Technologies for Big Data with PYTHON

marco milanesio

MS DATA SCIENCE 2020-2021

UNIVERSITÉ CÔTE D'AZUR

Who am I?

- PhD [2010]
  - Distributed systems
  - Network measurements and performances
  - Distributed storage
- @Inria
  - Optimization
  - Image processing
  - Spark
  - Federated learning
- @MDLab
  - Dev-ops
  - Implementation - optimization
  - Virtualization

Inria
inventeurs du monde numérique

Epione
e-patient / e-medicine

UNIVERSITÉ CÔTE D'AZUR
CENTER OF MODELING
SIMULATION AND
INTERACTIONS

UNIVERSITÉ CÔTE D'AZUR
MEDICAL DATA
LABORATORY

Who are you?

- Curious, open minded
- You know how to use a PC 😎
- Wanting to learn some cool stuff
- Not feared of tackling problems
- Not feared by errors
- (optional) Some coding experience
- (bonus) Some Python experience
- (bonus) "LMGTFY" skills 😎

- Ideal profile:
  - 50% data scientist: exploit the data
  - 50% software developer: you need to code

Course Overview

- Introduction

- Recap on Python3
- Basic data analysis
  - builtins
- Advanced data analysis
  - scikit-learn
  - tensorflow
- The BigData picture
  - Apache Spark

Course Overview

- First part
    - Syntax, data structures, types
    - Software development
    - Data management and cleaning

- Second part
    - Machine learning
    - Principles of functional programming
    - Spark & (Spark Mllib [maybe])

Course Overview

- REPL + scripts + notebooks
- 10 lessons
  - ~ 40% lectures
  - ~ 60% lab sessions
- Evaluation (to be defined)