EPFL

EE-490(h)	**Lab in data science**

Bouillet Eric Pierre, Choirat Christine, Obozinski Guillaume, Roskar Rok, Sarni Sofiane, Verscheure Olivier

Cursus	Sem.	Type
Data Science	MA2, MA4	Opt.
Electrical and Electronical Engineering	MA2, MA4	Opt.

Language of teaching	English
Credits	4
Withdrawal	Unauthorized
Session	Summer
Semester	Spring
Exam	During the semester
Workload	120h
Weeks	14
Hours	**4 weekly**
TP	4 weekly
Number of positions	

It is not allowed to withdraw from this subject after the registration deadline.

Summary

This hands-on course teaches the tools & methods used by data scientists, from researching solutions to scaling up prototypes to Spark clusters. It exposes the students to the entire data science pipeline, from data acquisition to extracting valuable insights applied to real-world problems.

Content

1. Crash-course in Python for data scientists

- Python packages: NumPy, Pandas, Matplotlib, Scikit-Learn
- Interactive data science with web-based notebooks
- **Project #1**: Curating data from a network of CO_2 sensors

2. Distributed computing with an Apache Hadoop distribution

- Understand main constituents: HDFS, Parquet, HBase, Hive, Zookeeper, Ambari, Spark, Spark Streaming, Yarn, Mesos, etc.
- **Project #2.1**: Prepare a sandbox distribution

- HDFS internals, best practices
- **Project #2.2**: Configure HDFS, prepare files used in subsequent projects, choose appropriate compression, etc.

3. Distributed processing with Apache Spark

- RDDs and best practices for order of operations, data partitioning, caching
- Data science packages in Spark: GraphX, MLlib, etc.

• **Project #3**: Large-scale processing of genomic data

4. Real-time data acquisition using Apache NiFi

• Stream processing using Apache Spark Streaming
• **Project #4**: Indexing tweets with NiFi and Solr

5. Final project - Summing it all up

• Tapping into live traffic data sources from a major city: Acquisition & curation of live traffic sensors, estimation of speed of traffic on different road segments, and prediction of congestion using Spark, HBase, Kafka.

Keywords

Data Science, IoT, Machine Learning, Predictive Modeling, Big Data, Stream Processing, Apache Spark, Hadoop, Large-Scale Data Analysis

Learning Prerequisites

Required courses

Students must have prior experience with at least one general-purpose programming language.

Important concepts to start the course

It is recommended that students familiarize themselves with concepts in statistics and standard methods in machine learning.

Learning Outcomes

By the end of the course, the student must be able to:
• Use standard Big Data tools and Data Science librairies
• Carry out real-world projects with a variety of real datasets, both at rest and in motion
• Design large scale data science and engineering problems
• Present tangible solution to a real-world Data Science problem

Transversal skills

• Demonstrate a capacity for creativity.
• Plan and carry out activities in a way which makes optimal use of available time and other resources.
• Write a scientific or technical report.

Teaching methods

• Hands-on lab sessions

• Homework assignments
• Final project

... using real-world datasets and Cloud Compute & Storage Services

Expected student activities

Students are expected to:
- STUDY: Attend the lab sessions
- WORK: Complete homework assignments
- ENGAGE: Contribute to the ineractive nature of the class
- COLLABORATE: Work in small groups to provide solutions to real-world problems
- EXPLAIN: Present ideas and results to the class

Assessment methods

- 60% continuous assessment during the semester
- 40% final project, done in small groups

Supervision

Office hours	Yes
Assistants	Yes
Forum	Yes

Resources

Virtual desktop infrastructure (VDI)
No

Bibliography

- **Python Data Science Handbook: Essential Tools for Working with Data** by Jake VanderPlas, O'Reilly Media, November 2016

- pyGAM - https://github.com/dswah/pyGAM

A list of additional readings will be distributed at the beginning of the course.

Ressources en bibliothèque
- Python Data Science Handbook: Essential Tools for Working with Data / J. VanderPlas

Websites
- http://www.datascience.ch