BIG DATA STATISTICS FOR BUSINESS

AY 2023-24

Giovanni De Luca
Parthenope University of Naples

Introduction to the course

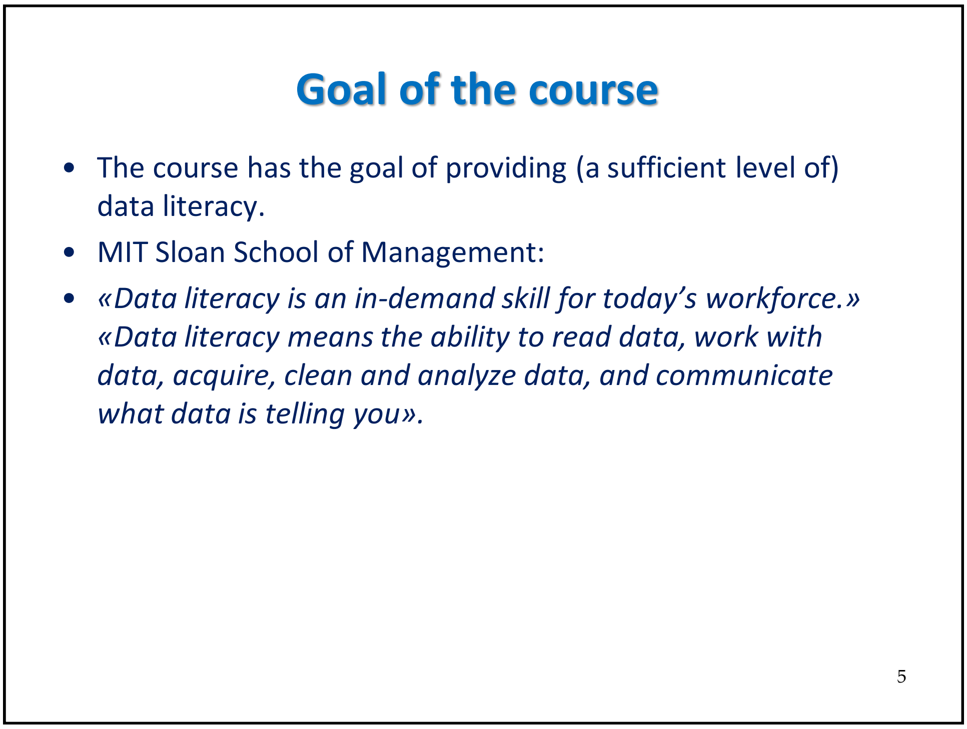
- We live in a society characterized by a large amount of data (data invasion, data avalanche).
- Where does the data come from?
- The development of multiple technologies has led to an exponential growth in the volume of data (firstly, social networks).

Introduction to the course

- The data are of great interest to companies.
- Data is said to be the new oil (or source of wealth) of the digital economy.
- However, a huge amount of data in itself is not information.
- This amount of data must be handled and analyzed so that it might transform into information.

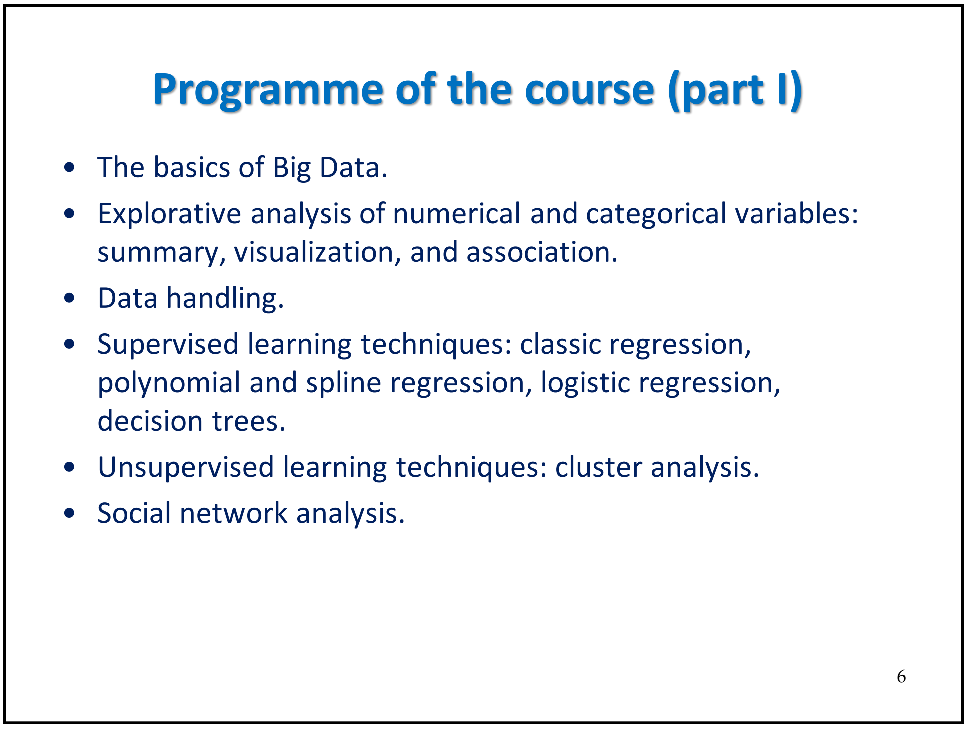
Introduction to the course

- Big data is more than high-volume data.
- We refer not only to very large datasets (actually, there is no minimum volume beyond which this term can be correctly used).
- We refer to complex data, too
- Not only numbers, but also images, videos, comments taken from social networks or GPS data.
- Size and complexity require specific tools.



Goal of the course

- The course has the goal of providing (a sufficient level of) data literacy.
- MIT Sloan School of Management:
- *«Data literacy is an in-demand skill for today's workforce.» «Data literacy means the ability to read data, work with data, acquire, clean and analyze data, and communicate what data is telling you».*



Programme of the course (part I)

- The basics of Big Data.
- Explorative analysis of numerical and categorical variables: summary, visualization, and association.
- Data handling.
- Supervised learning techniques: classic regression, polynomial and spline regression, logistic regression, decision trees.
- Unsupervised learning techniques: cluster analysis.
- Social network analysis.

Softwares

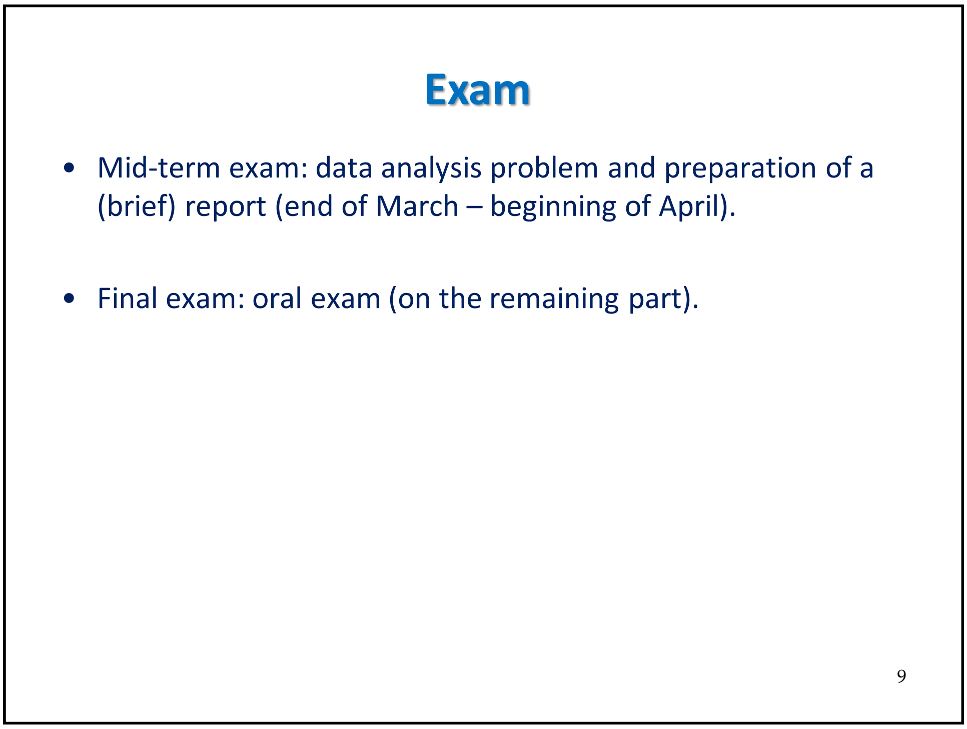
- Throughout the course, we will use two statistical softwares: Microsoft Excel and R.
- The R software can be found at https://cran.r-project.org
- RStudio is an integrated development environment (IDE) for R and can be found at https://rstudio.com

Textbooks

- James, Witte, Hastie, Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.

- Sedkaoui, Data Analytics and Big Data, Wiley, 2018.

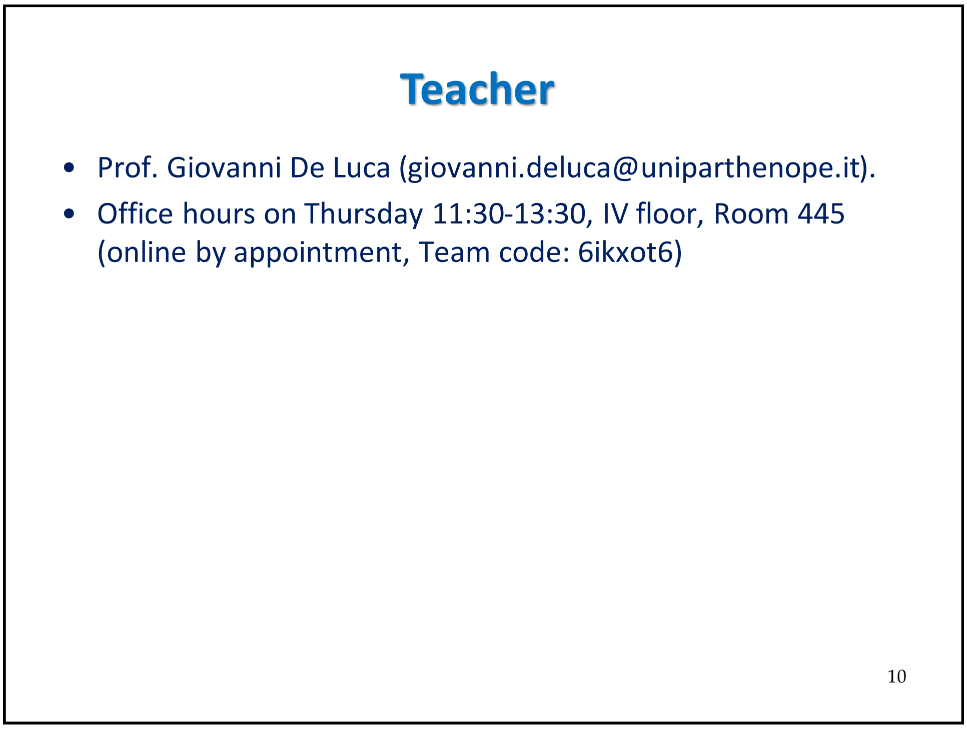
- Teaching material: slides, dataset, ecc.
  https://elearning.uniparthenope.it
  https://elearning.uniparthenope.it/course/view.php?id=2224



Exam

- Mid-term exam: data analysis problem and preparation of a (brief) report (end of March – beginning of April).
- Final exam: oral exam (on the remaining part).

9



Teacher

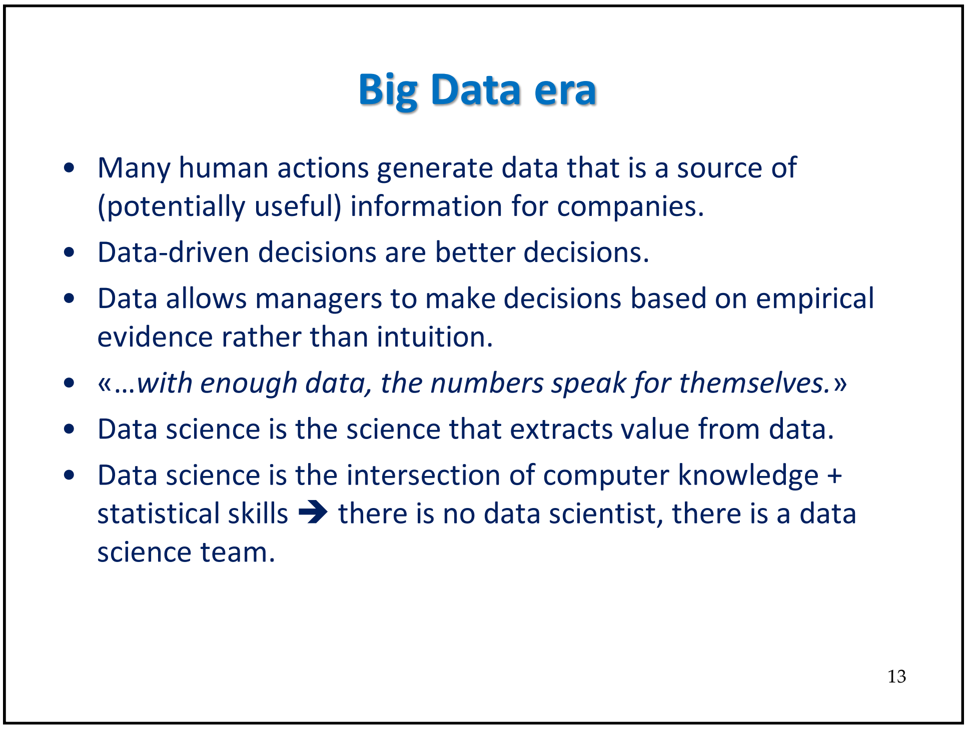
- Prof. Giovanni De Luca (giovanni.deluca@uniparthenope.it).
- Office hours on Thursday 11:30-13:30, IV floor, Room 445 (online by appointment, Team code: 6ikxot6)

10

The basics of Big Data

Big Data era

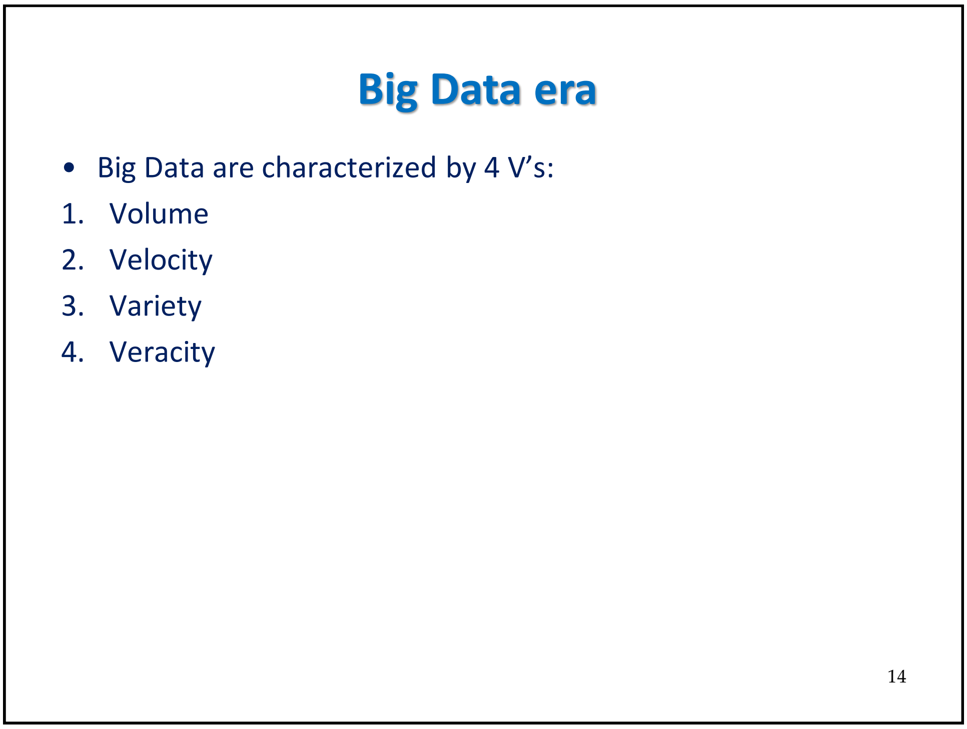
- The **Big Data** era is characterized by
  1. A large and complex quantity of data (“data avalanche”)
  2. The development of IT
- Big Data → better models → higher precision → smart and personalized business
- Examples: financial services, fraud detection, implementation of algorithms for trading, risk analysis, retail, CRM.



Big Data era

- Many human actions generate data that is a source of (potentially useful) information for companies.
- Data-driven decisions are better decisions.
- Data allows managers to make decisions based on empirical evidence rather than intuition.
- «…*with enough data, the numbers speak for themselves.*»
- Data science is the science that extracts value from data.
- Data science is the intersection of computer knowledge + statistical skills → there is no data scientist, there is a data science team.

13



Big Data era

- Big Data are characterized by 4 V's:

1. Volume
2. Velocity
3. Variety
4. Veracity

14

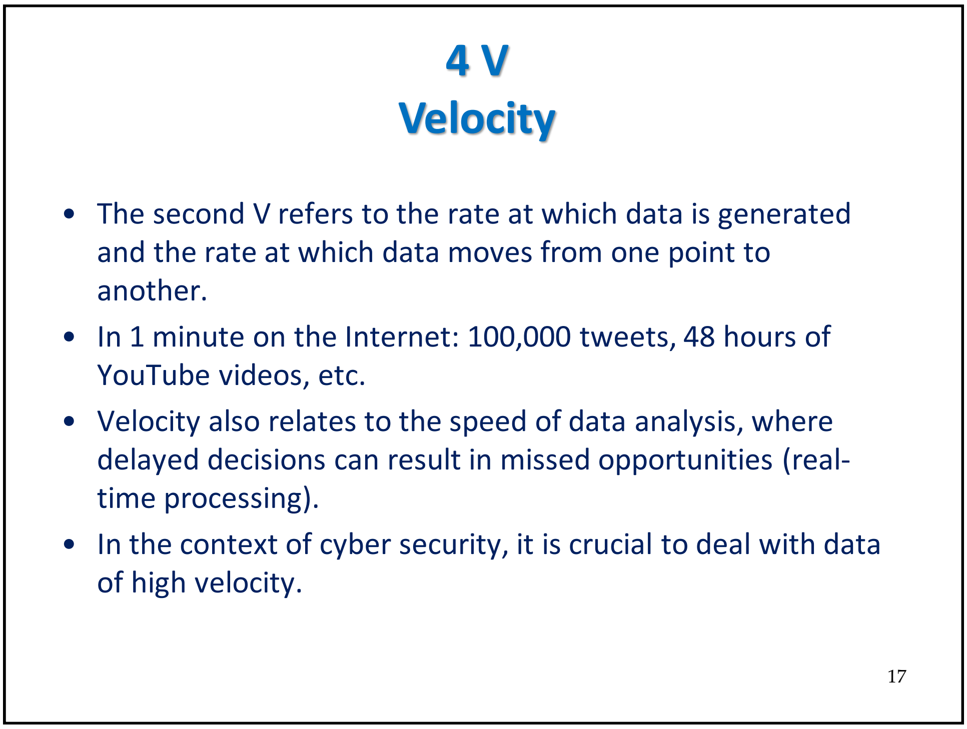
4 V
Volume

- The first V refers to the size of data generated minute by minute.
- It has been estimated that the data stored by the industries double every 1.2 years.
- Think about surveillance cameras installed in a major city, such as London.

4 V
Volume

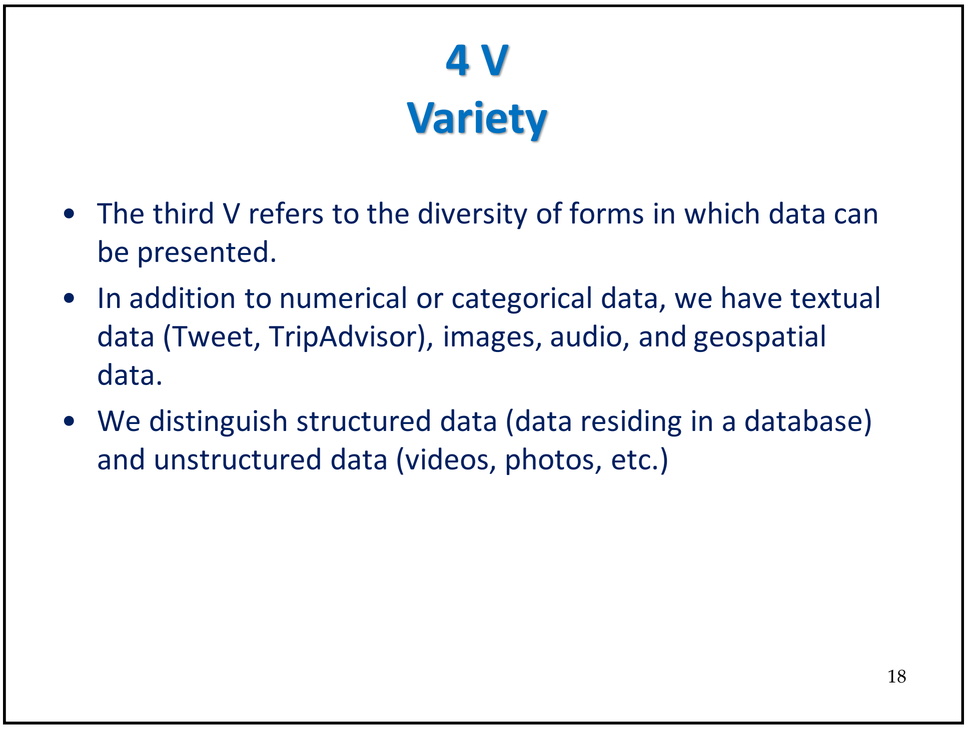
- According to the last Cisco Annual Internet Report:

- Nearly two-thirds of the global population will have Internet access by 2023. There will be 5.3 billion total Internet users (66% of global population) by 2023, up from 3.9 billion (51% of global population) in 2018.
- The number of devices connected to IP networks will be more than three times the global population by 2023. There will be 3.6 networked devices per capita by 2023, up from 2.4 networked devices per capita in 2018. There will be 29.3 billion networked devices by 2023, up from 18.4 billion in 2018.



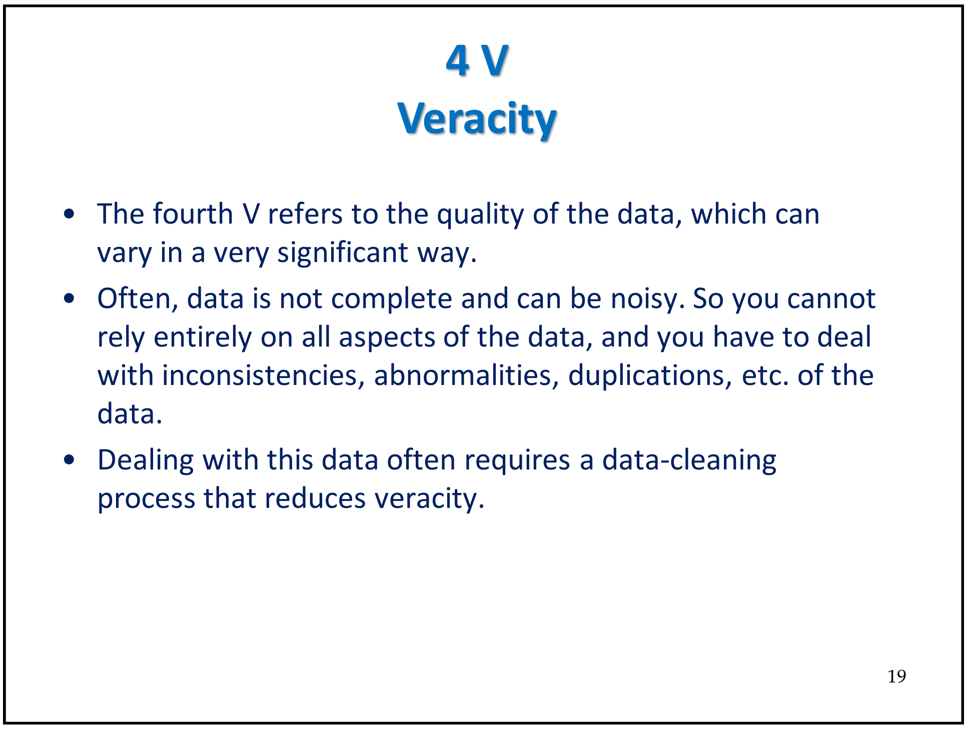
4 V
Velocity

- The second V refers to the rate at which data is generated and the rate at which data moves from one point to another.
- In 1 minute on the Internet: 100,000 tweets, 48 hours of YouTube videos, etc.
- Velocity also relates to the speed of data analysis, where delayed decisions can result in missed opportunities (real-time processing).
- In the context of cyber security, it is crucial to deal with data of high velocity.



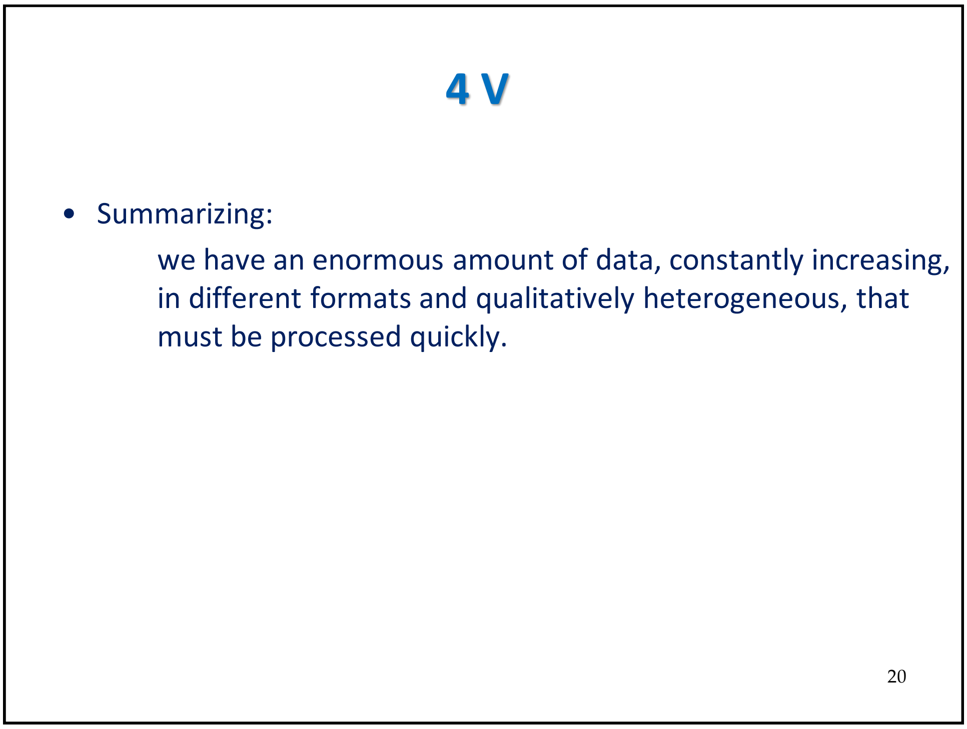
4 V
Variety

- The third V refers to the diversity of forms in which data can be presented.
- In addition to numerical or categorical data, we have textual data (Tweet, TripAdvisor), images, audio, and geospatial data.
- We distinguish structured data (data residing in a database) and unstructured data (videos, photos, etc.)



4 V
Veracity

- The fourth V refers to the quality of the data, which can vary in a very significant way.
- Often, data is not complete and can be noisy. So you cannot rely entirely on all aspects of the data, and you have to deal with inconsistencies, abnormalities, duplications, etc. of the data.
- Dealing with this data often requires a data-cleaning process that reduces veracity.



4 V

- Summarizing:

  we have an enormous amount of data, constantly increasing, in different formats and qualitatively heterogeneous, that must be processed quickly.