

# Large Scale Data Management

## Introduction

Silviu **Maniu**, LIG, Univ. Grenoble Alpes

## Logistics

Teaching **team**:

- Vania **Marangozova**, Prof. CS, UGA ([vania.marangozova@univ-grenoble-alpes.fr](mailto:vania.marangozova@univ-grenoble-alpes.fr))
- Silviu **Maniu**, Prof. CS, UGA ([silviu.maniu@univ-grenoble-alpes.fr](mailto:silviu.maniu@univ-grenoble-alpes.fr))

Two **complementary topics** (2 x 18 hours):

- Advanced Data Models and NoSQL (S. Maniu)
- Distributed Data Processing (V. Marangozova)

## Logistics

**Grading:**

- Project (25% of the final grade)
- Written Exam (75% of the final grade)

**Slot:** Fridays afternoons

**Site for the first part:** <https://silviu.maniu.info/teaching/>

## Topics Covered

**The challenges of Big Data** and distributed data processing

**Modern data models for Big Data**

- Relational data for analytics: Data Warehouses
- Principles of NoSQL and NoSQL systems

**Processing large amounts of data**

- Batch and stream processing systems

# The Data Deluge

## Many sources of data

- Sensors
- Social media
- Scientific experiments
- Industrial activity

# The Data Deluge

## Some numbers

Every 2 days, we create as much information as we did since 2013

- 90% of all data has been created in the last two years

**40K** search queries on Google every second

**45M** messages on WhatsApp every minute

**40 Billion** IoT devices by 2025.

**570** new web sites every minute

Largest database: **3.2 Trillion** rows (AT&T)

**40 TB** of data every second during an experiment at the Large Hadron Collider

# The Data Deluge

## Hardware capacity

### Storage

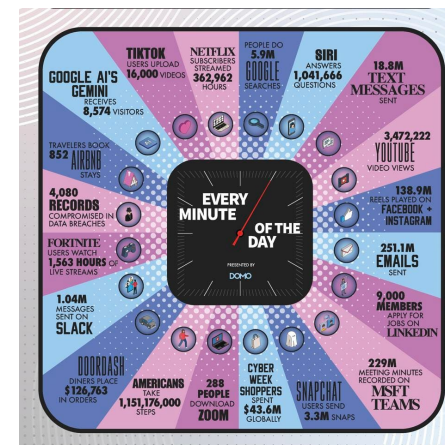
- All the music of the world stored for \$~500
- Large Amazon EC2 instance: 3.9TB of RAM, 8x7.5TB of SSD

### Computing resources

- Google data-centers: more than 2.5M servers (2016)
- Amazon capacity increase each day = size of Amazon in 2005

Huge opportunities for storing and processing data

# Today's Applications



# Large-scale Data Infrastructures



Figure 1: Google Data-center



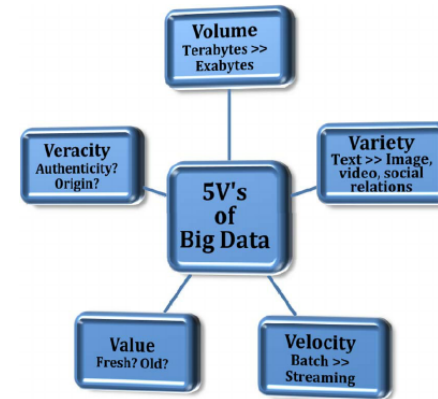
Figure 2: Amazon Data-center



Figure 3: Barcelona Supercomputing Center

# Big Data Challenges

## The V's



# Big Data Challenges

## The V's

**Volume:** Amount of data generated

**Variety:** all kinds of data are generated (text, image, voice, timeseries, etc.)

**Velocity:** Rate at which data are produced and should be processed

**Veracity:** Noise/anomalies in data, truthfulness

**Value:** How do we extract/learn valuable knowledge from the data

# Big Data Challenges

## The V's

In this course we are going to deal with:

**Volume Velocity Variety**

Questions to be answered:

- How to represent data in order to be able to analyze it?
- What properties do we need for systems in the Big Data era?
- How to build system and algorithms that can process huge amount of data in a timely manner?

# Main Resources

## Books

- [Designing Data-Intensive Applications](#). M. Klepmann (2018, new edition expected in 2026)
- [The Big Data Textbook](#). G. Fourny. 2025
- [The Data Warehouse Toolkit](#). R. Kimball, M. Ross. 2013

... and **scientific articles**, indicated in the respective lectures