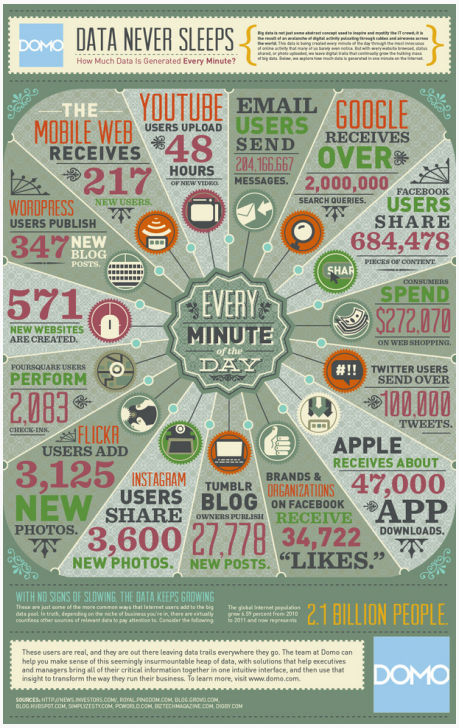
Introduction to Data Mining

Rita P. Ribeiro

Data Mining I - 2023/2024

Data is Everywhere



"We are drowning in data, but starving for knowledge."

From Data to Knowledge

- Data
  - Facts, numbers, or text that can be processed by a computer.
- Metadata
  - Data about the data itself such as logical database design or data dictionary definitions.
- Information
  - The patterns, associations, or relationships among all this data can provide information.
- Knowledge
  - Information can be converted into knowledge about historical patterns and future trends.

From Data to Knowledge

- Criteria to assess Knowledge:
  - correctness (probability, success in tests);
  - generality (domain and conditions of validity);
  - usefulness (relevance, predictive power);
  - comprehensibility (simplicity, clarity, parsimony);
  - novelty (previously unknown, unexpected)

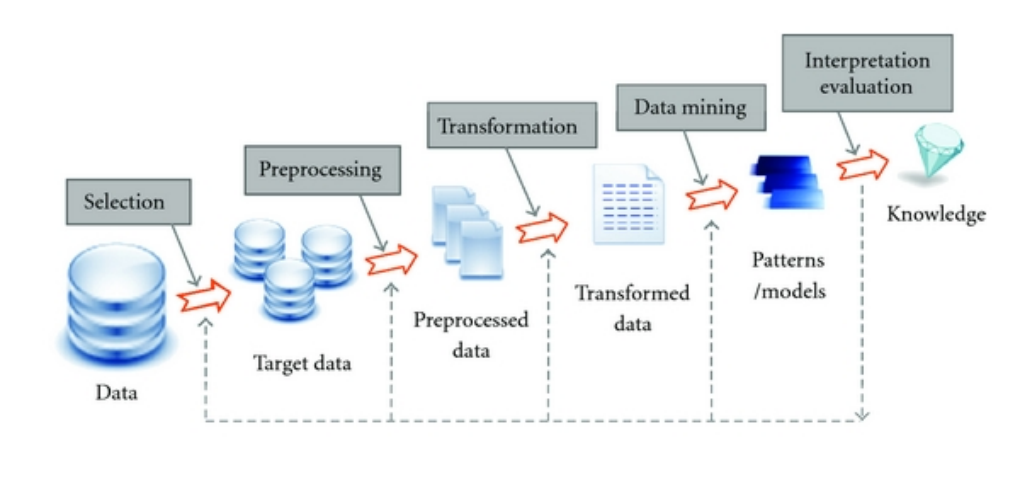
Data Mining is the process of discovery of knowledge from data!

Data Mining Definitions

- *"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner." — Hand, Mannila, Smyth, 2001*
- *" It is the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions." – Sumathi, Sivanandam, 2006*
- *"Data mining is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data." – Charu C. Aggarwal, 2015*

Data Mining: the core of KDD

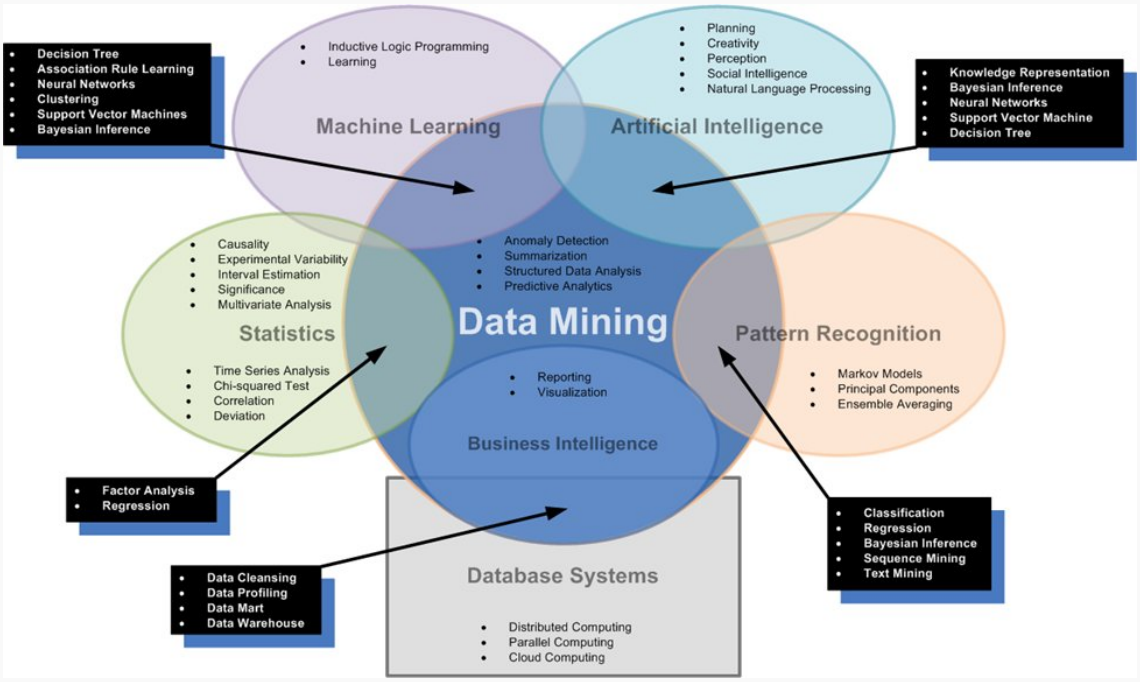
Knowledge Discovery from Data (KDD)



Data Mining and Big Data

- Data is being collected everywhere: transactions, social media, sensors, satellites, etc.
- Big Data has three dimensions described by the Three V's Gandomi and Haider, 2015:
  - Volume: massive, high dimensional, distributed data sets
  - Velocity: generated at high-speed
  - Variety: heterogeneous, complex
- Traditional techniques may be unsuitable due this kind of data.
- A key challenge for data mining is to develop techniques that can cope with Big Data.

Data Mining: a Multidisciplinary Area



{(source: data-analytics.swri.org)}

Some Data Mining Applications

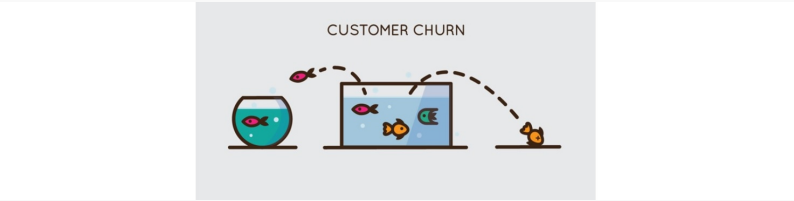
- Market management
  - Target marketing, customer relationship management, market basket analysis, cross-selling, market segmentation, trend analysis.

Some Data Mining Applications (cont.)

- Financial Industry, Banks, Businesses
  - Profitability analysis, risk management, sales forecasting, stock and investment analysis, customer retention.

Some Data Mining Applications (cont.)

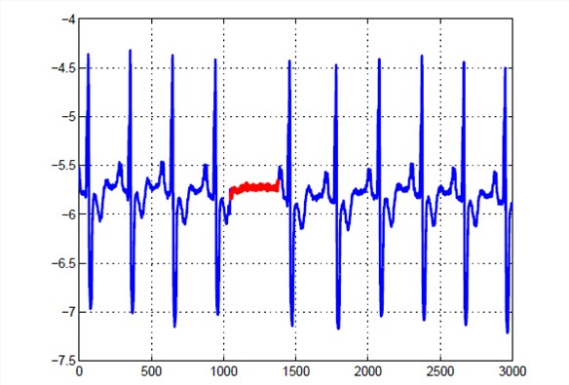
- Telecommunications and media
  - Response scoring, marketing campaign management, profitability analysis, fraud detection, and customer segmentation.



source: https://jtsulliv.github.io/churn-prediction/

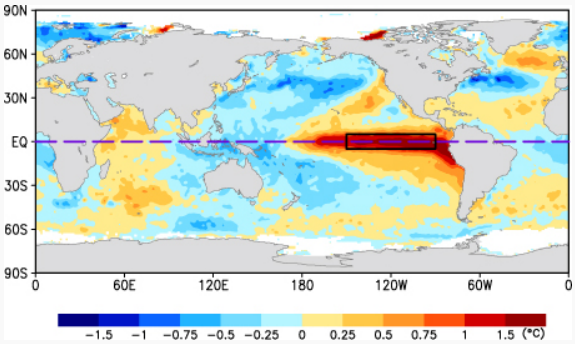
Some Data Mining Applications (cont.)

- Medicine, Pharmaceutical Companies, Health care
  - medical diagnosis, drug development, identify successful medical therapies, predict office visits, claim analysis, assisting health insurance organizations dealing with fraud.



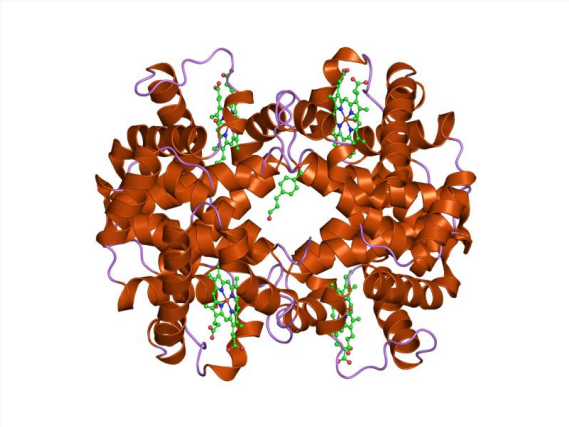
Some Data Mining Applications (cont.)

- Earth Sciences
  - Support climate change studies, forecast anomalous weather patterns, impacts in marine ecossystems, ensure seafood sustainability, predict harmful algae blooms, etc.



Some Data Mining Applications (cont.)

- Bioinformatics, Physics, ...
  - Microarray gene expression, protein family classification, sequence-based analysis, biochemical analysis, star galaxy classification.



CRISP-DM: a Typical Data Mining Workflow

- Cross-Industry Process for Data Mining (CRISP-DM)

Shearer C.: The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22.

CRISP-DM: Business Understanding

- Determine Business Objectives:
  background, business objectives and success criteria
- Assess Situation:
  inventory of resources, requirements, assumptions and constraints, risks and contingencies, terminology, costs and benefits
- Determine Data Mining Goals:
  data mining success criteria
- Produce project plan:
  project plan, initial assessment of tools and techniques

CRISP-DM: Data Understanding

- Collect Initial Data:
  initial data collection report
- Describe Data:
  data description report
- Explore Data:
  data exploration report
- Verify Data Quality:
  data quality report

CRISP-DM: Data Preparation

- Data Set:
  data set description
- Select Data:
  rationale for inclusion/exclusion
- Clean Data:
  data cleaning report
- Construct Data:
  derived variables, generated records
- Integrate Data:
  merged data
- Format Data
  reformatted data

CRISP-DM: Modeling

- Select Modelling Technique(s):
  modelling technique(s) and assumptions
- Define Test Design
- Build model(s):
  parameter settings, model description
- Assess Model(s):
  model assessment through experimental test, revise parameter settings

CRISP-DM: Evaluation

- Evaluate Results:
  assessment of data mining results w.r.t. business success criteria, approved models
- Review Process
- Determine Next Steps:
  list of possible actions, decision

CRISP-DM: Deployment

- Plan Deployment
- Plan Monitoring and Maintenance
- Produce Final Report:
  final presentation
- Review Project:
  experience documentation

Key Issues in a Data Minig Project

- Data Structure
  - what to measure? pre-processing steps?
- Model Structure
  - what type of model(s) should we build?
- Score Function
  - how to evaluate the obtained models?
- Optimisation and Search Method
  - how to search and optimise the models in the context of the selected structure?
- Data Management Strategy
  - how to handle the data efficiently during model construction and evaluation?

References

References

Aggarwal, Charu C. 2015. *Data Mining, the Texbook*. Springer.

Gandomi, Amir, and Murtaza Haider. 2015. "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management* 35 (2): 137–44. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques*. 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Moreira, João, Andre Carvalho, and Tomás Horvath. 2018. *Data Analytics: A General Introduction*. Wiley.