Applied data analysis

Brbic Maria

Summary
This course teaches the basic techniques, methodologies, and practical skills required to draw meaningful insights from a variety of data, with the help of the most acclaimed software tools in the data science world (pandas, scikit-learn, Spark, etc.)

Content
Thanks to modern software tools that allow to easily process and analyze data at scale, we are now able to extract invaluable insights from the vast amount of data generated daily. As a result, both the business and scientific world are undergoing a revolution which is fueled by one of the most sought after job profiles: the data scientist.

This course covers the fundamental steps of the data science pipeline:
Data wrangling
- Data acquisition (scraping, crawling, parsing, etc.)
- Data manipulation, array programming, dataframes

2024-2025 COURSE BOOKLET

Language of teaching
English

Credits
8

Session
Winter

Semester
Fall

Exam
Written

Workload
240h

Weeks
14

Hours
4 weekly

Lecture
2 weekly

Project
2 weekly

Number of positions

Applied data analysis Page 1 / 4
• The many sources of data problems (and how to fix them): missing data, incorrect data, inconsistent representations
• Data quality testing with crowdsourcing

Data interpretation
• Statistics in practice (distribution fitting, statistical significance, etc.)
• Working with "found data" (design of observational studies, regression analysis)
• Machine learning in practice (supervised and unsupervised, feature engineering, evaluation, etc.)
• Text mining: preprocessing steps, vector space model, topic models
• Social network analysis (properties of real networks, working graph data, etc.)

Data visualization
• Introduction to different plot types (1, 2, and 3 variables), layout best practices, network and geographical data
• Visualization to diagnose data problems, scaling visualization to large datasets, visualizing uncertain data

Reporting
• Results reporting, infographics
• How to publish reproducible results

The students will learn the techniques during the ex-cathedra lectures and will be introduced, in the lab sessions, to the software tools required to complete the homework assignments. In parallel, the students will embark on a semester-long project, split in agile teams of 3-4 students. In the project, students propose and execute meaningful analyses of a real-world dataset, which will require creativity and the application of the tools encountered in the course. The outcome of this team effort will be a project portfolio that will be made public (and available as open source). At the end of the semester, students will take a 3-hour final exam in a classroom with their own computer, where they will be asked to complete a data analysis pipeline (both with code and extensive comments) on a dataset they have never worked with before.

Keywords
data science, data analysis, data mining, machine learning

Learning Prerequisites

Required courses
The student must have passed an introduction to databases course, OR a course in probability & statistics, OR two separate courses that include programming projects. Programming skills are required (in class we will use mostly Python).

Recommended courses
• CS-423 Distributed Information Systems
• CS-433 Machine Learning

Important concepts to start the course
programming, algorithms, probability and statistics, databases

Learning Outcomes
By the end of the course, the student must be able to:
Construct a coherent understanding of the techniques and software tools required to perform the fundamental steps of the Data Science pipeline

Perform data acquisition (data formats, dataset fusion, Web scrapers, REST APIs, open data, big data platforms, etc.)

Perform data wrangling (fixing missing and incorrect data, data reconciliation, data quality assessments, etc.)

Perform data interpretation (statistics, knowledge extraction, critical thinking, team discussions, ad-hoc visualizations, etc.)

Perform result dissemination (reporting, visualizations, publishing reproducible results, ethical concerns, etc.)

Construct a coherent understanding of the techniques and software tools required to perform the fundamental steps of the data science pipeline

Perform data interpretation (statistics, correlation vs. causality, knowledge extraction, critical thinking, team discussions, ad-hoc visualizations, etc.)

Construct a coherent understanding of the techniques and software tools required to perform the fundamental steps of the data science pipeline

Transversal skills

Give feedback (critique) in an appropriate fashion.

Write a scientific or technical report.

Evaluate one's own performance in the team, receive and respond appropriately to feedback.

Teaching methods

Physical in-class recitations and lab sessions

Homework assignments

Course project

Expected student activities

Students are expected to:

Attend the lectures and lab sessions

Complete 2-3 homework assignments

Conduct the class project

Engage during the class, and present their results in front of the other colleagues

Assessment methods

- Homework
- Project
- Final exam

Supervision

Office hours: Yes
Assistants: Yes
Forum: Yes

Resources

Virtual desktop infrastructure (VDI)

No
Websites
• http://ada.epfl.ch

Moodle Link
• https://go.epfl.ch/CS-401