

# **Data-intensive Scalable Computing**

## **Introduction**

Pietro Michiardi

Eurecom

# Introduction and Motivations

# What is this Course About

- **The MapReduce Programming Model**

- ▶ Principles of functional programming
- ▶ Scalable algorithm design

- **In-depth description of Hadoop MapReduce**

- ▶ Architecture internals
- ▶ Software components
- ▶ Cluster deployments

- **Relational Algebra and High-Level Languages**

- ▶ Basic operators and their equivalence in MapReduce
- ▶ Hadoop Pig and PigLatin

# What is MapReduce?

- **A programming model:**

- ▶ Inspired by functional programming
- ▶ Parallel computations on massive amounts of data

- **An execution framework:**

- ▶ Designed for large-scale data processing
- ▶ Designed to run on clusters of commodity hardware

# What is Big Data?

- **Vast repositories of data**

- ▶ The Web
- ▶ Physics
- ▶ Astronomy
- ▶ Finance

- **Volume, Velocity, Variety**

- **It's not the algorithm, it's the data! [?]**

- ▶ More data leads to better accuracy
- ▶ With more data, accuracy of different algorithms converges