universidade de aveiro 

Mining Large Scale Datasets

2023-2024

(Adapted from CS246@Starford.edu; http://www.mmds.org)

Sérgio Matos - aleixomatos@ua.pt

Data Mining

- Data is everywhere
- Data contains knowledge ... and value

  Whether it is to improve health/well-being, sell stuff, or win elections :-)

- Data Mining = Extract knowledge from data

Data Mining

- Data is everywhere
- Data contains knowledge ... and value

  Whether it is to improve health/well-being, sell stuff, or win elections :-)

- Data Mining = Extract ACTIONABLE knowledge from data

Data Mining

Data Mining = Extract actionable knowledge from data

Given lots of data, discover patterns and models that are

- **Valid**: hold on new data with some certainty
- **Useful**: should be possible to act on the item
- **Unexpected**: non-obvious to the system
- **Understandable**: humans should be able to interpret the patterns

Data Mining

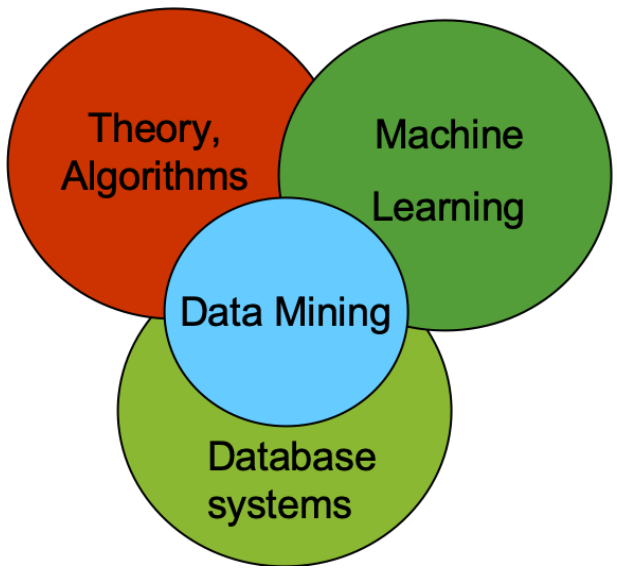
Data Mining = Extract actionable knowledge from data

- 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

Data Mining

Data Mining = Extract actionable knowledge from data

Some machine learning... but not only that



Data Mining

To extract knowledge, data needs to be

- Stored
- Managed
- Analysed

Data Mining

To extract knowledge, data needs to be

- Stored
- Managed
- **ANALYSED** $\leftarrow$ Focus of this course

We won't deal (much) with storing/managing
We won't cover ethics and privacy... very relevant aspects in DM

This class: MLSD

- Emphasis on **algorithms that scale**
  - Parallelization often essential
- Focus on
  - **Scalability** (big data)
  - **Algorithms**
  - Automation for handling **large data**
  - Use of computing architectures

This class: MLSD

- Different types of data:
  - High dimensional
  - Graphs
  - Streams of "infinite" data
  - Labeled data
- Different models of computation:
  - MapReduce
  - Streams and online algorithms
  - Single machine in-memory

This class: MLSD

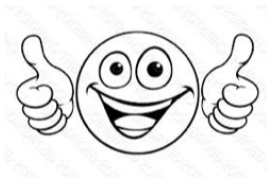
- Solve real-world problems:
  - Recommender systems
  - Market Basket Analysis
  - Spam detection
  - Duplicate document detection

- Learn/apply various "tools":
  - Linear algebra (SVD, Rec. Sys., Communities)
  - Optimization (stochastic gradient descent)
  - Dynamic programming (frequent itemsets)
  - Hashing (LSH, Bloom filters)

Course organization

- Theoretical exposition of algorithms and strategies
- Practical assignments based on Spark/Hadoop
- Python (or Java/Scala) is essential

- Self-study is highly encouraged.



Grading

- 3 practical assignments = 60%
- Final exam = 40%

The final exam will occur during the exam period ('Época normal').

Bibliography

"Mining of massive datasets". Leskovec, Rajaraman, Ullman, 2014.
http://www.mmds.org/

"Networks, Crowds, and Markets: Reasoning About a Highly Connected World". Easley, Kleinberg, 2010.
http://www.cs.cornell.edu/home/kleinber/networks-book/

"Spark: The Definitive Guide". Chambers, Zaharia, 2018.

"Data Algorithms with Spark". Parsian, 2022.

"Advanced Analytics with PySpark", Tandon et al., 2022.