Summary
This course teaches the basic techniques and practical skills required to make sense out of 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 a new breed of software tools that allows 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 Acquisition
- Variety as one of the main challenges in big data: structured, semi-structured, unstructured
- Data sources: open, public (scraping, parsing, and down-sampling)
- Dataset fusion, filtering, slicing & dicing
- Data granularities and aggregations

Data Wrangling
- Data manipulation, array programming, dataframes
- The many sources of data problems (and how to fix them): missing data, incorrect data, inconsistent representations
- Schema alignment, data reconciliation
- Data quality testing with crowdsourcing

Data Interpretation
- Stats in practice (distribution fitting, statistical significance, etc.)
• Working with “found data” (design of observational studies)
• Machine learning in practice (supervised and unsupervised, feature engineering, more data vs. advanced algorithms, curse of dimensionality, etc.)
• Text mining: vector space model, topic models, word embedding
• Social network analysis (influencers, community detection, 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
• Anonymization, ethical concerns

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 and the in-class quizzes. In parallel, the students will embark on a semester-long project, split in agile teams of three. 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 also 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.

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

Important concepts to start the course
Algorithms, (object-oriented) programming, basic probability and statistics

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.)

Transversal skills
• Give feedback (critique) in an appropriate fashion.
• Demonstrate the capacity for critical thinking
• 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
• In-class quizzes
• Course project

Expected student activities
Students are expected to:
• Attend the lectures and lab sessions
• Complete 2-3 homework assignments
• Complete 2-4 in-class quizzes (held during lab sessions)
• Read/watch the pertinent material before a lecture
• Engage during the class, and present their results in front of the other colleagues

Assessment methods
• 33% continuous assessment during the semester (homework and in-class quizzes)
• 33% final exam, data analysis task on a computer (3 hours)
• 33% final project, done in groups of 3

Supervision
Office hours Yes
Assistants Yes
Forum Yes
Others http://ada.epfl.ch

Resources
Virtual desktop infrastructure (VDI)
No

Websites