Knowledge Discovery from Data Bases
Computing Paradigms

KDD team
(contact: jgama@fep.up.pt)

LIAAD-INESC TEC, University of Porto, Portugal

September 2024

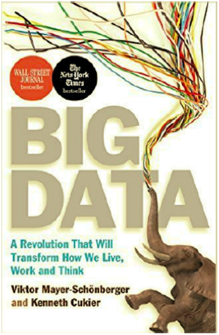
KDD: technologies that will change the world

Big data: The next frontier for innovation, competition, and productivity Report McKinsey Global Institute

McKinsey Global Institute

June 2011

Big data: The next frontier for innovation, competition, and productivity



WALL STREET JOURNAL bestseller

The New York Times bestseller

BIG DATA

A Revolution That Will Transform How We Live, Work and Think

Viktor Mayer-Schönberger and Kenneth Cukier

See the differences

- Machine learning used to take place behind the scenes:
  - Amazon mined your clicks and purchases for recommendations,
  - Google mined your searches for ad placement,
  - Facebook mined your social network to choose which posts to show you.
- Nowadays, machine learning is on the front pages of newspapers, and the subject of heated debate:
  - Learning algorithms drive cars,
  - translate speech,
  - Robots in Mars
  - Watson win at Jeopardy!

Team

- Rita P. Ribeiro, Univ. Porto (rpribeiro@fc.up.pt) (coordinator)
- João Gama, Univ. Porto (jgama@fep.up.pt)
- Paulo Azevedo, Univ. Minho (pja@di.uminho.pt)
- Bruno Veloso, Univ. Porto (bveloso@fep.up.pt)
- and, of course, Students

Contacts:

- LIAAD-INESC TEC, R. Dr Roberto Frias, 378; Porto
- www.liaad.up.pt

Program I

- Basic Concepts
  - Predictive Learning:
    - Distance-based, Search based and optimization-based algorithms
    - Ensemble models
    - Deep Learning
    - Evaluation Learning Algorithms
  - Outliers, Imbalance classes
  - Cluster Analysis
  - Pattern mining: association rules, sequence mining

Program II

- Advanced Topics
  - Advanced topics in Classification:
    - Novelty detection, structured output prediction
    - Semi-supervised learning,
    - AutoML
    - Explainability, Fairness, Trustability
  - Time-series Analysis
  - Predictive Maintenance
  - Natural Language Processing,
  - Text and Web Mining
  - Data stream analysis
  - Social Network Analysis
  - Big Data

Classes -draft

- Introduction to Machine Learning and Data Mining. Predictive Learning.
  1st Assessment: Kaggle Competition
- Classification Algorithms: Multiple Models
- Topics on Evaluation, ROC, PR curves. Imbalanced Domain Learning
- Advanced Topics in Classification: Semi-Supervised, Novelty Detection, Structured Output Prediction
- Text mining, Natural Language Processing
  2nd Assessment: NLP.
- Web Mining, Recommendation Systems
- Clustering
- Frequent pattern mining. Sequence mining.
- Data stream analysis, AUTOML
- Predictive maintenance
- Social network analysis
- PhD works in progress presentations.

Goals and Expected Results

At the end of the semester, students should understand Data Mining tasks, some methods and algorithms for each task, be able to apply these methods to specific data analysis problems and evaluate and criticize the results.

- Formulate a decision problem as a data mining problem;
- Identify the basic tasks in knowledge extraction from databases;
- Identify and use the main methods and algorithms for knowledge representation;
- Apply the main methods and algorithms for each mining task;
- Apply the main methods and algorithms in real-world problems and adapt to new contexts.

Teaching and Evaluation

- Theoretical and practical classes
- Evaluation
  - Home Work: Classification Kaggle Competition
  - Home Work: Text Mining

Evaluation: Kaggle Competition

Link for the Kaggle competition:
https://www.kaggle.com/t/e89d5972ff194ce2adefb8374997c7d8

- Groups of 2 students must perform the work.
- **Exploratory phase:**
  Exploratory data analysis of data: identifying outliers and anomalous examples. Use graphics tools to understand data.
- **Predictive phase:**
  You can use any tool or combination of tools for the predictions (Python, R, Excel, Weka, KNIME or RapidMiner). Each group should submit a report by 30 November 2024. Authors must upload the report as a PDF document. The report must have, at most, 12 pages.
- Criteria:
  - Critical analysis of the results!
  - Argumentation and justification of the choices made.

Teaching Material

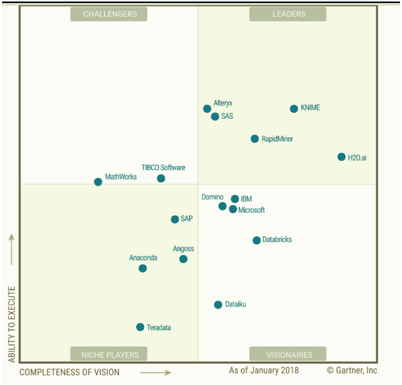
- Teaching Methods
    - Theoretical and practical classes
- Teacher’s Slides, reference papers, textbook
- Books
    - LIAAD Library

Bibliography

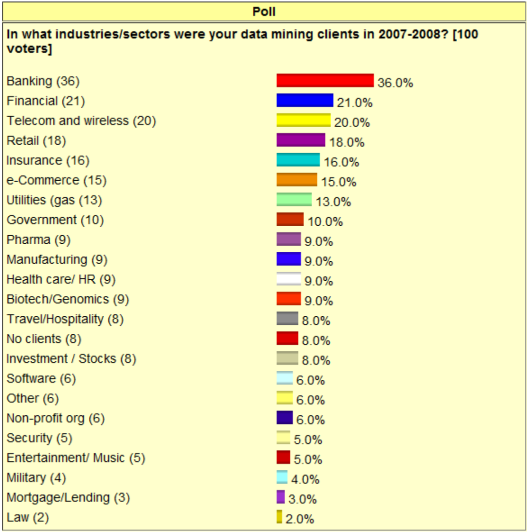
- **J.Gama, A. Carvalho, K.Faceli, A.Lorena,** *Extração de Conhecimento de Dados*,**Silabo, 2017**
- **Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006**
- Ian Witten, Eibe Frank; *Data Mining: practical machine learning tools and Techniques with java implementations*, Morgan Kaufmann, 2000
- Tom M. Mitchell, *Machine learning*, McGraw Hill, 1997
- J. Gama, *Knowledge Discovery from Data Streams*, Chapmann & Hall, 2011

Software

- R: http://www.r-project.org/
- Python: https://www.python.org/
- Weka: http://www.cs.waikato.ac.nz/ml/weka/
- Knime: http://www.knime.org/
- Rapid Miner: https://community.rapidminer.com/



Where is it used?

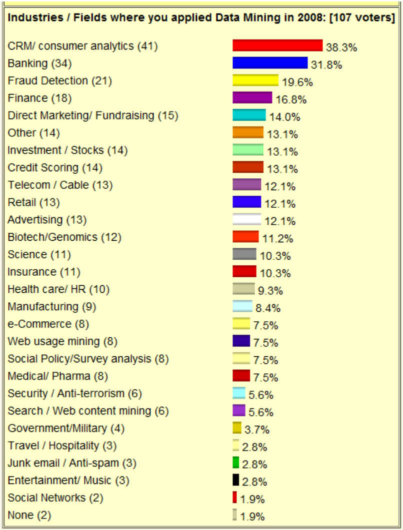


Poll

In what industries/sectors were your data mining clients in 2007-2008? [100 voters]

Banking (36)	36.0%
Financial (21)	21.0%
Telecom and wireless (20)	20.0%
Retail (18)	18.0%
Insurance (16)	16.0%
e-Commerce (15)	15.0%
Utilities (gas (13)	13.0%
Government (10)	10.0%
Pharma (9)	9.0%
Manufacturing (9)	9.0%
Health care/ HR (9)	9.0%
Biotech/Genomics (9)	9.0%
Travel/Hospitality (8)	8.0%
No clients (8)	8.0%
Investment / Stocks (8)	8.0%
Software (6)	6.0%
Other (6)	6.0%
Non-profit org (6)	6.0%
Security (5)	5.0%
Entertainment/ Music (5)	5.0%
Military (4)	4.0%
Mortgage/Lending (3)	3.0%
Law (2)	2.0%

Which tasks?


Industries / Fields where you applied Data Mining in 2008: [107 voters]
CRM/ consumer analytics (41) 38.3%
Banking (34) 31.8%
Fraud Detection (21) 19.6%
Finance (18) 16.8%
Direct Marketing/ Fundraising (15) 14.0%
Other (14) 13.1%
Investment / Stocks (14) 13.1%
Credit Scoring (14) 13.1%
Telecom / Cable (13) 12.1%
Retail (13) 12.1%
Advertising (13) 12.1%
Biotech/Genomics (12) 11.2%
Science (11) 10.3%
Insurance (11) 10.3%
Health care/ HR (10) 9.3%
Manufacturing (9) 8.4%
e-Commerce (8) 7.5%
Web usage mining (8) 7.5%
Social Policy/Survey analysis (8) 7.5%
Medical/ Pharma (8) 7.5%
Security / Anti-terrorism (6) 5.6%
Search / Web content mining (6) 5.6%
Government/Military (4) 3.7%
Travel / Hospitality (3) 2.8%
Junk email / Anti-spam (3) 2.8%
Entertainment/ Music (3) 2.8%
Social Networks (2) 1.9%
None (2) 1.9%