



Algorithms for Data Science

Introduction

Silviu Maniu

September 11th, 2020

Université Paris-Saclay

Table of contents

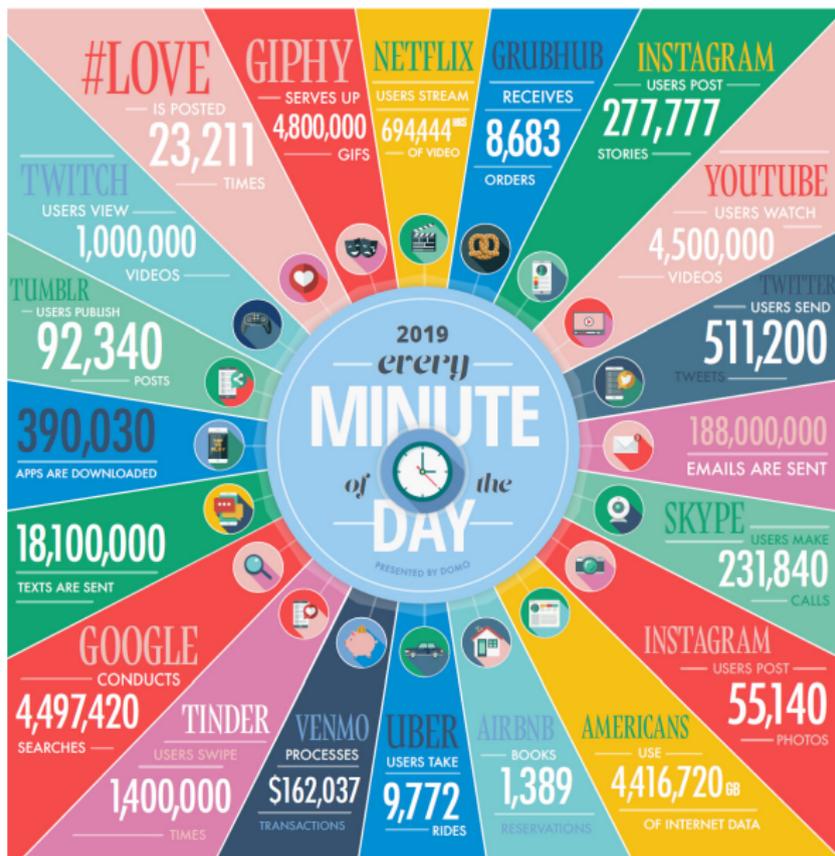
Data Mining

Course Objectives

Data = knowledge and value

In order to get the knowledge, data needs to be:

- stored
- managed
- **analyzed** – *the objective of this course*



Data Mining – use the most powerful hardware and the *most efficient algorithms* to analyze data

What does **analyze** the data mean?

– to discover **patterns** and **models**:

- valid, useful, unexpected (i.e., not trivial to find), understandable

A term that can be used interchangeably with **Big Data** or **data science**

Data Mining: Models

Data Mining relies on **modeling** the data – a way to generalize the knowledge we have **beyond the samples of data we currently have**

Two main views:

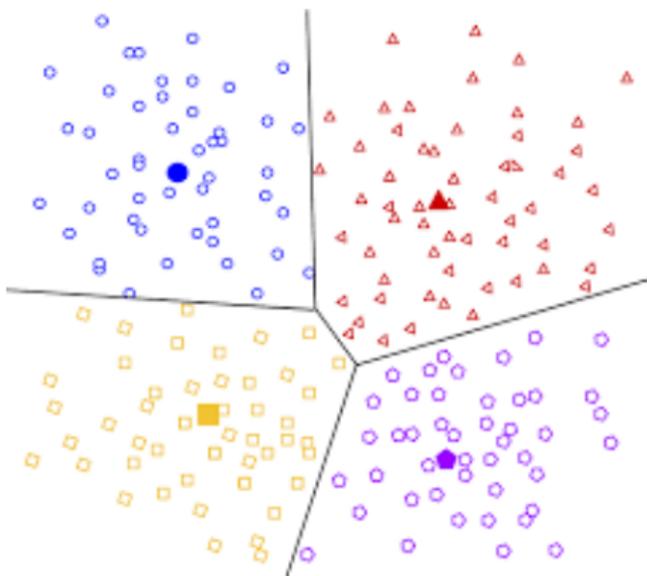
1. **statistical**: constructing a statistical model (e.g, fitting a distribution to data), machine learning (more complex functions over data) – look at data mining as a **statistical problem**
2. **computational**: summarization (summarizing the data succinctly and approximately), feature extraction (keeping only the most relevant features of the data)

– we care mostly about the **computational** case in this course

Data Mining: Tasks

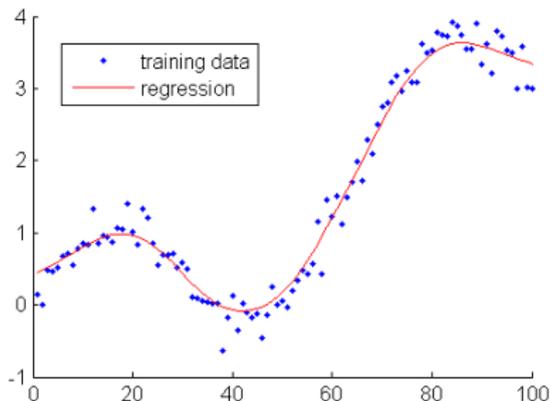
Descriptive methods: find interpretable patterns that aim to describe the data

- clustering, PageRank



Predictive methods: use features of the data to predict future values

– e.g., regression, recommendation



Data Mining Cultures

Large overlap of **Data Mining** with:

- **databases**: simple queries (SQL, based on mathematical logic), large data
- **machine learning**: complex models, (relatively) small data
- **theoretical CS**: randomized and approximation algorithms

In **databases**, data mining is large-scale analytic processing via queries – results are query answers

In **machine learning**, data mining reduces to **inferring models** – results are the model parameters

Dangers of Data Mining

If the analysis is not careful, a data analyst can use data mining and find **patterns that are meaningless**

Bonferroni's principle

- **informally**: the more data you have, the more likely you are that you find some pattern in the data (as it is more likely to occur randomly) – **bogus** (random data will always have patterns in it)
- **principle**: if the expected number of occurrences of the events you are looking for is significantly larger than the number of real instances – then *almost anything you will find is bogus*

One has to look only at events / data that are too rare to likely occur in random data.

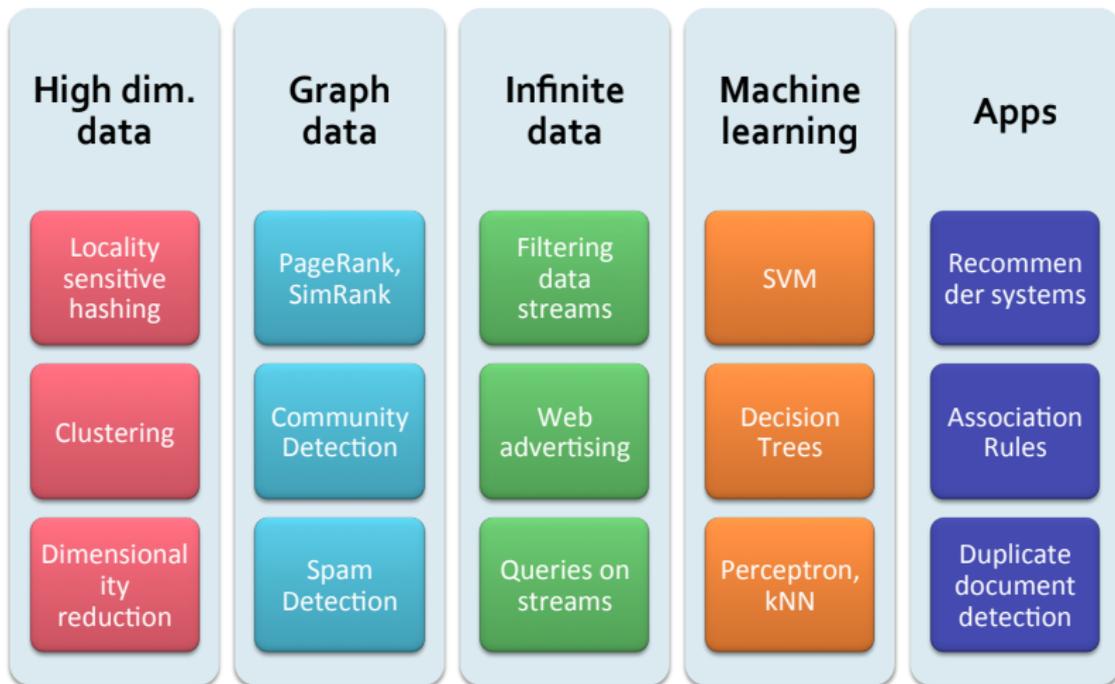
Another major danger: **the privacy-usefulness** tradeoff of using data to make decisions

Table of contents

Data Mining

Course Objectives

The Big Picture



Mining of Massive Datasets <https://www.mmds.org/>

Objectives

To present **data analysis methods** that are scalable to big data instances, *from a computational point of view*

Main topics in lectures:

- **item mining**: frequent items, finding similar items
- **advertising and recommendation on the Web**
- **data stream mining**

Structure

6 weeks of lectures/labs + 1 week of lab/evaluation; Fridays afternoon
14:00 – 17:45

Practical labs: follow lectures in alternate weeks (starting Week 2);
Python via Jupyter notebooks

Evaluation:

- **50% project** (programming assignment) – starting Week 4
- **50% in-lab evaluation(s)** – week(s) to be decided, but not before Week 4; replaces exam due to sanitary situation

Links

Class links: <http://silviu.maniu.info/teaching> and also on eCampus [https://ecampus.paris-saclay.fr; homeworks/deadlines](https://ecampus.paris-saclay.fr;homeworks/deadlines)
ony via eCampus!

Textbook: Mining of Massive Datasets, available at <https://www.mmds.org/>

Acknowledgments

The contents partly follows Chapter 1 of [Leskovec et al., 2020].

<https://www.mmds.org/>

-  Leskovec, J., Rajaraman, A., and Ullman, J. (2020).
Mining of Massive Datasets.
Cambridge University Press.