Big Data Analytics: Introduction

Mohamed Ndaoud

Presentation

- Assistant Professor of Statistics at ESSEC Business School
- Ex-member of the math department at University of Southern California (USC)
- PhD in theoretical statistics (X-ENSAE)
- Research interests:
  - High dimensional statistics
  - Robust statistics
  - Clustering
  - Learning theory
- Contact:
  - ndaoud@essec.edu
  - Office : Le Nautile, N324

10 INTERESTING FACTS ABOUT BIG DATA IN 2020

Each user will generate **1.7 MB** of new information per second.

Our data universe will go from 4.4 zettabytes to **44 ZB** (44 billion GB).

There will be more than **50,000 million** smart devices connected around the world, ready to gather and analyse data.

Big data technology solutions will bring **€206,000 M** to the European economy, representing an increase of almost 2% of the GDP.

The big data business will have a value of approximately **$9,400 M,** wich represents 10% of the global market of information management tools.

8 million professionals specialising in big data will be needed.

The amount of information generated by companies **will increase 75-fold,** so only the IT staff will increase 1.5-fold.

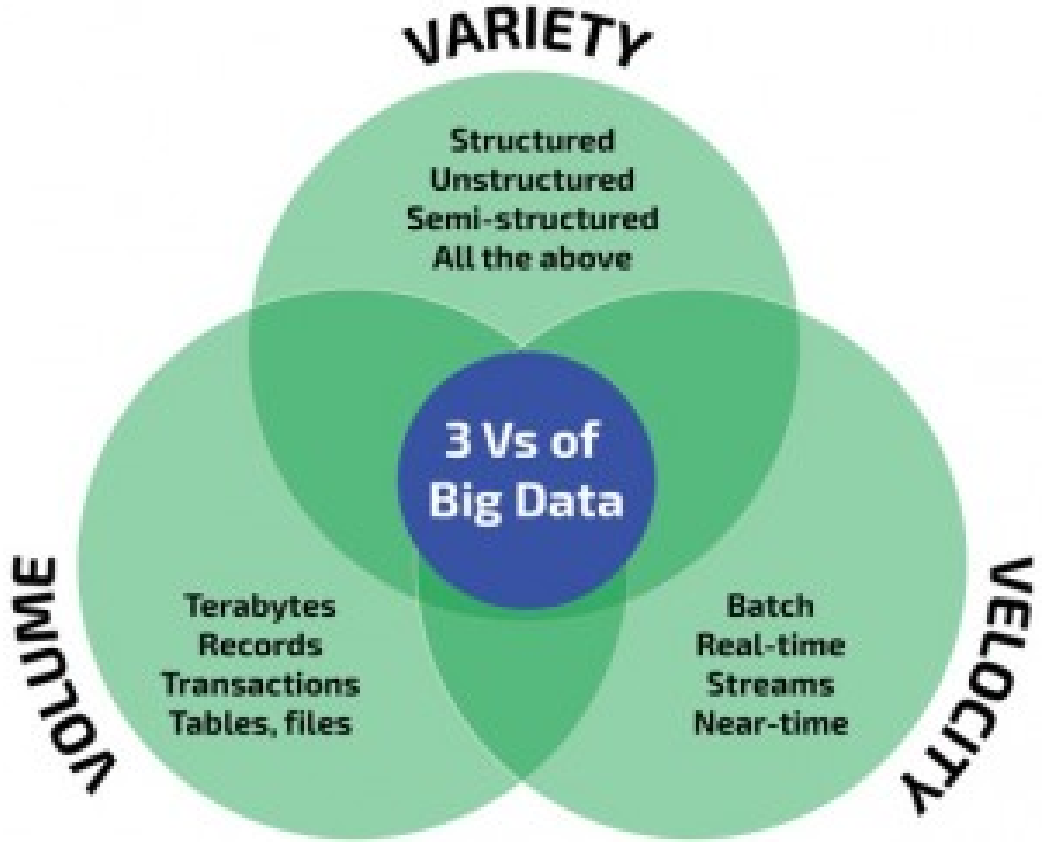
Each person will generate **1.5 GB** of data every day, which is equivalent to:

- **1.5** million messages
- **750** images
- Listening to music for an entire day
- **4** hours of videos

How big is Big Data?

- If the data can not be handle in "reasonable" or "useful" time by a system composed of a single node
- **Big Data is not only the size of the data set but also the speed of the processing**

Three V's of BigData



Veracity

- Big Data is messy
- Poor quality of data:
  - the source of information may not be reliable
  - human or technical failure
  - the data provided may also be intentionally wrong

Big Data needs to be sifted and organized by quality.

Ranking and Collaborative filtering

- Online commerce and advertisement: anticipating user tastes
- Companies routinely collect user rankings for various products
- Goal: predict user's preference.

Ranking and Collaborative filtering

- Online commerce and advertisement: anticipating user tastes
- Recommender systems are susceptible to manipulations
- Web companies need to deal with users who subvert the system in various way


POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
OFFICER ALGORITHM
NEW WAYS OF SEEING
a Gallery of Extraordinary
PLUS
AND
Lawrence Weschler
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

To extract the knowledge

Data needs to be

- Stored
- Managed
- **ANALYZED** (this class)

Data mining $\approx$ Big Data $\approx$ Predictive Analytic $\approx$ Data Science

What is Data Mining?

- Given lots of data
- Discover patterns and models that are:
  - **Valid:** hold on new data with some certainty
  - **Useful**
  - **non-obvious**
  - **Understandable:** humans should be able to interpret the pattern

Data Mining Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example**: Clustering
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example**: Recommender systems

Lifecycle of an Analysis Project

- **Clarify**
  - Become familiar with the data
  - Template a solution

Lifecycle of an Analysis Project

- Clarify
  - Become familiar with the data
  - Template a solution
- **Develop**
  - Create a working model

Lifecycle of an Analysis Project

- Clarify
  - Become familiar with the data
  - Template a solution
- Develop
  - Create a working model
- **Productize**
  - Automate and integrate

Lifecycle of an Analysis Project

- Clarify
  - Become familiar with the data
  - Template a solution
- Develop
  - Create a working model
- Productize
  - Automate and integrate
- **Publish**

DATA > RAM

Lifecycle of an Analysis Project

- **Subset**
  - Extract data to explore, work with
- Clarify
  - Become familiar with the data
  - Template a solution
- Develop
  - Create a working model
- Productize
  - Automate and integrate
- Publish

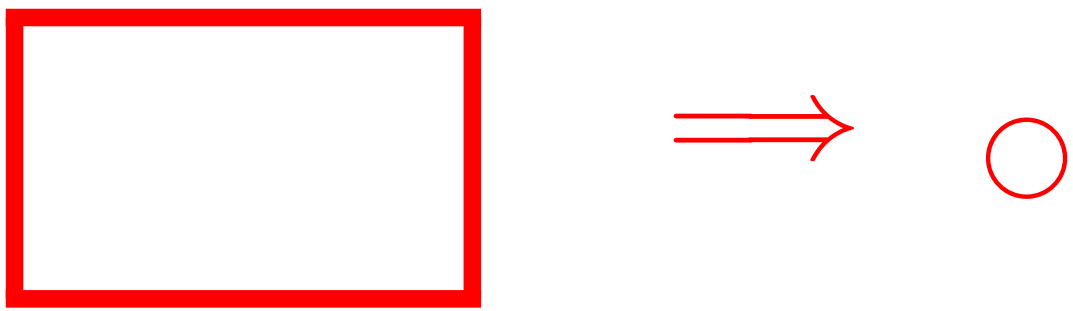
Lifecycle of an Analysis Project

- Subset
  - Extract data to explore, work with
- Clarify
  - Become familiar with the data
  - Template a solution
- Develop
  - Create a working model
- Productize
  - **Scale up the model to the entire data set**
  - Automate and integrate
- Publish

What are the problems
that analysts have with Big Data?

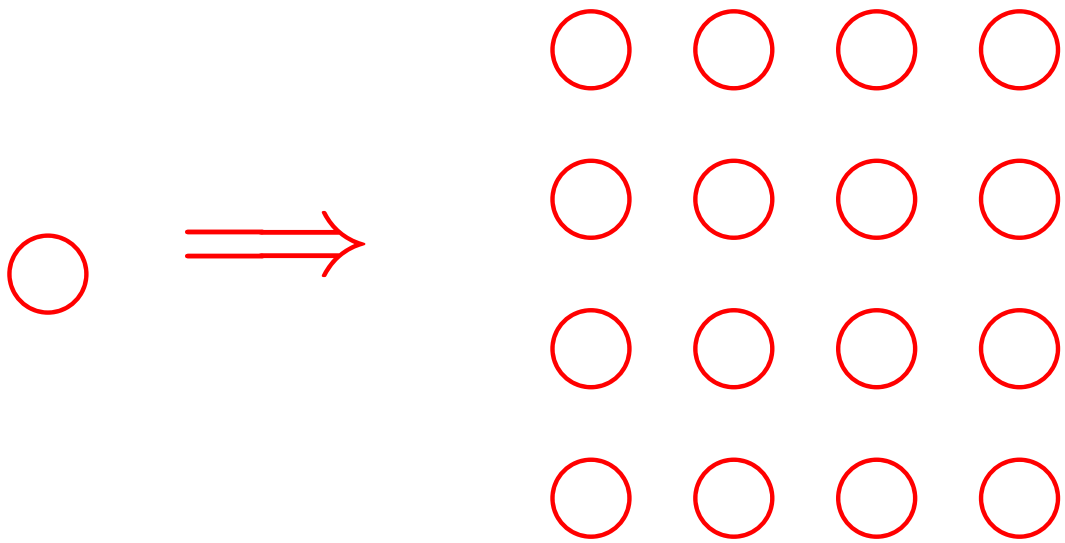
Analytic Big Data Problems

- **Class 1. Extract Data**
  - Problems that require you to extract a subset, sample, or summary from a Big Data source
  - You may do further analytics on the subset, and the subset might itself be quite large.



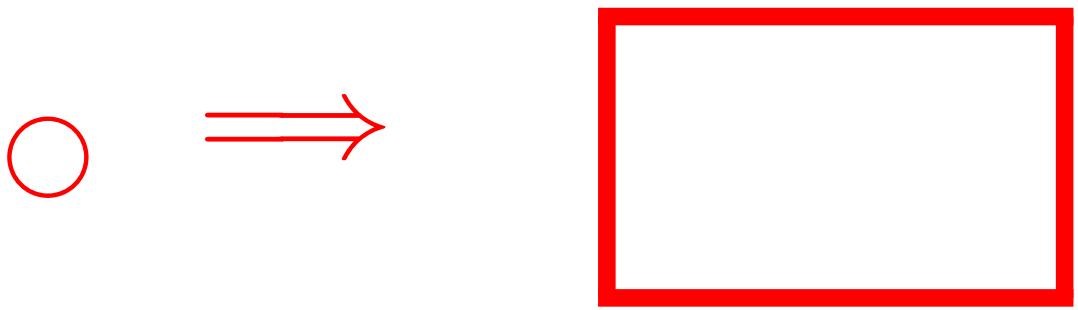
Analytic Big Data Problems

- **Class 2. Compute on the parts**
  - Problems that require you to repeat computation for many subgroups of the data.
  - You may aggregate the results once finished.

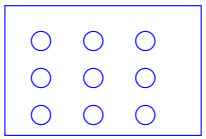


Analytic Big Data Problems

- **Class 3. Compute on the whole**
  - Problems that require you to use all of the data at once.
  - These problems are irretrievably big; they must be run at scale within the data warehouse.



Lifecycle of an Analysis Project

- Subset: **Class 1**  $\Longrightarrow$ 
- Clarify
- Develop
- Productize: **Class 2 and 3**  $\Longrightarrow$ 
- Publish

Data Mining: Cultures

- Data mining overlaps with
  - Databases: Large-scale data, simple queries
  - Machine learning: Small data, Complex models
  - CS Theory: (Randomized) Algorithms
- Different cultures:
  - DB: queries that examine large amounts of data
  - ML: data - mining is the inference of models
    - Result is the parameters of the model

This class stress on

- Automation for handling large data
- Algorithms
- Computing architectures

What we will learn?

- Mine different types of data:
  - high dimensional
  - graphs
  - infinite/never-ending
  - labeled
- Solve real - world problems:
  - Recommender systems
  - Market Basket Analysis
  - Spam detection
  - Duplicate document detection
- Various "tools":
  - Linear algebra (SVD)
  - Hashing (LSH, Bloom filters)

LSH(相似性查找)：Locality-Sensitive Hashing
Bloom Filters（高效检索）

Readings

- J. Leskovec, A. Rajaraman and J. D. Ullman
  Mining of Massive Datasets (2014)
  Free online: http://www.mmds.org
  Chapter 1 "Data Mining"
- Learning Spark by Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia.

Course Logistic

- Course website:
  - Lecture slides + notes
  - Exercises and solutions
- 2 Homeworks 50%
  - Theoretical and programming questions
- Final exam 50%
  - December 12th
- Final grade=
  max(Final exam; 0.5×Final exam +0.5× Homeworks)