IIT Mandi

Course Title: Introduction to Data Science for Experimental Scientists
Course Code: CS502P
Credits: 1-0-3-3
Prerequisites: Consent of Teacher
Intended for: MSc/PhD in Chemistry, Biology, Physics. Not open to B.Tech.
Elective or Core: Free elective
Semester: Any

1. Course objectives

Experimental science relies on measurements by precision instruments. Many of these instruments are computer-controlled and make available a large amount of data – input parameters and measured results – in digital format. Analysis and interpretation of this data is a key part of experimentation. To this end, the techniques for data collection, analysis and visualisation are an important part of the experimentalist's toolbox.

This course is intended to introduce Python programming, and the use of built-in statistics and machine-learning libraries for typical problems in experimental chemistry, biology and physics. The practicum format ensures that the student will be able to use these tools after this course. Lab assignments will include analysis of very large datasets from real lab experiments.

The course assumes no prior knowledge of programming. It is not intended for engineers, computational scientists and those who already are familiar with programming.

2. Course Content

Module	Week	Lectures	Lab
1	1	Introduction to Data Science; model of computation; program = data structures + algorithms	The Python programming environment (OS and IDE); "Hello world" in Python
2	2	Expressions: scalar variables, operators, precedence, data types	Python as a calculator; input/output of numbers; type conversion
3	3	Decisions: if-else; nested decisions; flow-charts	Classification using if-else; debugging using print, breakpoints, execution of selected statements
4	4-5	Top-down and bottom-up program design; iterations – definite and indefinite; arrays	Trying it out using the console; flow-charts to code; initiallising an array; filling an array with input numbers; printing an array
5	6-7	Functions and modules; stats, rand, matplotlib modules Review of statistics Visualisation of data	Filling an array with random numbers; distributions; Computing statistics of an array of numbers; Plotting line, scatter, bar, histograms using matplotlib

Module	Week	Lectures	Lab
6	8-9	Collections: list, set, dict, Numpy array File I/O; exceptions	Use of list, set and dict; Reading data from a file into a Numpy array
7	10-11	CSV file format; Pandas dataframe; Regression and interpolation; measures of goodness of fit; visual inspection – boxplots	Reading `.csv` into a Pandas dataframe; Fitting curves to given datasets
8	12-13	Data collection; cleansing; formats; units	Gridding of datasets; conversion of file from one format to another
9	14	Case studies from experimental chemistry, biology and allied areas	Lab exam

Text books

1. Michael Dawson, *Python Programming for the Absolute Beginner*, 3rd Ed, Premier Press, 2003 (Chapter 1-7)
2. Jake Vanderplas, *Python Data Science Handbook*, O'Reilly, 2016 (Chapters 1-4)

References

1. Montgomery & Runger, *Applied Statistics and Probability for Engineers, 3rd ed.*, Wiley, 2003
2. Jose Unpingco, *Python for probability, statistics and machine learning*, Springer.
3. Ben Stephenson*, The Python Workbook: A Brief Introduction with Exercises and Solutions*, Springer, 2014

3. Similarity Content Declaration with Existing Courses

The course is a subset of IC152. IC152 is intended for BTech students and assumes a strong background in mathematics and familiarity with computers. CS502P is not open to BTechs, it assumes no knowledge of computers and only 10th standard mathematics.