Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur
Proposal for a New Course

Course No: CS698B
Course Title: Fundamentals of Data Engineering – Part I (Data)
Per Week: [3-0-3-0] Lectures: 3 (L), Tutorial: 0 (T), Laboratory: 3 (P), Additional: 0 (A)
Credits: 12
Duration: Full Semester
Course type: PG-DE
Proposing Department: Computer Science and Engineering
Proposing Instructor: Purushottam Kar
Pre-requisites: None, but familiarity with basics of linear algebra, calculus and prob-stats assumed.
Course Description:

a. *Objectives*: Data handling is essential for ML applications with Python being the contemporary choice of programming language for designing and building ML applications and services. This course is the first part in the DOT trilogy of courses (Data-Ops-Things) and will introduce students to Python basics, followed by techniques for data collection, curation and comprehension using Python libraries. The course will also offer hands-on lab experience involving digital data sources.
b. *Content Overview*: There will be an equivalent of 40 lectures of 50 minutes each and 12 labs of 3 hours each. A weekly breakup of lecture and lab content is given below.
c. *Evaluation*: Evaluation will use a combination of lab coding exercises, lab exams, take-home coding assignments, coding projects, and pen-paper quizzes and exams.

Weekly breakup of content:

Lecture content (40 lectures)			
SNo	**Broad Title**	**#**	**Topics**
1	The Bare Necessities	8	*Elements*: indentation, keywords, good identifiers, name-object system (Python) vs variable-value system (C, C++, Java) *Operators*: unary, binary, assignment, conditional, operator precedence *Expressions, statements*: simple, compound, execution wrappers (with) *Loops*: while, for, notion of invariants, nested loops, break, continue *Functions*: args (positional, kw), return, namespace, scope, recursion
2	Built-in Goodness	8	*Built-in datatypes*: logical, integer, float, character, logical/Boolean, string, tuples, lists, dictionaries, sets, mutable vs immutable types *Built-in ops*: concatenation, repeat, length, unpacking, zipping, list comprehension, indexing, slicing, broadcasting, typecasting, I/O
3	Notebooks and Libraries	2	*Setup*: iPython, notebooks, notion of execution cells, imports, aliasing, IDEs and dev environments (VSCode, Jupyter, Colab, Binder) *Basic operations*: using numpy, scipy for numerical computations
4	Pro Tips	2	*Code hygiene*: comments and documentation, writing requirements *Resources*: effective use of API docs, community fora (Stack Overflow, GitHub etc), etiquettes when engaging at online fora *The future of dev?*: Assistants (ChatGPT, Copilot, Gemini, Claude etc)
5	Data Storage and Processing	4	*Formats*: CSV, Excel, Parquet, SQLite, JSON, sockets, etc *Processing*: ETL (extract-transform-load), data scraping, preprocessing
6	Data Visualization and Analysis	10	*Visualization*: plots (line, scatter, bar, stacked, box-whisker), error bars, confidence intervals, aesthetics, accessibility (colorblindness, alt-text) *Analysis*: Primitives (sampling, classification, regression, clustering, dimensionality reduction, data augmentation, anomaly detection) *Implementation*: Use of Python libraries (numpy, sklearn, scipy, pandas, openpyxl, sqlite3, matplotlib, seaborn, plotly, bokeh)
7	Look Ma, I am a Dev!	6	Building native/web apps for data viz., games, productivity, services etc using Python/JS libraries such as PyGame, FastAPI, D3, pipelines, data transformers, trigger/event-based programming,.
	Total	**40**	

Lab content (12 labs): (examples given below are illustrative)

1. Familiarization: IDEs and programming environments
2. Python basics: creative use of built-in datatypes and operations, loop invariants, conditionals, built-in and user-defined functions
3. Data storage, processing, visualization, and analysis
    a. Perform data cleanup, missing data imputation, visualization on static data from a spreadsheet or SQL database, or live data from an online source
    b. Implement supervised and unsupervised data analysis on cleaned data
4. App development
    a. Build a web app offering ML-as-a-service
    b. Build a live app by feeding live data into a pre-trained model e.g. face detection system by piping camera feed (depending upon hardware availability)
5. Use of coding assistants (depending upon license availability) and debugging tools

Lab Equipment: Sufficiently many PCs with internet connectivity, and appropriate software (browsers, IDEs such as VSCode, and Python + IPYNB runtimes with libraries) will be needed. Specific experiments may need specialized hardware.

Short summary for inclusion in the Courses of Study booklet: the course will introduce Python basics and Python-based techniques to collect, curate, and comprehend data.

Textbook: There is no textbook for this course.

Course proposer:
Date:

Convener DPGC:
Date:

The course is approved/not approved
Chairperson, SPGC
Date: