of money was none. This was intersection of two different domains of information, and their merging wasn't humanly understood since this intersection had no historical precedent. Fast moving of money by new apps—in private clubby social network environments—precipitated a crisis that was long in making in financial governance. Two endeavors intersected, venture business exploited it for nearly a decade, and then precipitated a moving event starting with the collapse of the Silicon Valley bank ending a long period of one kind of flow. This is dynamics at work.

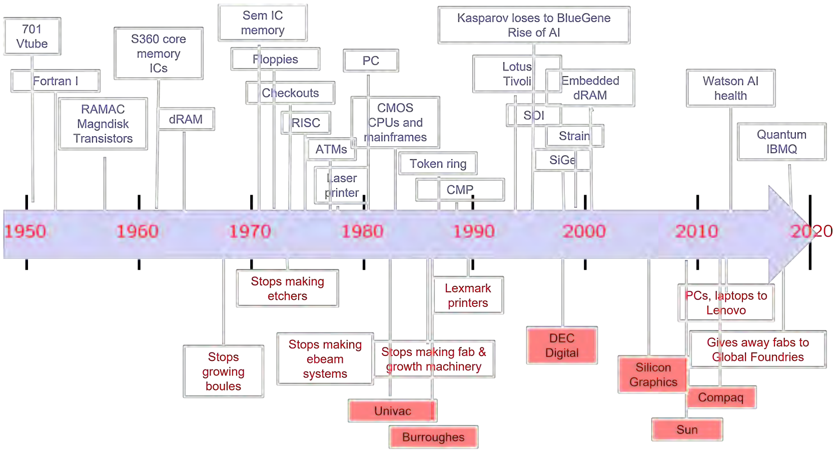


Figure 6.2: A world line of *IBM* in the world with semiconductor emphasis. The top half is for some important events of introduction of technologies, many invented at *IBM*, and bottom is when *IBM* stopped doing something, sold off businesses, or where competitors disappeared.

The story in short form that I want to relate is that of *IBM*, where I had the incredible luck and joy of working after schooling. Caution is warranted, the story is just to see lessons in it, and it is a personal perspective. To me it illustrates some of the junctures in time and space where choices have to be made because the existence and growing is of a dynamic open boundary system. As a company, *IBM*'s origins go back to Hollerith and punch cards in 1880s. Tabulating, clocks, and accounting machines was its business. It found its footing through the use of punch cards in census and the 7 3/8 × 31/4 *in* 80 rectangular holes punch card that became an industry standard. Its first computer, 701, was based on vacuum tubes circa 1952 a year behind Univac in its introduction. It flowered in the decades that followed. The first compilable high-level computer language, magnetic storage, and the first scalable computers with reusable programs (the *IBM* mainframe S360) all appeared in the next decade, and a formidable period for the company—not too different from that of Google, Apple, Microsoft, with their walled gardens—was born.

The blossoming was a combination of an incredible promotion and marketing machine, a company that served its customers well and provided a very sheltering umbrella for the best in the world to perform research, where research was these folks' main interest, with some of it flowing to the world. It was in mid-50s that *IBM* Research Laboratories were set up. There were innumerable technology advances, from the first scalable—same program running on all the machines—to magnetic drives, dynamic memories, reduced instruction set computing, to *SiGe* and others all appeared from *IBM*. Even the very first successes of machine learning/artificial intelligence such as geometric proofs, or the Blue Gene machine beating Kasparov in chess towards the end of last century are from *IBM*. Through this cycle one can see *IBM* being ahead of others in technology discovery, technology usage, and also getting out of tasks that became commoditized or were not necessarily central. It used to make much of the equipment used in technology. It iteratively stopped them. Any computing business that became competitive in time, where technology was now available through all, it left or had to leave. Laser printers or laptops or personal computers, and others. By 2020, even the entire fabrication operation—a capital-intensive and demanding task—was given away to Global Foundries. There is much much more of course behind the story.

For a long time, the freedom, with sound management practices of promoting excellence, made the research environment a frothy and exciting place where ideas were constantly bubbling. At the same time, as others started catching up, the same bane of large companies—do not rock the cash cow—also played, and this slowed *IBM*'s entry, even if the earliest work and inventions took place at *IBM*, into new areas. Reduced instruction set computing and not employing its own processors and operating system for personal computers eventually reflected in the breaking open of a walled garden in advanced computing at one end and personal computing at the other. Large size, and resulting bureaucracy and hierarchy slowed the decision making and the law of large numbers and averaging of

Emanuel Piore, chief scientist of Office of Naval Research where the early ideas of computers—von Neumann's computer was supported by them as was Eckert-Mauchley's ENIAC at University of Pennsylvania— came to place *IBM* on strong scientific footing. The laboratory was first at Columbia University, moved to Yorktown Heights, and then expanding to San Jose and Europe. During my time, I could watch Ralph Gomory—an applied mathematician, John Armstrong—an optical scientist, Jim McGroddy—a semiconductor scientist, and Paul Horn—a condensed matter physicist—as heads of research. Gomory, in particular, stood out in being able to stand up for bringing out ideas from the laboratory into the world by promoting them in face of special interests. Today's smart cities go back to traffic flow of 70s as an applied mathematics development in *IBM* Research. In writing Python codes, I constantly see the techniques and syntax that goes back to the programming approaches that came from that time. I noticed that the more you know a subject, it is certainly true that the more you see the potential, but also that the more is your bias towards the subject. Successes and failures both get amplified in consequent business decisions. By the 1970s, computing was going through far faster changes as it democratized, and ideas such as interactive computing pioneered by *MIT* Lincoln Laboratories and taken public by Ken Olson and Harlan Anderson through *DEC* at the beginning of this process—*UNIX* being a major long term software offshoot that lives underneath most of computing today—and so many others at the intersection of semiconductors and computing started taking over. The dictum ``Ideas escape from research,'' certainly seemed to prevail towards the end of my time.

The white collar culture, nobody can be fired for buying *IBM* in the hay days has the same whiff of white babu culture transforming to brown babu culture of at least during my first two decades of life that I spent in India.

distribution became ascendant.

This story is the story of a dynamic system and open boundaries constantly at play. *IBM* still exists more than hundred years later, perhaps will even prosper again as it changes. Over time, most of its competitors have come and gone. *IBM* is not a media and advertising company though that is apparently where most of the money in computing is made these days. It may very well be a wise decision. Serving other businesses with technology and expertise to make their own business succeed is a more robust long-term business than a business that is steeped in extracting profit through monetization of customers by advertising and invading their privacy. The latter is primed for missteps and supplanting. Free access to invade privacy and to create automated profiles through artificial intelligence is also a seed for failure if people walk away and one is left with a very expensive infrastructure to support. *IBM* has faced this often, when computing approaches change, and old legacy systems need to be continued to be supported because they are central to the business. This is often the reason for the need for renewal. If is easier for business-to-business processes. Advertising companies come and go.

The oldest longest-extant small company is the Italian arms company Baretta founded in 1526, that is, nearly five hundred years ago. Violence never goes out of fashion.

The message from this experience is that dynamics and open boundaries are pervasive through all human endeavor, as individuals and as an enterprise. Vibrancy comes, success comes, and is reinforced through conservation of flow, of flow of bright young people streaming in, new ideas growing and uprooting old dogma, opening and exploration of new territories, and that the lessons of the *two marshmallows principle* must be constantly remembered.

The world changes around us, we must change, interesting problems move outwards under a strong entropic force, and so must individuals, companies and countries.

6.3 *India in sepia and in the world*

I offer here a narrow information-based view of India now—my own reading of it and its evolution within the world seventy five years after independence. This is now several generations following the freedom. People are better off than in my times when decisions were so often based on how to eke a living. There is more self and passion today.

But nothing is vertiginous. Progress happens through confluence of a flow where nothing is a bottleneck and the entirety is limited by the slowest part of those inputs that constrain. This could be due to skills, or due to lack of capital, or due to poor communication or transit infrastructure, or due to the time-scale of processes, and so many other factors. It all depends on the particular technology-based

Personally, in this extended period of living in India following long periods of absence, it is quite evident that this is a very different country from the one nearly five decades ago when I left. It has enormous self belief, it does not think anymore in reverence of white sahibs and brown sahibs even if *VIP* culture is still persistent, it does not look at western publications such as Financial Times or Wall Street Journal or The Economist or institutions such as the the World Bank or International Monetary Fund as the fountain for wise prescriptions having realized that they are all self centered. India stands for its own interests as it sees fit in the world context and through its own experiences of post-freedom period. This is very different from my times till 1976. There is no submissive non alignment but an assertive non alignment.

social uplifting task at hand.

Growing and rising is a task replete with challenges, and it is in this midst that one must view the question of semiconductors. Development and growth is a flow that is across the entire chain. People, ideas, development, products, social structure, transit, inter-country relationships for trade, time, et cetera all matter.

I offer the following informational view.

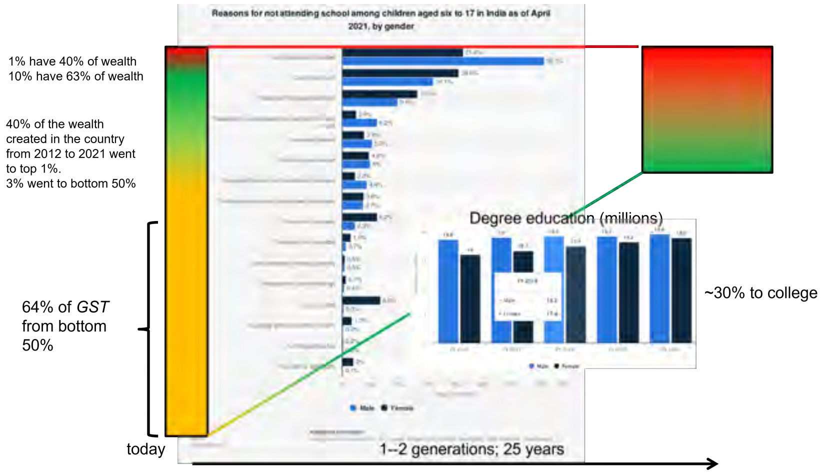


Figure 6.3: Wealth and taxation and representative issues of education today. I have added to it an idealized projection of how the wealth across the populace needs to be in a developed nation which is not an unfair idealized goal a century after freedom.

Today's India (Figure 6.3) has 40% of the wealth in the top 1% and 63% of the wealth is with the top 10%. 40% of the wealth that was created in the last decade went to the top 1% and only 3% went to the bottom 50%. It is the bottom 50% that pays 64% of the goods and services tax, which covers everything involving daily necessities. This is very incongruent. It is a system that is making the money flow to the rich rather than the larger populace.

For population at large, education is the major tool for improvement and for getting out of their low-income circumstances. But, the education system is such that a large number of young do not make it past the primary school for reasons of cost, interest, poverty, quality of schooling, and others. A skilled workforce requires college education, whether academic or vocational. Yet only $\sim$ 30% of college-age Indians are in college. This does not account for what the colleges are teaching, and the quality of that education. If one wants an educated India where everyone has a chance to be creative and productive and satisfied, this spread must compress, and it must compress in a generation.

One can also assess the proposition that all-are-created-equal in democracy through what is happening to females, who are usually the ones left behind in societies, most societies being male dom-

Measures for of incongruency are, of course, non trivial. But a ratio of top earning to median earning and top to the lowest earning within any grouping—a company or a nation or other subsystem—is not a bad start. Science tells us that a factor of few Euler's constants is a sign of nonlinearity and not normal. This is the fallibility and fungibility of humans. We may be able to tackle clean water, enough food, decent medical care, but then take on the harder things, love, safety, aspirations, natural kingdom, and that bar always seems to be so high.

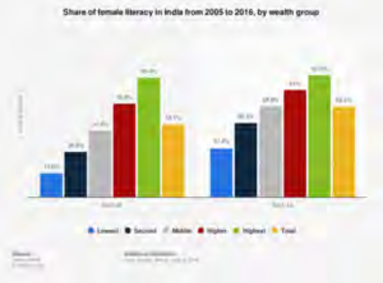


Figure 6.4: Female literacy by income group and changes over a decade.

inated. The poverty- and access-driven paucity is reflected in the share of literacy (Figure 6.4) . The poorest have only 1/3rd of female literacy. This is not equality of access, or democracy at work, or a nation that is putting sufficient effort to the most important factor—education—in a community's growth.

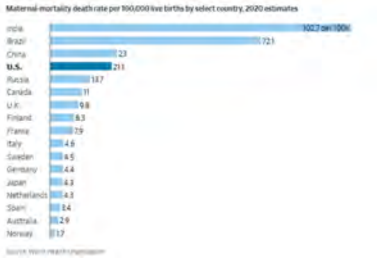


Figure 6.5: Maternal mortality in select countries of the world.

Another good measure, besides education, for people-oriented governance is healthcare, and a representative measure is maternal mortality (Figure 6.5) which is a gauge of caring for mothers as well as caring for the new born. India is factors of ten or more worse than developed countries, particularly European countries, which tend to be more people centric. Not on this chart, but even a poor country such as Costa Rica has one-fourth of the maternal mortality of India. China, about as large, has a mortality of 1/5th of India.

So, to understand the broader outline of India's development dynamics, the underlying statistical measures of how money is spent, I chose to build a cohort group that was at about the same state of development at the time of India's freedom. A viewing of this also gives a few example instances of what may happen given the if-then type questions, as well as one that is fascinating. What does one do and what the characteristics are that cause growth to happen but then stagnate before one is on equal footing with the developed nations. This is the *income trap*, which can be at low income or even mid income, and one that the emphasis and the nature of the country and its interactions in the world are likely to determine.

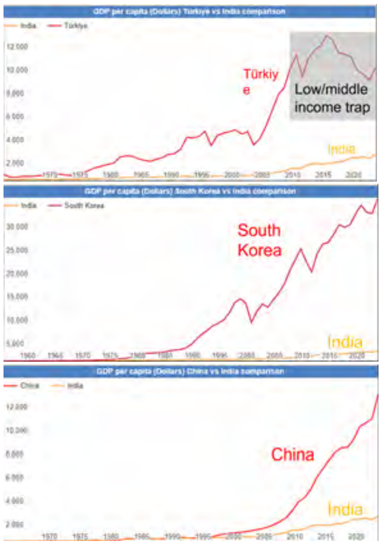


Figure 6.6: A comparison of *GDP*/capita for Türkiye, South Korea and China with India over a major period following India's independence to modern times. Source is countryeconomy.com.

The cohort group (Figure 6.6) is Türkiye, a country that Ataturk democratized following the fall of the Ottomon empire, China, which was under quite strict Communist governance following Mao Zedong's overthrow of the Kuomintang, and South Korea, which appeared as a nation following the Korean war, and was often under military dictatorship for a few decades.

A number of observations can be made, a few of them I find particularly interesting and remarkable. The cohort groups all show two points where rapid change occurs. There is an initial rise, this is followed by a plateauing, and then another rise takes place. *There are two inflection points.* India seems to have had its first, but not yet its second. So, an interesting change to look for ii about now and should be observable in a few years. With the deglobalization efforts underway in *USA*, the slow ending of dollar at the center of all finance, and the rise of a multipolar world, the prognosis is one of good chance.

South Korea has continued to grow getting to nearly half of European standards, with the changes taking off in 1980. This is the time when its steel industry and ship building became competitive with that of Japan. Soon after, in the next decade, it embarked on its semiconductor—primarily through memories—mission. Gradually the other areas: chemicals, materials, telecommunications, and now biology have become ascendant. This is the second inflection of South Korea. For China, the first changes were once Mao's cultural revolution ended, the change was slow like India's early decades. With Nixon's opening up to China as a counterfoil to Soviet Union, Deng Xiaoping's ascendancy, and the focus on becoming the manufacturer to the world—starting with the Shenzen experiment—came the first clear inflection. The next inflection is around 2008 when the western world was trapped in a financial crisis brought about by the capitalist easy money and boom-and-bust, not dissimilar to what is playing out right now, while China started graduating from manufacturing to higher-value creation by corporations that designed and built their own and started competing with the Western cohorts on equal footing. China keeps moving on and there are many lessons in this that we must explore since India's trajectory looks so far as one that is twenty years behind China's.

Türkiye took off in the 70's, and again has a double inflection around the 2005 time frame. So far, for these examples, one sees noticeable change taking place for all these countries around 2005. This is most likely globalization and *USA*'s use of lower labor countries for low-end manufacturing, from plastics to small appliances to clothing, with different categories important for different countries. South Korea and China however have kept growing by bootstrapping to higher value industry. Türkiye does not seem to have. This is a low/middle income trap, where now for nearly a decade-and-a-half there has been bouncing around and flattening. Türkiye now also has significant inflation.

This raises an important question of what causes countries to fall in an income trap, where after reaching a level of income, the growth stops. What makes or breaks growth of countries. Any country wishing to be advanced with its entire population enjoying a comfortable humane life needs to avoid the trap by understanding its causes.

6.4 Leitmotifs, self incarceration and the income trap

In science there is an important concept enshrined in the phrases *short-and-long, fast-and-slow, and heat-and-work* that speak to the flow in the midst of complexity. The concept holds for many dimensions. Its behavior is most immediately visible in time. Short, or fast, is like nitpicking or scattering, where lots of little events here and there cause unpredictable outcomes and hinder movement and flow. Like walking in crowds versus an early-morning or late-night walk without the crowds. The short avoidance/bouncing events are fast events, are frictional, and they slow one down. It is heat that takes away from the ability to do useful work. Actions have consequences. Causes have effect. Sometimes we react. Sometimes we act. Sometimes we are impulsive. Sometimes we are deliberate. Fast and slow is a Kahnemann phrase. Gut actions are a system 1 quick action based on past experiences. They are programmed in. Slow is deliberative, where we think through, work through stories from which we learn, and then act. They are analytic. Short is an immediate effect. Long is an effect that plays out in time. What looks right in short can be terrible long term. Scattering is a fast short phenomena. In walking through crowds with a lot of fast scattering, there is much heat and very little productive work in the long term. Sometimes short, fast and heat are useful too. For example, if one is walking, trying to cross the road, and suddenly a red car comes roaring down, one must quickly step back. There is no thinking required, no working through what if questions. These are all examples of various connections between short or fast events and long and slow effects. Long effects linger. Drop a stone in the pond, a quick event, and that wave spreads out for a very long time and has an effect that can be felt far far away.

The income traps are such a long term effect. Looking at the world, in European region, which by and large are well-off nations, one can see a number of countries, Türkiye that we are discussing, but also Greece and Italy that are stagnant. Each of these three countries has tourism as a major industry. Tourism is a service industry, it provides an employment that is at best low and middle income for most of the participants except the hotel owner or the tour operator owner. On top of this Türkiye is in the midst of secular-religious cultural fights and a lower key conflict with Kurds. Greece has constant right-left debates, changes of the government every so often, and has still not outgrown the angst left over from the second war, the civil war, and the military dictatorship that followed. Italy seems to have similar leftovers of fascism and tourism is now broadened to gastronationalism, with the latter contributing nearly half of the *GDP* of nation.

Eric Hobswam says that ``When a community finds itself deprived of its sense of identity, because of whatever historical shock or fracture with its past, it invents traditions to act as founding myths.'' The Italians have invented there's, the English too have their coronations, royal soap operas, the Northern Ireland union centered sagas, the Turkish folks have the Armenian genocide, Kurds nation, and Islamic-secular kerfuffles. These are all sources of friction, the short, and all they do is cause heat with nothing productive at the end. They become the alter ego of the income trap.

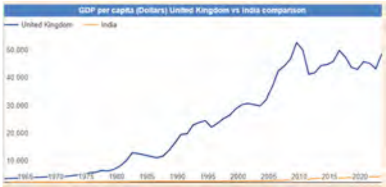


Figure 6.7: UK and India comparison of *GDP*. Source is countryeconomy.com.

To the test the thesis of living in the past and fabrication of identities, UK provides a good test opportunity. Sure enough, it has been in a middle income trap (Figure 6.7). Like Türkiye for the past two decades. UK has not been able to shake off the loss of the empire and outgrowing its easy looting of the wealth of other people. Manufacturing is not appreciated and has mostly disappeared except for some remnants in Rolls Royce airplane engines. The class, feudal, and financial chicanery that helped it rule and impoverish giant nations are now its own problems whose most ill effect has been Brexit, an issue that is a constant—a short— irritant everyday in the society. It is trapped in its own jail of beliefs and deprived by its faux empire identity. This constant fluctuation witnessed in politics, Brexit, meddling in other nations, London as the sole focus because of its finance as a background in the daily life is very unproductive. It is these self incarcerations that have left these countries not striving to rise but to just duke it out within constantly.

There is a lesson in this for India. Amartya Sen lauds the constant Indian discussing and debating rather than doing things by calling the people as *the argumentative Indian.* Practiced within limits, the debating and rational civilized fighting is productive and leads to proper course corrections. Unrestrained, it is just producing heat, and doing everything useful inefficient through that friction. It slows one down and it also slows others down.

For progress, the simple normal things that one needs to live should be free of friction and taken for granted so that the energy is placed in useful work that is needed past the daily civil human existence.

This is my lesson from Eric Hobswam and a look at the statistics of nations. The trap is is largely tied to cultural factors. There are events and traditions that one cannot get over since they have been so ingrained since they are part of inculcation in family, in schooling and by religion. All the people involved in past injustices may be long gone, but injustice still live today and in turn keeps continuing the tradition of doing injustice. This is the invention of traditions and creation myths and they become leitmotifs. *It is representative of failures rather than success.*

As an example, in Indian context, while it is true that as recently as 1700, India accounted for about 24% percent of global *GDP*, not different from what what *USA*'s is or Europe's is or China's is today, while India's is only 3%, let one not forget that the governance was in decay. In the north part of India, there was steadily increasing extraction from the populace by Delhi, Akbar and others included, in the midst of insularity, excessive religion, poor education and modernization, while Europe was undergoing the science and art renaissance. The perpetuation of power and glory meant creation of

I should add here some current Indian context from my experiences. I have benefited tremendously through others who have helped me wade through bureaucracy of the financial systems. The systems have undergone tremendous improvement with robust safety measures, less corruption, and is now straightforward for those with smart phones. As a visitor, I have had to rely on others after struggling on my own for a few days. You need local credit cards, local bank accounts, local phone access to manage. It makes sense for daily living and honest governance. But I find phones slow me by interfering in my own attempts at productive work. I am also a privacy fanatic and protect myself from Googles and Apples. So, perhaps it is my own self-built trap. But, where I stay on the campus, just across the institute wall are are two temples that sometimes blast all night long, and regularly attempt to indoctrinate in evening and early morning. I can not figure out how children can get good sleep and be prepared to learn the next morning, or how can the parents be at their best at work. Religion should be a personal belief, not a source of irritation to others. This is short and long, friction, and inefficiency in India the same way as it is in Türkiye or Northen Ireland or Israel. Not solved, it is going to lead to a trap.

zamindars, the impoverishment of the landless, while Delhi hid itself in music and poetry, whose eventual culmination was the British, the opium trade, the indenturing—a euphemism for slavery, emigration, and all that Bankimchandra and Sharatchandra and Premchand so eloquently talk about. This is the reality of flow. And of course one can trace this farther and farther back of which unfortunately less and less is likely to be what was real.

Myth building helps avoid really facing information, so one should look at some of the statistics to understand what the bottlenecks in growth can be.

	Türkiye	South Korea	China	India
Annual *GDP* ($*M*)	817,508	1,797,810	17,744,640	3,176,296
GDP/Capita	9,654	34,744	12,564	2,257
Debt/*GDP* (%)	41.8	51.33	68.06	89.18
Debt/Capita	4,036	17,968	7,164	1,704
Deficit/*GDP* (%)	−3.86	−0.02	−9.72	12.76
Expenditure /Capita ($)	3012	9046	3726	588.5
Export/*GDP* (%)	26.15	35.61	18.97	12.46
Education /Capita (%)	395	1487	347	**56**
Education /Budget (%)	12.41	24.98	11.45	12.75
Health/Capita ($)	291	1214	337.9	**19.9**
Health/Budget (%)	9.69	13.42	9.07	3.38
Density	108	515	147	428
Life expectancy	75.85	83.5	78.08	70.15
Population	84,680,273	51,736,000	1,412,360,000	1,407,563,842

Table 6.1: Financial comparisons, important ones normalized, in some important categories for cohort comparison. The numbers are either from 2021 or 2022, and the sources are countryeconomy.com and www.statista.com

Returning to India, and its policies, the potential-of-trap issue, and of flow with respect to the cohort, and United Kingdom, issues that have been highlighted, Table 6.1 provides some of the important financial statistics of the four countries that are being compared. For India, with a *GDP*/capita factors of 4 or more smaller than the other comparisons, it is not surprising that similar ratios hold for expenditures, that is, funds going to infrastructure, railways or defense, or interest payments, but, the ones connected to the two items I called out as central to the trap arising in the flow and well being are education $ per capita and health budget $ per capita. These are even more than the reduced factors of other categories. These are factors of 10 worse for health per person and it is not surprising that education and health outcomes are abysmal, and it is quite a stretch to imagine that a vibrancy comparable to even the poorest of Europe can be achieved in a decade and a half—of the order of a generation—for the country. Furthermore, it also implies the flattening following a spurt unless this is rapidly corrected.

Compared to the cohort, a more suitable place to draw lessons and emphasize the relationship between emphasis and outcomes, is to compare to China. It is a big country and till the 70s, the comparison between China and India was similar for most categories of development. China today has established itself as a leading technological nation bootstrapping beyond its origins of change in the lower value manufacturing. The technological change could not have been possible without the simultaneous development of the entire flow system.

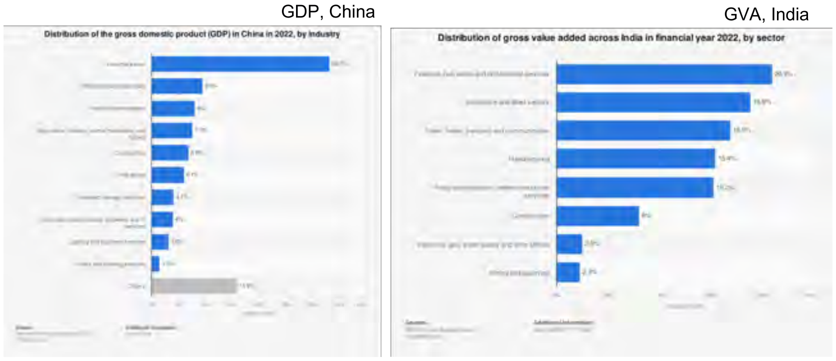
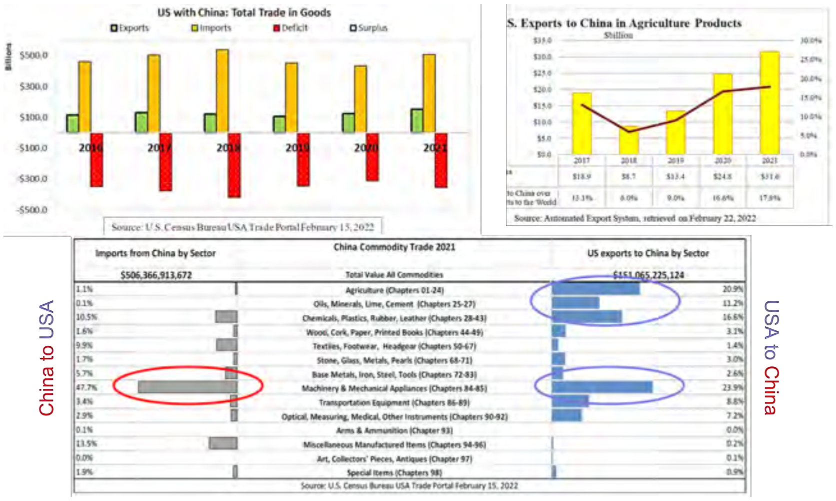


Figure 6.8: China and India comparison of the sources of gross domestic and value products. *GVA* (gross value added) adjusts the *GDP* (gross domestic product) and also shows the impact of subsidies and taxes on products.

China's wealth is far more from industry, trade and finance (Figure 6.8). India's is broadly across a number of efforts. Manufacturing, the highest value undertaking in terms of a real product of real human enterprise, is 1/3rd for China, but only 1/7th for India. For India, finance, services, and real estate is 1/5th and agriculture is just a little less. Real estate building and agriculture are low pay poverty employment, and that is where much of the employment of the nation, that of the bottom half is with which the sepia tale started with. Of course, wealth generation and the flow towards higher income is weaker since India's emphasis is still at the low end of the monetary contribution chain. The education deprivation keeps this flow highly constrained and will continue to do so as a Catch 22.

China is also instructive when one compares it to *USA* to probe its success and its the ability to keep the flow intact (Figure 6.9). China exports more than 3-times more than it imports from *USA*. Nearly half of these exports are machinery ranging from electrical, tools, construction, to other higher-end factory floor machinery. Only about a quarter of imports from *USA* to China are machinery. *USA*'s other major items are agriculture—a low value added product—and materials such as oils to cement. China exports items that bring more profit and higher wealth to the nation. So many of *USA*'s exports are



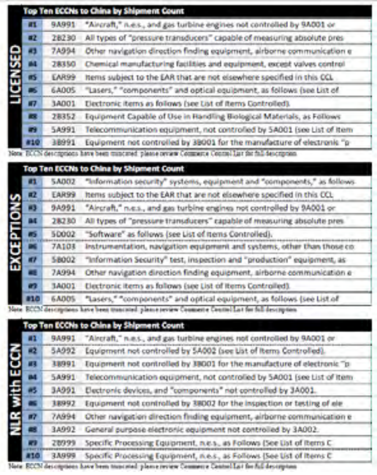
Imports from China by Sector	China Commodity Trade 2021	US exports to China by Sector
$506,366,913,672	Total Value All Commodities	$151,065,225,124
1.1%	Agriculture (Chapters 01-24)	20.9%
0.1%	Oils, Minerals, Lime, Cement (Chapters 25-27)	11.2%
10.5%	Chemicals, Plastics, Rubber, Leather (Chapters 28-43)	16.6%
1.6%	Wood, Cork, Paper, Printed Books (Chapters 44-49)	3.1%
9.9%	Textiles, Footwear, Headgear (Chapters 50-67)	1.4%
1.7%	Stone, Glass, Metals, Pearls (Chapters 68-71)	3.0%
5.7%	Base Metals, Iron, Steel, Tools (Chapters 72-83)	2.6%
47.7%	Machinery & Mechanical Appliances (Chapters 84-85)	23.9%
3.4%	Transportation Equipment (Chapters 86-89)	8.8%
2.9%	Optical, Measuring, Medical, Other Instruments (Chapters 90-92)	7.2%
0.1%	Arms & Ammunition (Chapter 93)	0.0%
13.5%	Miscellaneous Manufactured Items (Chapters 94-96)	0.2%
0.0%	Art, Collectors' Pieces, Antiques (Chapter 97)	0.1%
1.9%	Special Items (Chapters 98)	0.9%

Source: U.S. Census Bureau USA Trade Portal February 15, 2022

Figure 6.9: China and India comparison of the sources of gross domestic and value products. *GVA* (gross value added) adjusts the *GDP* (gross domestic product) and also shows the impact of subsidies and taxes on products.

in categories that are commonly to be found in exports from emerging countries, with agriculture and minerals and other resources from land dominant. Transportation—aircrafts being one of the major categories of export for *USA*—is less than 10%. The physical work occupations and products dominate.

One can perceive in these statistics the reason for tremendous worries in *USA* of it being supplanted by China. So what can *USA* do? Its actions over the past few years is to increasingly throttle the high-end export (Figure 6.10) where all the wealth comes from. This is aircraft engines, sensors, transducers, some chemicals, lasers, some telecommunication, nearly all the different things related to information and how efficiently can people move to do things, so aircrafts. There is a broad set of categories, nearly all related to higher-value technology products and their manufacturing, and of nuclear industry.



LICENSED

Top Ten ECCNs to China by Shipment Count		
#1	9A991	"Aircraft," n.e.s., and gas turbine engines not controlled by 9A001 or
#2	2B230	All types of "pressure transducers" capable of measuring absolute pres
#3	7A994	Other navigation direction finding equipment, airborne communication e
#4	2B350	Chemical manufacturing facilities and equipment, except valves control
#5	EAR99	Items subject to the EAR that are not elsewhere specified in this CCL
#6	6A005	"Lasers," "components" and optical equipment, as follows (see List of
#7	3A001	Electronic items as follows (see List of Items Controlled).
#8	2B352	Equipment Capable of Use in Handling Biological Materials, as Follows
#9	5A991	Telecommunication equipment, not controlled by 5A001 (see List of Item
#10	3B991	Equipment not controlled by 3B001 for the manufacture of electronic "p

Note: ECCN descriptions have been truncated; please review Commerce Control List for full description

EXCEPTIONS

Top Ten ECCNs to China by Shipment Count		
#1	5A002	"Information security" systems, equipment and "components," as follows
#2	EAR99	Items subject to the EAR that are not elsewhere specified in this CCL
#3	9A991	"Aircraft," n.e.s., and gas turbine engines not controlled by 9A001 or
#4	2B230	All types of "pressure transducers" capable of measuring absolute pres
#5	5D002	"Software" as follows (see List of Items Controlled).
#6	7A103	Instrumentation, navigation equipment and systems, other than those co
#7	5B002	"Information Security" test, inspection and "production" equipment, as
#8	7A994	Other navigation direction finding equipment, airborne communication e
#9	3A001	Electronic items as follows (see List of Items Controlled).
#10	6A005	"Lasers," "components" and optical equipment, as follows (see List of

Note: ECCN descriptions have been truncated; please review Commerce Control List for full description

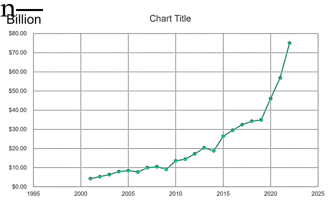
NLR with ECCN

Top Ten ECCNs to China by Shipment Count		
#1	9A991	"Aircraft," n.e.s., and gas turbine engines not controlled by 9A001 or
#2	5A992	Equipment not controlled by 5A002 (see List of Items Controlled).
#3	3B991	Equipment not controlled by 3B001 for the manufacture of electronic "p
#4	5A991	Telecommunication equipment, not controlled by 5A001 (see List of Item
#5	3A991	Electronic devices, and "components" not controlled by 3A001.
#6	3B992	Equipment not controlled by 3B002 for the inspection or testing of ele
#7	7A994	Other navigation direction finding equipment, airborne communication e
#8	3A992	General purpose electronic equipment not controlled by 3A002.
#9	2B999	Specific Processing Equipment, n.e.s., as Follows (See List of Items C
#10	3A999	Specific Processing Equipment, n.e.s., as Follows (See List of Items C

Note: ECCN descriptions have been truncated; please review Commerce Control List for full description

Figure 6.10: Export control categories from *USA* to China. Source: www.bis.doc.govindex.phpcountry-papers2971-2021-statistical-analysis-of-u-s-trade-with-chinafile .

Taiwan of course finds itself in the middle of this historic fight. The foundation of information technology in all its physical implementation—semiconductors—is at the center of this tension. *TSMC* is the world's global foundry (Figure 6.11). 1/3rd of world's silicon chips, designed by others, come from it. This includes for nearly all of the products from Apple, particularly its mainstays of iPhone and the Macs. It has 13 foundries. From its formation in 1987—many of my *IBM* colleagues in silicon technology efforts went back to participate and to start other companies that feed or feed from semiconductor manufacturing and have had a satisfying life—it has grown to the current dominance. It has 65,000 employees, so it is a high value $1*M* per employee output. All this started from a humble beginning by an enterprising Morris Chang leaving Texas Instruments, one of the



USA: CHIPS Act is roughly $280B

Figure 6.11: *TSMC* revenue over its lifetime. data from https://companiesmarketcap.comtsmcrevenue .

semiconductor integrated circuit inventor company. Texas Instruments still exists with $21*B* per year output. Intel, for comparison here is $62*B* revenue.

The formation of foundries was a brilliant idea, not as much appreciated as it should be, for breaking open semiconductor technology access to clever circuits and hardware designers who could now build and forge their own directions unconstrained by the large companies. This was the blossoming that led to rapid progress in communications, such as of the incredibly powerful modern mobile phone, as well as in computing, where new architectures and new approaches could be started and implemented. Qualcomm is a supremely successful example of the former and Nvidia of the latter. Qualcomm had $42*B* and Nvidia $27*B* in revenue in recent year. Neither builds any semiconductors, but all the cellphones and all the internet search infrastructure, and supercomputing, video, and other infrastructure all rely on the work from these companies. The foundry manufacturing made it possible. Only Samsung as a foundry appears as a competitor to *TSMC*. It too is next door to China. The semiconductor technology has become so complex that Intel is now generations behind and *USA*.

6.5 *Dirigisme, walled gardens, and commanding the high end*

China with its 1*B* users and $6.6*T* digital economy—much larger than *USA*'s—is also where much of *TSMC* output goes, since all the large-and-small hardware builders use China as their integration and assembly source. This includes Qualcomm and Nvidia. For *USA*, this makes China the bogey man. The Chips Act initiative makes $280*B* of funds available in an attempt to rejuvenate the semiconductor industry in *USA* with strings attached to minimize exploitation of the funds, which is the norm all the world over when a pot of money becomes available.

For any government policy in an industry that has been science- and technology-based ideas driven, and capitalism as a freedom driven system, this is an enormously powerful instance of dirigisme.

Dirigisme has many different faces. China has always practiced it. And now *USA* is recognizing it. But, evidence of its existence in the West, and specially in the *USA*-China supremacy contest has been around for quite some time. The Huawei episode in Canada, where Canada arrested the Huawei official at *USA*'s behest, and later Huawei's banning from so many countries is dirigisme in practice in the various forms.

In science and engineering, one can make best guesses based on current knowledge, but in the end, it is only the result that tells one

Dirigisme is a directive role by the state. In capitalism, the state usually only plays a regulatory role. Many economies—Japan, India, UK, European union, China, nearly all emerging countries—have been practitioners of dirigisme. It can work in social balancing objectives, but in technology, where ideas are the most important element and dynamic changes are constant, the success possibilities reduce substantially at the higher value end by state interference. New technologies are invented by competitors, choices can be wrong, support means picking one over the other, and this in turn reduces competitiveness. As the semiconductor industrial knowledge became more pervasive, even as *IBM* shed many of its factory efforts, its reliance on New York state support, New York being interested in maintaining employment, became an albatross. Decisions that should have been made, and clever new options followed, were differed. It makes the pain far more intense when the finality arrives. *IBM* inventions in spun-off environments would have been major businesses in their own right. Thermo Electron Corporation and other companies have shown how spun-off enterprises can flourish and yet help the original organization.

if what one pursued was good thinking or bad thinking. This is what Bertolt Brecht was telling us in asking us not to be too confident in science. That science can reduce possibilities of errors, but that it cannot eliminate them. The memory of a period in 1980s, when *IBM* was still doing remarkably well, is still intense with me. Everybody, including *IBM*, were very scared of Japan's 5th generation project. This was an effort of massive proportions—Fujitsu, *NEC*, *NTT*, Toshiba, etc. —and the state working together to make the most powerful supercomputers that will eliminate the status quo. Fujitsu in particular was a worthy competitor to *IBM*. In the project, lots of very special purpose chips were designed, and so on. Meanwhile, out came *UNIX* from Dennis Ritchie and Ken Thompson at Bell Laboratories working on *DEC* computers, *VLIW* practiced on specialty microprocessors from *IBM*, and the distributed and massively parallel systems era started. It was not long before one could even assemble a supercomputer in the basement of one's house should one so desire and had the knowledge. All the prominent Japanese computers are now gone out of computing as such. Dirigisme failed. But dirigisme certainly has succeeded for China. Dirigisme also helps bootstrap in known technologies that are still useful. It is in the highest end, where the idea itself is dictated by the policy where it is subject to causing a catastrophe.

A lead in technology, therefore that a strength in the highest value efforts that build a nation's wealth, along with other power dynamics, underly these *USA*-China internecine conflicts.

The US emphasis on *TSMC*, Taiwan and the Chips Act, Huawei and others, all the export controls over semiconductor manufacturing and design tools, the potential ban on Tiktok, et cetera, are all examples of the important role that semconductors have in the most important technology of these times. Global power of course is another.

Why semiconductors and where exactly is the wealth from semiconductors so that one understands what that entails in pursuit of higher value and growth of a nation?

At the simplest level, of course the wealth is in the higher end. *TSMC* may be a $75*B* dollar company, but it makes many other companies of the same size possible through the design and hardware chain built on top of the technology. This answer is fallacious in that it is the country of the higher value product that is gaining. Taiwan, in some ways, is in this trap and at the center of a rivalry, where what it has is central to the well being of others. This tension, of course, is also its safety valve. *It becomes a country worth defending and protecting.*

The lesson here is that the highest value comes from building and then controlling a *fenced garden* of some essential or addictive human need.

IBM did that in the past with mainframes. When that world became only a small part of the human landscape of information usage, others adeptly stepped in. Apple now is a master of a garden with its iPhone and the *iOS*, and with the laptops, et cetera, and the mac*OS*. Google is a master of its fenced garden with Android for phones made by many manufacturers, and of Chrome, and of the Cloud infrastructure it has built by providing free access to it to the masses. Microsoft, starting with Windows, going through a weak period, now does the same with Azure and Cloud that forcibly pull people into its walled garden. Facebook used its social network skills to rope people to its universe. Looks free, is useful, what is there not to like? WhatsApp works free. Apples and Googles can extract 30% for allowing every loaded software into their garden, and then extract a cut on every transaction conducted through it. Facebooks, Apples, Googles can bombard you with advertising exploiting your psychological profile learned by spying on your transactions and your ``free storage.″ Amazon can do the same in its selling, and since it knows what sells for all the sellers on its website, it can even undercut and put the small merchant ouf of business by creating its own equivalents. This is *the top of the value chain.*

Top of the value chain is very profitable, but also filled with serious moral, ethical, and social hazards. Dirigisme has a very proper place for tackling the latter as India's electronic transactions edifice has been by protecting the meager profits of kirana shops from disappearing into the companies at the top of this food chain.

Amazon	$514B/yr		Alibaba	$135B/yr	
Apple	$388B/yr		Huawei	$110B/yr	
Google	$283B/yr		Baidu	$20B/yr	
Tesla	$82B/yr		BYD	$52B/yr	
Meta/FB	$117B/yr		ByteDance(TTok)	$58B/yr	
			Tencent(WeChat)	$81B/yr	
Twitter	$4.4B/yr	328M users	Sina(Weibo)	$2.1B.yr	340M users
Nvidia	$27B.yr	H100	Biren Tech		B100
	$27B.yr	7.7×10^{10} trx			7.7×10^{10} trx
		7 *nm*			7 *nm*

Table 6.2: Major *USA* and China information-based company comparison.

An interesting view of the *USA*-China, the former has a GDP of $23*T* with 330*M* people, the latter of $18*T* with 1.4*B* people, can be seen in some of the important information-centric companies comparison of Table 6.2. For each of the dominant company in one of the technology-utilizing area, there is at least one competitive company based in China. Some *USA*-based companies are significantly large, but the Chinese companies are starting to bite at the toe, and the *USA* environment monopolist, while China's is not, so also an interesting dynamic. This is one piece of the technology-underlying tension.

The second is in the controlling, that is, the walled-garden part. In this advanced technology-value-control dynamics, the current situation in a few different domains is

- Large cloud providers: Amazon, Google, Microsoft, **Alibaba Cloud**,
- Dominant desktop OS providers: Microsoft, Apple and various *Linux*,
- Dominant mobile OS providers: Google and Apple,
- Chip companies: *TSMC*, *Samsung*, Intel, *Global Foundries*,
- Design companies: Nvidia, *Broadcom*, Qualcomm,
- Electronic design automation software: Cadence, Synopsis, Mentor Graphics, Avant!,
- Social networks: Meta, Whatsapp, Snap, **TikTok, WeChat**,
- Car companies: Tesla, *Hyundai, Toyota, VW, Mercedes, BMW*, **BYD**, **Gili**, et cetera,
- Airplanes: Boeing, *Airbus*, **Comac C919**, and
- Networking hardware: *Nokia, Ericsson*, **Huawei**, Cisco.

Hard work and speed can overcome much adversity. Many countries find this out in wars. Economic war is another one of similar wars where a nation's place in space and time is at stake.

The bold lettering is for Chinese companies, and the italics are non-*USA* companies. Certainly *USA* still controls through the standards a fair domain, but China is starting to attack the top of this value chain, aircrafts included, and for control, as all this recent Tiktok fuss, or the Huawei fuss has been about. China, a communist country, has more internal technological competition than *USA*, which is dominated by monopolies. It is building dominance in electric vehicles. Huawei and *SMIC* created one of the most advanced semiconductor technology node process that even Intel is still trying to master to overcome the *USA* sanctions. Germans better beware in the automobile industry.

Figure 6.12 shows in billion Yuans, the market valuations of the major technology-centric commerce and hardware companies of China. There are many more information-exploiting companies in China. Competition is alive in a Communist nation. Controlling a market gives large money monopoly. For example, Amazon, Google,

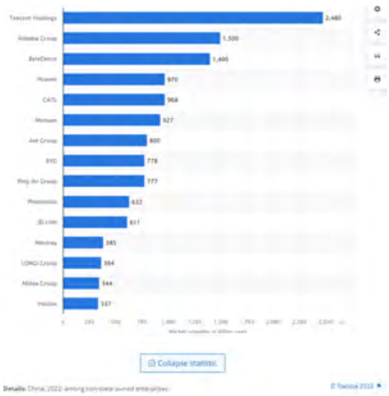


Figure 6.12: The revenue of major Chinese technology-centric commerce and hardware companies in billions of Yuan. Source: www.statista.com.

Facebook, Snapchat, Pintrest, just five companies, had $380 billion in advertising revenue in 2022. This is one aspect of Capitalism that does seem to stand in contradiction with the belief of free enterprise. It is a notion that we will revisit since it is one that dirigisme can address.

Nevertheless, collectively portrayed in this description is a trend that there are very serious changes afoot and that the world is now at a major turning point in the post-WWII order. For every *USA* company, there exists at least one, if not more, Chinese companies that can compete. They can extract high value and not leave it with the merchants that some of the organizations in this collective are. Much of the value in information infrastructure draws on system dominance, hardware dominance, and access to semiconductors as Figure 6.13 shows for shareholder return . The top 7 sectors draw their return by building on semiconductors.

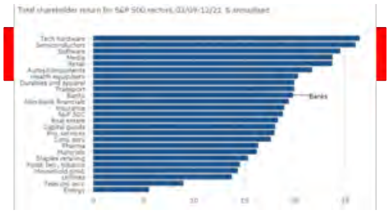


Figure 6.13: Returns of the semiconductors-dominated and -based information structure. Source: www.ft.comcontent939e819e-8381-4fee-8639-439847a196b3.

This collective has lessons for India. When and if semiconductors come to India, India will have to do this to exploit its semiconductor advantage and to protect itself from dirigisme of others.

We viewed *TSMC* as the $75*B* revenue company. Figure 6.14 shows the revenue changes over the years of semiconductor companies of the world, with the list including some who design but get their hardware built. The contributions of semiconductor manufacturers and designers keeps rising. I would suspect that with the difficulty of the practicing of the technology so large, this trend will not change, and barring world calamities, Samsung, *TSMC*, ST Micro, and others are well placed.

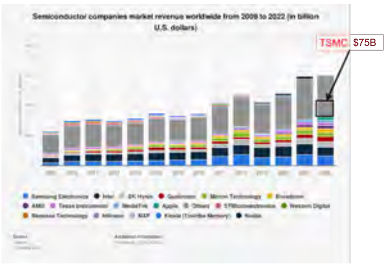


Figure 6.14: Market revenues of semiconductors and semiconductor designing companies of the world. Source: www.statista.com

India can join this list, but I am going to argue that it should not join them as just a manufacturer in the *TSMC* sense, but as an ecosystem, meaning one that has other higher value wrappings around the semiconductor.

6.6 *Free at last: Design and open software*

Dirigisme is a help as one is climbing a value chain and building accepted and immutable technologies. It does not work once one is at the high end and needs to break one's own path. There are a number of issues buried in this and let me discuss my view through example discussions.

Consider the first level up from semiconductor manufacturing. The design of the circuits that form a subsystem of a hardware. This involves tools of design, testing, verification, et cetera, but more importantly complex design where timing, heat, energy, power, noise, robustness, speed, compatibility, fitting into other people's designs of something else of which the design may be a part, and so on, all matter. Design is not a trivial undertaking. Lot of software, lot of intense knowledge and experience comes to bear when one needs to provide robustness, security, speed, low energy, and all the other design constraints. It is accumulated experience and it needs good team work. It is also pervasively needed. No semiconductor is useful without a good working design, be it in the cloud, or the edge of the network, or in your hand. This is where quite a bit of money is made sitting and designing and testing to get things right, with *TSMC* or Samsung making it for you.

It happens in two ways. One is where one is designing special-purpose systems, where the design is very integrative and much more specific to the task. The second is where one can share designs because of the commonality of the tasks being performed.

AI/ML-oriented designs fall in the first bracket. NVidia appeared as a graphical processors maker of the early video processing engines so that young people could play their games of blood and gory and glory. Pretty soon, it was realized that this streaming way is also very useful in fast high end computing such as those needed in supercomputers for nuclear, or biological, or other such problems. *CUDA* as a way of programming was born. Cryptocurrency mining—an incredible fooling of our society similar to the other addictions—helped. The processors started becoming mainstream. Nvidia kept improving and these are the processors employed in *AI/ML* tasks. NVidia is far far more valued than Intel.

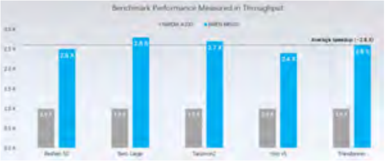


Figure 6.15: A comparison of *BR*100 versus the previous generation Nvidia highest-end processor used in *AI/ML* tasks. Source: HotChips'22.

But, here too, with good effort, one can supplant. Nvidia's competitor in China is nearly as good. Biren Technology is a company founded in Shenzen in 2019, does fabless design for *AI* and high performance computing. Its highest end processor design is BR100, made in 7 *nm* using 7.7×10^{10} transistors. All of these employ large amount of fast static random access memories (many 100s of *MB*), and flow architectures for speed with large bandwidths within the chip and to the chip. The total assemblage dissipates 550 *W*. It is just as powerful as the latest Nvidia highest performing chip. If Biren can keep getting its chips, China will do fine despite Nvidia not being allowed to sell H100 to China under the new export controls.

I also remember a particular *IBM* angst during days of competition with *DEC*/Digital. *IBM*'s technology was 3 generations ahead, but *DEC*/Digital performance was equivalent through cleverer design. They employed dynamic circuits. *IBM* used old static approaches with re liance on transistor shrinkage rather than also iterating design.

Never underestimate human ingenuity and ability to cleverly improve. Indeed in design and technology, it is when one is under the severest of constraints and stress that one produces the best designs. When life is easy, and one can depend on just working on improving one aspect of the design—for a long time, it was just the size, later node, of the transistor, with transistor improving every generation—one inevitably produces poor designs of a complex system with many properties interacting.

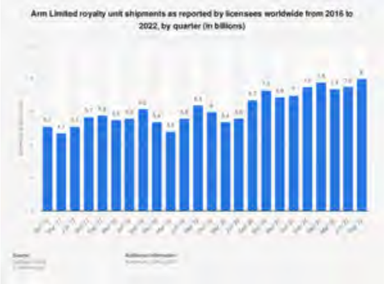


Figure 6.16: *ARM*-licensed shipments in recent time.

The second design approach's example is reuse and employment of standards across many companies. For not-so-high-end, these ideas too can flourish despite the walled-in gardens created by so many companies. Mostly, these are low-end or medium-end designs being used in large numbers in all the electronic automation one sees all around. *ARM* (Figure 6.16) is pervasive right now. Nearly 70% of the world non-memory semiconductor designs are based on *ARM* currently, nearly 80 *B* designs per year. Most of these go to China. *ARM* is used from cloud to edge, with *CPU*s in the middle. All the Macs, iPhones, the Chrome machines, et cetera, are *ARM*-based designs from Apple and Google. Google's cloud too uses a lot of *ARM*-based designs from Google.

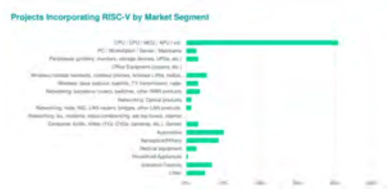


Figure 6.17: *RISC-V*-based design spectrum for different applications.

The change and opportunity is in *RISC-V* (Figure 6.17), which is a new open instruction set architecture accessible to all with no giridisme interference. Plenty of designs are already available on Github. About a third of current chip projects in application-specific and field-programmable, processors, controllers, et cetera, employ at least some *RISV-V*. Europe is strongly with *RISC-V* designs. There are plenty of non-high-end designs that happen in Europe. This is a slow-and-steady replacement of *ARM*. When I was at *ETH*, the entire floor of the large building where I sat was occupied by Prof. Benini's students and post-doctoral fellows—must have been at least 50—deep in the task of *RISC-V* designs. He is not alone as China seems to have adopted *RISC-V* to free itself of controls. 10 *B* chips, half of them from China, a volume that is about 1/5th of *ARM*'s was *RISC-V* based in the past year. Chinese academy of sciences is on a 6 *mo* cycle of upgrading designs. Among the major design firms are Starfive and SiFive that make a variety of *CPU*s (from low end to high end), Alibaba—like Google—making its own custom built processors, even porting operating systems such as Android on to it, Ali Pingtou, another Alibaba offshoot that makes processors with high end, multi-core, multi-cluster, cloud and *AI* objectives, Baidu, and of course Huawei. This is an army of bright people behind *RISC-V*.

Just as *UNIX* changed computing through its openness giving world-wide freedom to clever people, or foundries gave world-wide freedom to chip production, *RISC-V* is giving freedom to combine chip design, chip production and software freedom through the merging of all these trends in one product. *ARM* is formally owned by Softbank, but it is a very complicated ownership with China's *ARM* part also in charge. Softbank wishes to make *ARM* a public company, but there are these ownership issues involving Japan, *UK*, *USA* and China in the background, all at a time when *ARM* is also losing to a new competitor. The flattening of *ARM*'s output is quite telling.

The writing is on the wall for *RISC-V* designs to keep improving and becoming the main computing-structures design in the next few years as it continues to progress and more and more people join implementing their own designs. *ARM* has lost this race.

I hope that I have convinced you of this thesis of dirigisme—exercised with caution—as an operating principle in the new world.

USA-China is but one example. Any competition—an inevitable outcome of a nation trying to raise itself to higher echelons—will face it. It is a north-south, west-east, past colonizers-colonies, first world-third world strife for getting an upper hand whose major catalyzing reason is economic. A recent example is of South Korea and Japan. South Korea, for example, has done well by focusing on semiconductors. Japan-South Korea have historically been frenemies due to the deep-seated angst from World War II and has had its own dirigismic incidents with Japan within this decade.

Dirigisme can be rewarding when one is at the bottom and getting going on an new path. Protection provides a chance to establish oneself. But, in turn, it also is a trap. Protected industries become comfortable in status-quo. This is the story of India of the oligarchs-politics industrial complex for many decades after independence. No desire existed to keep rising that is essential to avoid the income trap, and to take control of future. The long arm of this story exists in India's information information industry. It is a back-end office of information tackling for the West. There exists no equivalent of Baidu, Huawei, SiFive, Biren, Tencent, BYD, and so many other high technology companies across various industries that China has built without having been the world's back-office supplier.

Open software is one clear way of becoming free of external influences and for exercising control that can be leveraged to becoming the higher-end source for the world. One needs to be competent in understanding the pitfalls, the hidden trapdoors, the bugs, and everything else that complexity has, but it removes a nation or a community from being blocked in its aspirations by another.

6.7 *Education for free flow and leadership*

This stress on openness and using open work of others just stresses the importance of education as a flow-driven and leadership-driven requirement of technology even more. Germany, as I had stressed in the past writing, designs for the needs of the entire flow and statics of a needs chain, thus making it very efficient. Problems do not crop as often as bottlenecks, friction and unnecessary heat-like waste in energy in effort avoided, much of the energy goes into productive purposes such as education of the young in a sheltered environment, their freedom to choose, and sufficient numbers of educated and

I have stayed away from the defense aspect of dirigisme. I don't know enough. These are all secrets of the dark halls that mortals cannot observe.

Samsung and Hynix are the major semiconductor entities in South Korea. Both are world's memory source (with China and *USA* following), the former is also a foundry—like *TSMC* with similar high-end capability in nodes, with *IBM* as a major customer, but in addition, there is DB Hitek, that is in high voltage, analog-digital, as well as SOI products and other specialist technology, a bit like On semiconductors. The Covid period taught us that even simpler semiconductors—of past generations—are just as vital. Automotive semiconductor demand could not be met. Much of it depended on products from world over, including China, and in other semiconductors and specialty tasks.

Wars, even if short—few years—events, leave behind a long wave of repercussions. Long and short too takes place as in between occupiers and occupied. Japan placed restrictions on specialty materials, coatings such as polyimides and specialty gases such as hydrogen fluoride that are essential to patterning and growth of silicon structures from being exported to South Korea in 2019. It is not clear at all what the cause is (see www.wto.orgenglishtratop_edispu_e/cases_eds590_e.htm but competition must factor in. The result a few years later (see asia.nikkei.comBusinessTechSemiconductors-Japan-export-curbs-pay-off-for-South-Korean-chip-materials-makers) is that chip materials industry blossomed in South Korea. This episode leads me to the corollary that dirigisme, if one is the underdog, is a very effective tool worth many many marshmallows. Not many give credit to George Fernandes—then the Minister of Industries—who ejected *IBM*, along with Coca Cola, out of India in 1977 for exercising exclusive control in their Indian operations. This was a seminal event in the establishment and growing of India's software and computing industry. As an aside, throwing out Coca Cola was also useful in mitigating the debilitating consequences of excessive sugar consumption and the depletion of water from aquifers that soda industry employs.

trained people who join industry or join academia come about at the end of the chain keeping the growth alive and vibrant. This is essential to continuation and its quality is important to achieving leadership.

Education has its own very unique character. One needs a large body of core learning that makes the later part of creativity and intellectual vibrancy possible. Multiplication as a rote learning is essential to being able to use it as a tool and to move on to higher objectives. Calculus is essential to any understanding of complexity and how to deal with it since second-order coupled equations are pervasive. Exploring behavioral patterns that give us some idea of what is just right, or what will cause an undershoot or overshoot, eventually all goes back to calculus. The effects are integrative, so if one doesn't know differential calculus and integral calculus, it is hard to take a person seriously who is entirely relying on historical patterns that may not really hold true in the present setting. Dynamics and open boundary conditions are all at play. These essential learnings of mathematics and different disciplines make the freedom for being creative possible.

In addition to this scientific analyticity, there is a very essential need to *truly appreciate that what one doesn't know is a far far larger information space than what one knows*. This goes back to Brecht and his statement of science and limiting errors. This what-we-do-not-know part comes from not just posing questions and acting on them, which we all do, but truly asking questions of principle that are much deeper. It is through such an approach that one arrives at new points of view. This is insight and some understanding of order in a dynamic environment. It enriches the experience and together, all these, make it possible to think through more confidently, and be creative.

In the 19th century, two models of modern education arrived in the world. The first was the Humboldt educational model that the university is the environment where students turn to being autonomous individuals and world citizens, so the education must be organized to raising independent multi-dimensional thinkers who care about the natural world possible. To a large extent, the German model has stayed true to this objective with additional modern specializations and world excellence in all intellectual endeavors planted at the top in graduate education and research in the university and at the institutes. The second model was with the arrival of Johns Hopkins University with research and modern problems, medicine as a science being the seed at that time, forming the core of the education endeavor. Both have been successful, yet times are different too, and education too needs to fit into its local environment, the needs of the society, and in science and technology, the industry's future needs that certainly Germany seems to have a good historical record of figuring out.

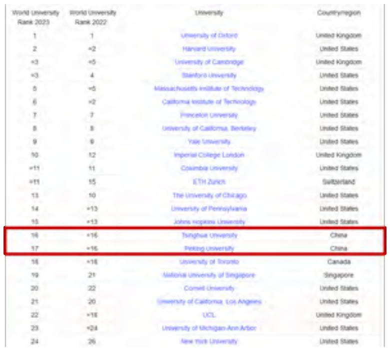
In the modern world. In the Hopkins model, this can lead to problems of overshoot and undershoot as technologies cycle, students, susceptible and for many, driven by the vector of where the media is directing their mind, are often optimizing future earnings expectations and the inner appeal to them of the subject.

The modern need for being leaders in science and technology is for a blending of both the Humboldt foundations and the Hopkins research drive, but with the needs of modern science and technology in both. Critical thinking, problem solving, asking deep questions, self management, working with others, being adept at technology use, and the core literacy of sociology, psychology, and other such explorations as well as what the artists and writers have to say, are important in this modern education.

At the turn of the century, both China and Germany outlined policies and effort to place their higher education and research in the upper echelons. China's objective was to place 15 universities in the world's top 100, Germany's was to build up a select set of centers of excellence. A decade-and-a-half later, China's success (Figure 6.18) with its effort are quite visible. Tsinghua and Peking universities are very clearly in the very best group of the world. I even suspect that the common lists with their different criteria, although they are gamed by all, are not truly reflective of intellectual excellence at the institutions. My list will have *MIT*, Caltech, Stanford, and *ETH* at the very top, and I will not shortchange University of Chicago, which has a remarkable emphasis on developing a keen intellectual mind. These lists also do not reflect Germany's eminence. Between its Max Planck Institutes for the exploratory and Fraunfhoffer Institutes for the practical, and such historic places as Munich, Aachen, Gottingen, and others, and looking at the number of Germany-based Nobel Prize winners over these decades, it is enormously successful without being so measured by the faux criteria. Tsinghua and Peking are constantly in the news with developments in modern areas such as quantum computing and cryptography and networks as well as in biology. There exists even a quantum key distribution based nearly 5000 *km* length secure fiber infrastructure. This is not trivial technology and is a compelling example of success in technology and science from the universities. This is also reflected in the well-ranked scientific publications (Figure 6.19). Again, as with ranking, this is not a very convincing statistic, some of the most powerful work is only recognized as such decades later since it is too esoteric or too few people in that domain, but it is certainly evidence that in all the

For India, from what I can see, it is data sciences for the past decade. The education for it, however, is mostly a higher level of multiplication table learning in the form of coding skills. Not the deeper nature of what the meaning of the data is, the fallacies there are in statistics, and others. Such large-scale training in a subject that suddenly crashes because technology moved elsewhere or the froth did not work out, can lead to large disappointments. *AI/ML developments are primed to displace rote coding*. Semiconductor education has had this boom and bust cycle in *USA*. I keep my fingers crossed for *AI/ML*. Certainly the prognosis from all the layoff news from Meta, Google, Microsoft, *IBM*, et cetera, doesn't look as good at least for a few years at the moment.

The Hindi expression that comes to mind is ``bhediya dhasan,˝ that is, herding of sheep in English. I was accused to be party to this phenomenon by a very bright *IITK* student friend of mine when I moved from being a student of physics to electrical engineering. What has kept getting progressively clear to me is that I loved both, the physics—of really figuring unknowns out—as well as engineering—of doing something real with hard work, mind and hands. Fortunately, my life kept evolving as an entanglement of physics and engineering, so I never really left either, the horizons just kept expanding, and I am grateful for this luck. This is very Humboldt like with a Hopkins top hat.



World University Rank 2023	World University Rank 2022	University	Country/region
1	1	University of Oxford	United Kingdom
2	=2	Harvard University	United States
=3	=5	University of Cambridge	United Kingdom
=3	4	Stanford University	United States
5	=5	Massachusetts Institute of Technology	United States
6	=2	California Institute of Technology	United States
7	7	Princeton University	United States
8	8	University of California, Berkeley	United States
9	9	Yale University	United States
10	12	Imperial College London	United Kingdom
=11	11	Columbia University	United States
=11	15	ETH Zurich	Switzerland
13	10	The University of Chicago	United States
14	=13	University of Pennsylvania	United States
15	=13	Johns Hopkins University	United States
16	=16	Tsinghua University	China
17	=16	Peking University	China
18	=18	University of Toronto	Canada
19	21	National University of Singapore	Singapore
20	22	Cornell University	United States
21	20	University of California, Los Angeles	United States
22	=18	UCL	United Kingdom
23	=24	University of Michigan-Ann Arbor	United States
24	26	New York University	United States

Figure 6.18: Times higher education's list of the premiere universities of the world. Source: www.timeshighereducation.comstudentbest-universitiesbest-universities-world.

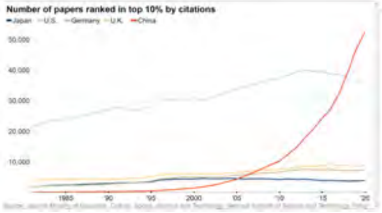


Figure 6.19: Top 10% papers count.

currently widely pursued science areas, China has placed itself at the top of list. So, both in its education and research quality, it has bootstrapped itself up to be on par with Western world.

India's, unfortunately, is a sad tale. Even as there are individuals at scattered institutions, who can be seen as leading thinkers and doers, most institutions and most teaching just does not bring out the passion and joy of learning and discovery and all the breadth of advanced science and engineering that is constantly expanding wider. The education is too stuck in the past, employs too much rote methods, and far too much of testing and grading that usually goes together with the rote methods. The young mind is still forming its neuronal network and its thinking and probing and doing style till the age of 30, of which the college year age is the most vital. This is when one has left the security of home, one is an explorer and traveler in the vast unknown, and it is the curiosity about this unknown that needs to be promoted, buttressed, and supported with learning and learning style. Instead the students get beaten down into submission, and only a few escape this tyranny. The consequence is that independent of what engineering or sciences one have pursued, the student becomes an income-oriented cog in the data machine. No Indian company can be contrasted with any of the Chinese successes in leadership of new directions and new products that lead the world.

Rote teaches reproducing, copying, and refining.

The educational methods of India need a Humboldtian revolution.

To this I add the autre of dirigisme. Dirigisme can lead to failures, as the 5th generation project and export controls have been for Japan. India too responded to this excellence push effort by establishing an initiative of institutions of eminence. It is a heart-breaking story. I looked up information on the current state. One of the institutions of eminence is Jio Institute, which at the time of the start of the initiative had no students, but 52 acres of land. Today, it has 2 graduate programs, 120 students, and 6 faculty. Pathetic. Buildings alone do not make an institute. Creative students and faculty, hard work, world recognition of the work, local impact, and serving humanity does. Even Caltech, one of the smallest superb institutes of the world, with about 2500 students (2/3rd graduate and 1/3rd undergraduate), has 300 faculty, occupies about 400 acres in the middle of the city of Pasadena. And each output from Caltech over the past 100 years is a gem worth reading.

Excellence is excellences, no compromises, strong expectations, commitment to research and good teaching, and constant striving for getting better and better at the top of the world's research. There is a complete disconnection between education and aspirations, and the disconnect in the present form, guarantees a lower income trap where one is going to remain a follower.

This serves to debate what an aspirational country that wishes to reach back to its place in the upper echelons of the world that it occupied in distant past must do. The lessons of this semiconductor-oriented discussion are just as appropriate to many of the other areas of science and technology—biology, transportation, agirculture, et cetera—that the country will need to excel at.

Being a leader in semiconductors and using semiconductors to improve one's lot and not be caught in a trap requires a lot more than building a fab or two at the highest advanced technology point current at that time. Facilities become obsolete in half a decade, at which point they certainly can be repurposed for lower points of the value chain, such as for transportation needs, or others. But, to make wealth and grow requires the ability to exploit that most advanced technology. Today, this is now being driven by *AI/ML* such as in Nvidia's or Biren's technology, or the high-end networking hardware such as of Huawei. In five years, there may be an introduction of quite different pieces of technology, which also needs to be created and developed. There is also a value chain in exploiting the semiconductors, providing for the semiconductor technology, and to exploit the semicconductor technology in any significant way, being the standard setter who also controls the operating systems of record. All this requires a wide and broad ecosystem spanning skills, education and infrastructure.

This ecosystem is one for both the physical semiconductor value chain as well as for the industry that makes it possible and education that makes the continuing progress come about.

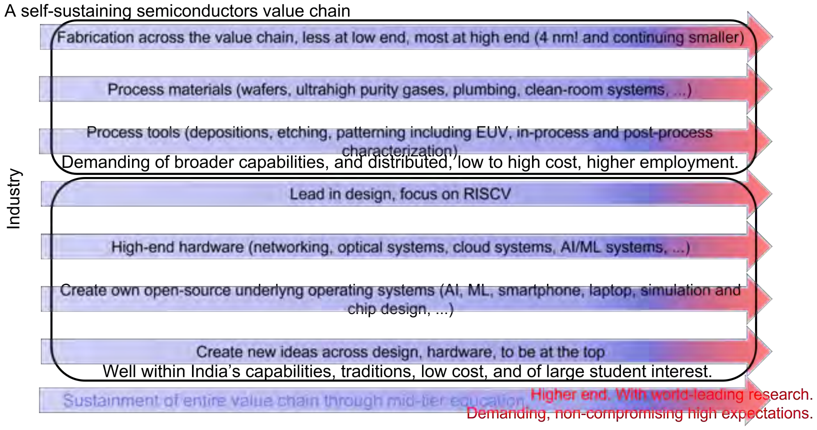

A self-sustaining semiconductors value chain
Fabrication across the value chain, less at low end, most at high end (4 nm! and continuing smaller)
Process materials (wafers, ultrahigh purity gases, plumbing, clean-room systems, ...)
Process tools (depositions, etching, patterning including EUV, in-process and post-process characterization)
Demanding of broader capabilities, and distributed, low to high cost, higher employment.
Industry
Lead in design, focus on RISCV
High-end hardware (networking, optical systems, cloud systems, AI/ML systems, ...)
Create own open-source underlyng operating systems (AI, ML, smartphone, laptop, simulation and chip design, ...)
Create new ideas across design, hardware, to be at the top
Well within India's capabilities, traditions, low cost, and of large student interest.
Sustainment of entire value chain through mid-tier education
Higher end. With world-leading research.
Demanding, non-compromising high expectations.

Figure 6.20: Semiconductors and semiconductor-related higher value efforts with an education chain that supports it.

Semiconductor plants at the high end cost $\$10B$ investment. Yet they employ only a few thousand employees. Those that keep the automated machinery humming, those who are doing the plumbing, characterization, checking, technology day-to-day fixing, supplies, in-and-out scheduling and so on. This nets as a cost of $\$5M$ per employee. In this form it makes no sense. It also makes no sense if by dirigisme, when a country undercuts another country's strength, critical supplies or necessities that make a plant hum are cut off. This highest cost effort becomes very productive and useful if one expands it to build the entire value chain (Figure 6.20) that makes semiconductor technology possible. These are much lower cost, employ many people, and the expertise then makes the country a supplier to the world in a far broader domain. In this, I would include process materials, such as wafers, ultrahigh purity chemicals gases and liquids, the vast plumbing, automation, electric, cleanliness and safety infrastructure that goes together into an automated modern clean room. Just as important is the creation of a process and characterization tool industry: tools for patterning such as extreme ultraviolet, and the automated multi-chamber plasma-based gas or even liquid employing tools, optical and electrical tools for in-process and post-process characterization, and others. Europe—Sweden, Austria, Germany, Switzerland, Holland—have shown remarkable success in these tasks through their engineering talent and that talent's recognition as a very important profession. These are specialized technology-demanding tasks that are necessary for a fabrication facility to work. Between all these tasks, one has a demand for broader set of capabilities and expertise from nearly all the engineering disciplines, these ancillary industries can be distributed around the country, some are low cost, some higher, but they all provide large well-earning employment.

India should remember incidents such as nuclear sanctions from Western countries and Japan, or space research sanctions by United States.

Memories, dynamic and non volatile, go through boom and bust cycles as demands change. But, more transistors appear in memories than in the logic, a ratio bordering on 80%. Our cell phones need to store for fast access. All the transactions need to store the data in a way that it can be quickly processed. This is all in memories in close proximity of the processors. Even in processing such as the most advanced tensor processors, working with data—this data must reside as it is processed—is in memories built into the circuit. In this case right next to processing parts for immediate access. Memories may be very economically cyclical, but without memories, computing and communications and automation and all electronics cannot exist.

Micron technology just said that its revenue decreased by 50% and had a loss of more than $\$2B$ in the quarter. The last quarter was similar too. China, in a dirigismic act, has placed Micron memory products under review for security purposes, similar to what *USA* has done to a number of Chinese companies' products. When computing demand drops, usually when economies are struggling, memories reflect that as compounding. An organization needs a spine to get through booms and busts.

Even succeeding just at this part of semiconductor industry is not going to be rewarding to a country.

The value creation and large-scale talent need is higher up in the value chain, for which semiconductors are the foundation. China has been very successful at this building out through the vast array of companies that compete on equal terms with the *USA* counterparts. This is deployment of intellectual heft on top of a physical structure that is by its nature demanding and capital intensive.

To take advantage of the semiconductor technology capability, one has to look at where it is used. This is the walled gardens that have been built by companies either through the designs themselves or the software that underlies the usage. I place the following as the important components of this strategy.

- Lead in chip design. This is the immediate demand and usage creator of the semiconductors. Processors, memories, sensors, graphical engines, all the edge-of-the-network creations, all the microcontrollers that run appliances, any system with any level of automation needs design. Except for the highest-end hardware, these are commonly usable designs that can be integrated across various platforms. The most important necessity for such designs is common instruction sets, and other common standards. We have seen how the world has moved to nearly 1/3rd of the designs in *RISC-V*. There is a tremendous talent pool of engineers, such as those involved in data tasks, who can also learn these tasks. India needs to lead in design and this is do'able using *RISC-V*—an open platform—by being the place where the talent exists for what is the most essential integrative task in the use of semiconductors.

- Design needs design tools. Tools that work with accuracy in thermal, electrical, symbolic, behavioral domains for digital, analog, and optical signaling. Such tools are complex, a good challenging software and mathematical task to which *AI/ML* is going to bring clever new methodologies.

One of the early sanctions on China was the denial of high-end design tools. No tool, no future, except if one figures out new ways or develops one's own.

- Make the designs, the chips, and the hardware that is at the high end. This list is broad: network, optical networking systems, cloud systems of computing, the *AI/ML* systems and the supercomputing systems that use the highest end processors. This is where a large bulk of the profit, despite fewer number of products, is.

- None of the above succeeds without control of the operating systems. Google, Apple, Microsoft succeed by trapping individuals in their domain. A large country with a larger user community is quite capable of creating its own open-source underlying operating systems applicable to smartphones, computers, simulations and chip design, and the *AI/ML* and the supercomputing tasks. As well as new ones that will come that it must create by excelling in education and research. Through its success in the *UTI* and Aadhar systems, India has shown that such control also leads to lower cost for the populace and security at the same time.

This above collection is well within India's capabilities and traditions. It is talent demanding, not capital demanding. As a set of tasks that are intellectual and demand thinking, it is one that is likely to be of large student interest.

The sustainment of this large value chain requires both mid-tier talent and high end talent. So, there is much in these tasks that requires an education system that does need to change and be competitive in the modern advanced world. None of the Indian institutions are there, select faculty at select universities and institutions are.

A world-class university in research and education—Caltech or *MIT* like—has to be non-compromising and demanding from both its faculty and its students. These are institutions that are not copy-and-paste and executive bullet point elevator pitch institutions that are inherently biased and fallacious result of Occam's razor. Institutions need to have original thought, creative thought and practice, and need to fit in the local environment and its needs in space and time. There have to be world-class standards that must be strictly adhered to. The world-class standards are not publication metrics of numbers of publication, but research output that others sit up and notice and follow up on. Education is the instilling of creative exploration as a natural act in the student. The purpose of education is to fire up a young person's mind of the immense possibilities ahead with the learning they have achieved. It is not about stuffing them with ‵‵facts.″ Neither Caltech nor *MIT* started as major universities. Both were vocational schools about hundred years ago. Caltech started as Throop and *MIT* as Boston Tech. Good things take time and care. Fast and slow need to be balanced. It is not possible to hire top-class faculty in one fell swoop. There are not that many people beyond 4σ. Culture building and maintaining takes time and constant effort and discipline and awareness of the two marshmallows principle. Always hiring somebody brighter than oneself is one principle with which one cannot go wrong.

We as faculty like to work with problems that stay within our control. These are inevitably highly constrained and we all have our own ways of finding an optimal spot which suit our methods and objectives. This worked before, but I do not believe it will work if everybody continued doing this. The science and technology world has changed. A large fraction of the most interesting problems are now complex. They are integrative. We now have tools to deal with complexity. *AI/ML*, 500 years of learning, hardware, new ways, can all be brought to bear on the difficult problems. Hardware design, chip design, materials design, process design, integration, systems, and use of systems by people are all complex integrative task to which many directions of technology feed. These are perfect problems for

Using Josephson junctions for computing was one of the high-risk research efforts at *IBM* Research in the 70s and 80s. There were many lessons from this effort, similar to that from an earlier effort of deploying tunnel diodes in the 1960s, technical in that directionality, gain, fan in and fan out, signal-to-noise ratio, reproducibility and robustness, et cetera, all really matter in integration. But more important, from Josephson effort and some from the silicon-oriented effort, was a short-and-long message for success. One cannot expand people numbers rapidly no matter how time sensitive a task. One ends up with a normal distribution for that grouping. Quite a few of the maximum likelihoods ascended the organizational structure, in turn, hurting the entire research-and-beyond enterprise. When the going gets tough, it is the folks with inner strength—strength accumulated from many domains but with the important characteristic of belief and passion backed up by learning and experience—that make it through the walls. Lucent, the vestige of old *AT&T* and remnants of the Bell Laboratories, couldn't because of a hardcore salesman with golf course fetish was in charge. *IBM* had Akers. He gave Bill Gates the keys to the personal computer operating system, Gates immediately acquired Gary Kildall's *DOS*, repackaged it, and crossed the moat into the castle in a Trojan horse. *IBM*'s recent story with *AI* has been of Watson, named after the luster accrued from the victory of ‵‵*AI*″ over Kasparov in a chess game in the 1990s. It pushed Watson *AI* hard in medicine with salesman-like false promises and failed. Medicine and health care are not the same as chess. Judgments, specifics, experience, past, details, and so much more of what one does not know matter in making life-and-death decisions of cancer care. It is not chess with hard rules and all the necessary information right there on the board. Health is real world. Chess is a parlor game. From what I see, *IBM* has learned this lesson with its offerings of different *AI* models for different limited domains. This kind of understanding and foresightedness is critical in the management. Foundational learning and people skills both matter. This drama plays out in academia too. People with little research contributions, no aspirations towards learning and pushing teaching frontiers, and beholden to Google searches—as gate-keeping messengers of a system—eventually rot great organizations. Seeing long is hard. My own experience of of finding the most academic of the environments in an industrial laboratory has amused me ad infinitum. The European universities come closest. The private *USA* universities are mostly non-profits for profit.

deploying learning in *AI/ML* that has only arisen in the past half a decade. We must embrace complexity as a way to become leaders. Systemically, of course, this also means mitigating friction, creating painless processes, have an appropriate organization that rewards and shelters, and encouragement. Success, change and well being of the nation will be the reward.

For a flow of the growing society's march to high value requires this stress at the very best institutions, but also a stress—as in Germany—of a high-expectations education in the march from childhood to adulthood. So, pre-school, primary schools, middle schools, high schools all matter just as much as college education. This education has little chance to succeed purely by respecting a profession, as ancient cultures often do. It requires infrastructure and well-paid teachers that attracts the group of college graduates with a love for education, who are qualified, and who can pass on and demonstrate that love to the growing mind.

The country needs entrepreneurs, breadth of ideas being tried out, and also top-class faculty. The change that I see from my time is that while in my time, the best were interested in being in intellectual life and aspired to be faculty members, today it is high-paying jobs. We need to keep promoting and keep creating and tweaking mechanisms for starting up of new ideas from the bright young people who also are fortunately interested in staying in the country which is increasingly a country of opportunities. It needs to be done with processes that prevent exploitation and financial chicanery that is all too common. It also means India needs to find ways to get past its oligarchic nepotistic ways that are unfortunately still too common and so easy to see not just in commerce, but also in media, and the culture, which unfortunately has been reduced to primarily a Bollywood culture.

Independence of thought, valuing opposing thoughts to improve, not defensively reacting to every opposition, making mid-course corrections upon acquiring new insights, are all necessary virtues of good leadership. Future is never quite predictable. One can only make good judgments and good judgment come through wisdom. Wisdom is decision making based on information and insights and belief in oneself acquired through past successes.

The mechanisms of the higher end higher value science and technology seeding, germination, and growing need to be centered on ideas, ability, and intellectual depth.

A technology-society reason to particularly stress this point is another example of short-and-long: that the time-scale at which technology changes is far far shorter than that of the governance in society. There are constant disconnects. It is, for example, visible right now in the social impact of social networks on loss of social

There is a beautiful anthropological book by David Graeber titled *Bullshit jobs. A theory,* (ISBN 978-1-5011-4331-1, Simon & Shuster (2018)) that is an expansion of the arguments of a similarly-titled essay that is available on the internet and was written during the period of the Occupy movement. Occupy movement happened during a very difficult financial period for people, while Wall Street was, as it does all the time, making hay. He argues that it is easy to figure out what an essential job is. Just look at the salary for it. In *USA*, this applies to the nurses, the check-out people, the food servers, et cetera. Hard work and subsistence living. School teachers are such essential workers. Bullshit jobs are the ones where the person goes home every night, he may be rich, but he wonders as he lies on his bed trying to go sleep what his contribution to human life is. Bullshit is a proper dignified civil word. The essence of bullshit is not that it is false, but that it is phony. A liar needs to keep the truth in view in order to concoct the lie. The bullshitter is utterly indifferent to truth. I have met such people in industry, academe and government. Since I was at a great research company in my early working life, my bullshit rank order is government, academe followed by industry. Academe is quite a confusing place because of the conflict between not-for-profit, teaching and research, and keeping the business of academe going. Too many people, specially in administration, who move things from here to there, and from there to here. Bullshit, appears in this respectful place in another great small book, *On bullshit,* by Harry Frankfurt published by the Princeton University Press.

learning by children, polarization in the adults, the new forms of addictions brought by internet and cryptocurrencies, none of which the governance has been effectively able to respond to. Education, intellectual heft, and model governance are essential to avoid major collapses arising in the rise of the modern rise of technology and what it makes possible.

6.8 *Judging and awarding world-leading efforts*

I note fallacies appearing in funding award mechanisms, on one side where Occam's razor fallacy prevails with power invested in an individual, or on the other side, a collective that usually ends up being people who have time or who are themselves in search of such funds. In *USA*, for example, much of *NSF* panels are composed of people starting on their research journey and evaluating proposals that have much in common, with only a few breaking new paths. The breaking new paths naturally doesn't find as much appeal since a normal distribution function is evaluating a normal distribution function. A median of a collection of such reviewers should hardly be expected to reward ambitious projects. It is important to sanity check ambitiousness. Is it on firm grounds, what has been missed, what is a counterfactual, and if it passes this test, is it by a person who can be expected to follow through despite the numerous fundamental and practical difficulties that arise in any creative undertaking. If it passes these tests, then it should be funded. *NSF*, by and large, fails this test. It is only with a program manager who is astute enough, has experience, knows and appreciates the ebb-and-flow of research who can overcome the systemic inertia.

Awarding funds for an intellectually demanding and expected to be world leading task can never be a democratic spread-the-wealth mechanism. An averaging of judgment by a normal distribution inevitable leads to an averaging of final results. This is the story of human history.

There are deviations from this model that succeed for some time before they too get overcome with gathering moss. A successful one has been *DARPA* led by a rotating group of people who have sampled the world of academia and industry and who are interested in national service as also intellectual pursuits of changing world in their specialties in their lifetime. Fair sums of money, a rigorous time-bound checking, a focus on the most advanced areas of technology, and a view to its move to society. It may be driven by defense, but it ends up in the society. The progress in the integrative high-end technology of self-driving cars, where *LIDAR*, multidomain signaling or fast machine learning and *AI* is the essential leading to the present robotics and *AI/ML* of the world, or in plant-based generation of vaccines leading to low cost vaccines, or the autonomous drones which required low energy flying technology and is now useful in improving agriculture go back to *DARPA*-seeded ideas. They are a success because of the ambitiousness with soundness of ideas, the selection, and the no-compromise expectation and reward. There is no room in this for excuses or flattening a distribution. The money can be better spent as projects show failures. Failures are learning tools but shouldn't be unproductive sinks of precious funds and talent.

6.9 Last word

I have written about a matter of inference that is replete with complexity and is subject to large errors. We are all adults. Many class-structured societies have a tendency to treat others down with faux praise, or worse. The scientific way is to draw on all the information and past experiences across all the domains that make life and living and the concept of country a social matter. In a society with wide disparities, and of historic angst, the idea of growth balancing short and long objectives can be hard. But, history and times and a position and flow of the current state in the global interplay is a zugswang opening for India.

A prior zugswang for India was when the *USA* sent off the English, and the English chose India as the next target. I am inextricably entangled in both, and in a way, that should not be surprising as a Markovian connection.

We are, I believe, at the threshold of another order from disorder, not unlike what one sees sometimes in sciences if phase transition, one where there will be an order defined by military and security and related alliances, another by economic related by some schema of rank ordering of high value to value creations by the education-industrial-social framework of a country, and the the third of a digital order, where technology companies—not unlike British and Dutch East India company—will set the flow of information that is at the heart of the the functioning of the physical world. History teaches us lessons, philosophy teaches us a way to think through such complexities of existence, and information teaches us to minimize our biases. Each country will have to find its path.

I started with Bertolt Brecht's remark to view science as a way to limiting errors rather than providing perfect answers. I believe this. So, I end with another of Brecht's poem, *An den Schwankenden*, that has often given me solace in my darker moments. The following is the last stanza of the poem. This too I believe. In an English translation,

To the waverer
. . .
Whom do we still count on?

Are we just left over, thrown out of the living stream?
Shall we remain behind? Understanding no one and understood by no one?
Have we got to be lucky?
This you ask.
Expect no answer other than your own.

This is the lesson I learned from my student days at *IIT* Kanpur. Erwarte keine andere Antwort als die deine. Danke shön.