BIG DATA COMPUTING

ID's last digit: 5 – 9

Francesco Silvestri

Department of Information Engineering

University of Padova

silvestri@dei.unipd.it

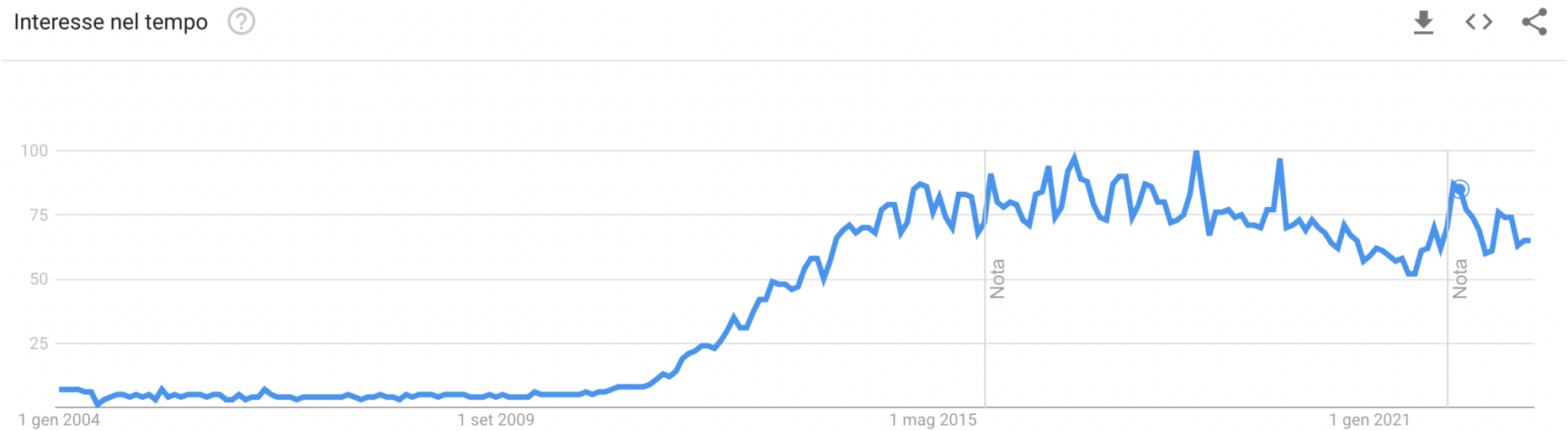
OUTLINE

1. Big Data Phenomenon
2. Computational Challenges
3. Organization of the Course
4. Administrative Issues

Who am I? Francesco Silvestri

- Experience:
  - 2006-2009: PhD University of Padova in Computer Engineering + visiting scholar University of Texas at Austin
  - 2010-2016: Post-doc University of Padova and IT University of Copenhagen
  - 2016-2019: Assistant professor at University of Padova
  - Since 2019: Associate professor at University of Padova
- Research:
  - Big data algorithms: how to efficiently extract information from big-data?
  - High performance algorithms: how to exploit modern computer architecture for big-data?
- Real life:
  - 3 kids + 1 wife
  - Love Denmark and biking cargo-bikes
  - Sports: barely jogging and swimming; play with kids.

Big Data Phenomenon



From: *Google Trends*

Big Data: term proposed in 2005 from Roger Mougalas (O'Reilly Media)

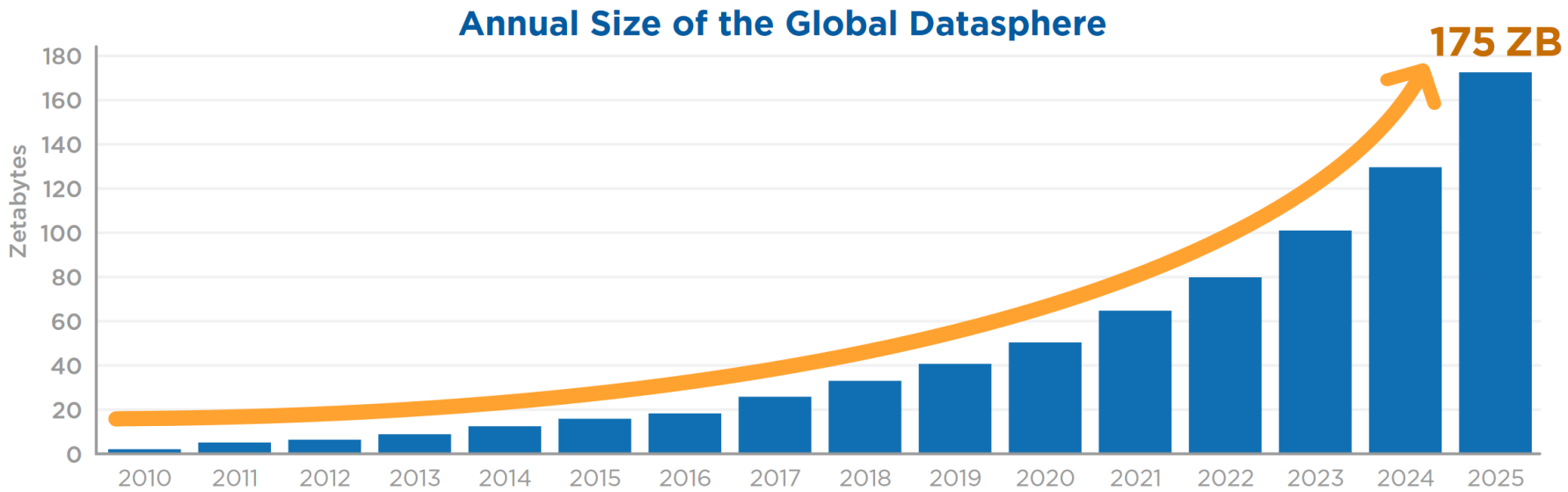
Big Data Phenomenon

"Space is big. Really big"
(*Douglas Adams, The Hitchhiker's Guide to the Galaxy*)

Why is DATA growing so much?

- Technological progress:
  - Growth of storage capacity
  - Growth of comunication bandwith
  - Growth of computing capacity
- Reduction of ICT costs
- Pervasiveness of digital technologies: scientific research, health, business, politics, social interactions, ...

Big Data Phenomenon



From: *The Digitization of the World (IDC, 2018)*

How big is 175ZB?:

- 1 ZettaByte (ZB) = 1 trillion GB = 10^{12} GB;
- 175 ZB $\equiv$ 23 parallel stacks of DVD from Earth to Moon;
- Downloading 175 ZB at 1Gb/s takes $>$ 43 million years

Big Data Phenomenon

The world continuously collects huge amounts of:

- Physical data: from sensors, telescopes, particle physics experiments.
- Biological/medical data: from genetic studies, patient monitoring, epidemic evolution analyses.
- Human activity data: from social networks, mobile devices, internet/web traffic, IoT systems.
- Business data: from online stores, customer profiling, bank/credit-card/financial services, quality-of-service monitoring.

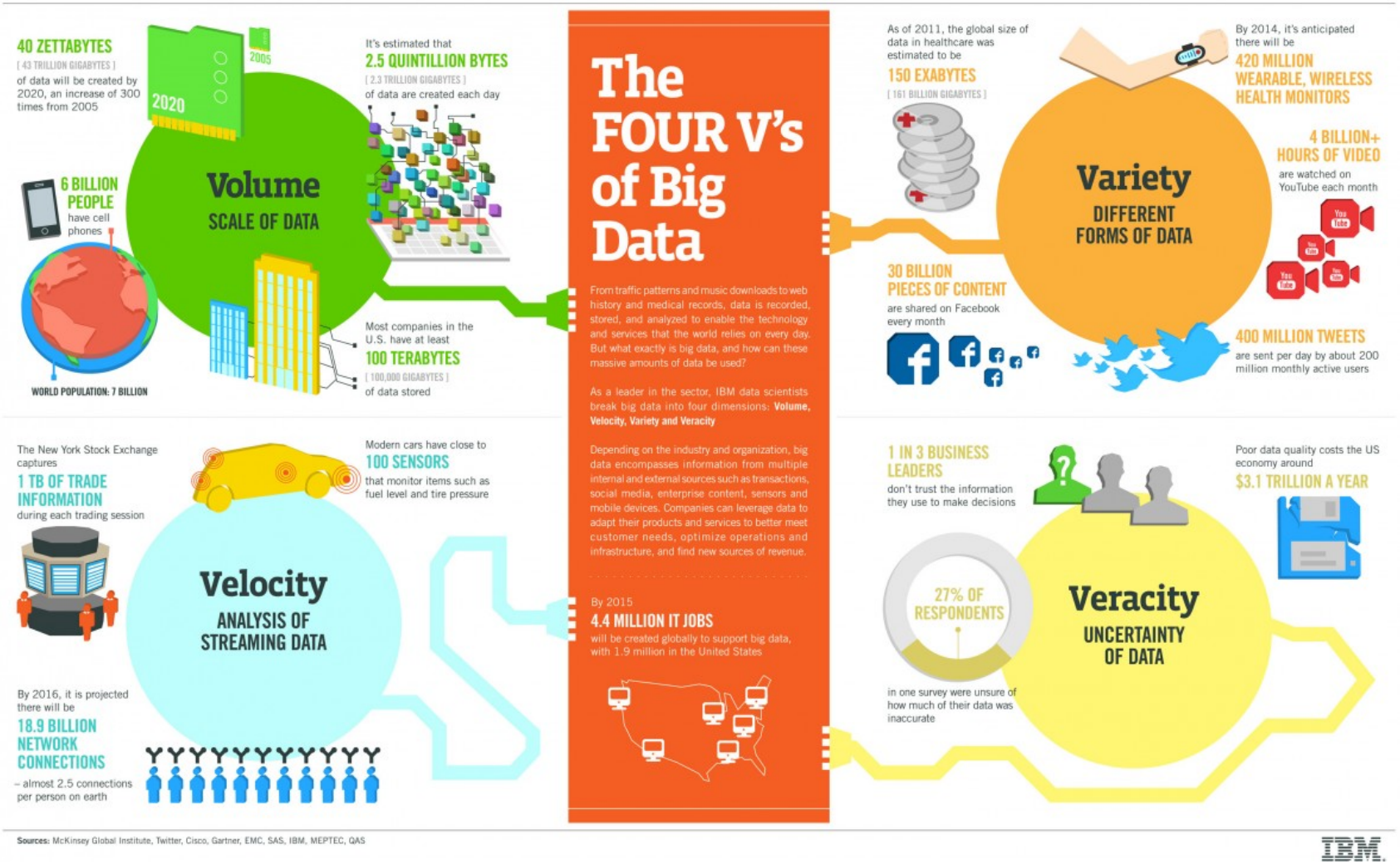
Big Data Phenomenon

The term Big Data relates to two distinct issues:

- ISSUE 1:
  - Data are produced everywhere;
  - Automated analytics are required (vs human inspection);
  - NEED: data selection/preparation procedures, adequate analysis tools.
- ISSUE 2:
  - Massive datasets must be processed;
  - Traditional (algorithmic) approaches are unsuited;
  - NEED: *feasible* and *efficient* methods to process massive data, novel computing frameworks.

This course focuses on ISSUE 2!

Computing Challenges



Source: IBM Big Data & Analytics Hub

Computational Challenges

- **Volume:** processing huge datasets poses several challenges and requires a data-centric perspective.
- **Veracity:** large datasets coming from real-world applications are likely to contain *noisy, uncertain data*, hence accuracy of solutions must be reconsidered.
- **Velocity:** sometimes, the data arrive at such a high rate that they cannot be stored and processed offline. Hence stream processing is needed.
- **Variety:** large datasets arise in *very different scenarios*. More effective processing is achieved by adapting to the actual characteristics of data.

The above issues require a

paradigm shift w.r.t. traditional computing.

Computational Challenges

To tackle the above challenges effectively, one needs:

- Platforms with:
  - High storage capacity and computing power
    $\Rightarrow$ parallel/distributed architectures
  - Moderate costs
  - Ease of programming and management
- Focus on accuracy-resource tradeoffs, to cope with size, noise, and uncertainty of data
- Data-centric view
- Data stream processing (sometimes)

Big Data Computing Course

What will we learn?

1. Novel computing/programming frameworks for big data processing: theory and practice
2. Key techniques to process large-scale data
   - Rigorous setting (provable guarantees)
   - Application to fundamental data analysis primitives

Specific topics

1. Frameworks: Distributed (MapReduce, Apache Spark) and Streaming
2. Techniques with applications (in parentheses):
   - Partitioning (data distribution)
   - Coresets (unsupervised learning);
   - Sketches (estimation of moments, set memebership)
   - Locality sensitive hashing (similarity search);

Organization of the Course

Subdivision into classes

Students of all programs are subdivided into two parallel classes based on their ID's last digit (*same syllabus, homeworks, and exams*)

- Class A (prof. Pietracaprina): last digit 0-4
- Class B (prof. Silvestri): last digit 5-9

Lectures

- Slide sets are made available in advance for each topic.
- *Attendance and active participation are strongly encouraged*.

Organization of the Course

Exam

- Homeworks: programming assignments (6+1 points)
  - Groups of 2-3 students (even from different classes)
  - 3 homeworks, approximately one every 3 weeks.
  - Use of Apache Spark on individual PCs and Cluster.
  - Bonus point if team registers by the deadline *and* all homeworks submitted by their deadlines.
- Final written exam (26 points)
  - Must be taken only after returning all homeworks!
  - To pass the exam: written exam ≥ 13 and final grade ≥ 18.
- Oral exam: at teacher's discretion, but compulsory if last homework returned after Session 1.

SEE DETAILED RULES IN THE COURSE MOODLE

Organization of the Course

Exam Sessions

Written exams are scheduled in the following dates (also found in the Course Moodle):

- Session 1: June 18 2024, 9am
- Session 2: July 15 2024, 2pm
- Session 3: September 5 2024, 9am
- Session 4: January/February 2025 (t.b.a)

IMPORTANT: *No additional exams sessions will be scheduled, independently of specific individual needs. It is the student's responsibility to organize her/his work and plan the exam well in advance.*

Organization of the Course

Required background

- Java or Python programming
- Basic algorithmics: asymptotic, worst-case analysis; fundamental algorithms and data structures; (e.g., lists, queues, stacks, hash tables, maps/dictionaries)
- Basic math tools, combinatorics, and probability.

Reference Textbooks

- J. Leskovec, A. Rajaraman and J. Ullman. Mining Massive Datasets. Cambridge University Press, 2014.
- A. Blum, J. Hopcroft, and R. Kannan. Foundations of Data Science. Cambridge University Press, 2020.

Administrative Issues

Online tools

- **Course Moodle:**

  https://stem.elearning.unipd.it/course/view.php?id=8801

  - Announcements and student forum.
  - Infos: contacts, textbooks, exam rules and sessions.
  - Lectures diary.
  - Material: slides, videos, exercises, articles.
  - Preliminary exams grades.

- **Uniweb:** Official exam lists and final grades.

- **Exam Moodle (only one for the two classes):**

  https://esami.elearning.unipd.it/course/view.php?id=5621

  - Formation of groups for Homeworks
  - Submission of Homeworks.

Administrative Issues

Contacts and office hours

- Teacher (prof. Francesco Silvestri): silvestri@dei.unipd.it
- TAs (Filippo Bragato and Mohammadmahdi Ghahramanibozandan): bdc-course@dei.unipd.it

Office hours are by appointment (via Email). Teaching assistants should be contacted by email and only for questions related to homeworks.

TODO: As soon as possible

- Register in the Course Moodle (no password required)
- Register in the Exam Moodle (password:)
- Form groups of size at most 3 for the homeworks **by March 15**. Once a group is formed, it must be registered in the Exam Moodle using the Group registration link. Bonus point needs registration by the deadline.