UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

CPGEI

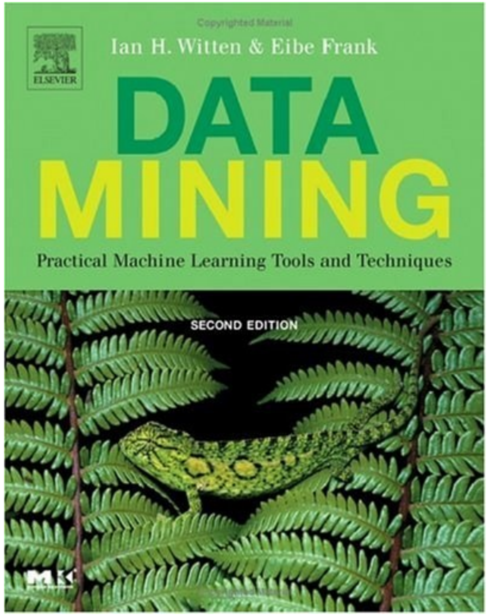
Prof. Heitor Silvério Lopes
Prof. Thiago H. Silva

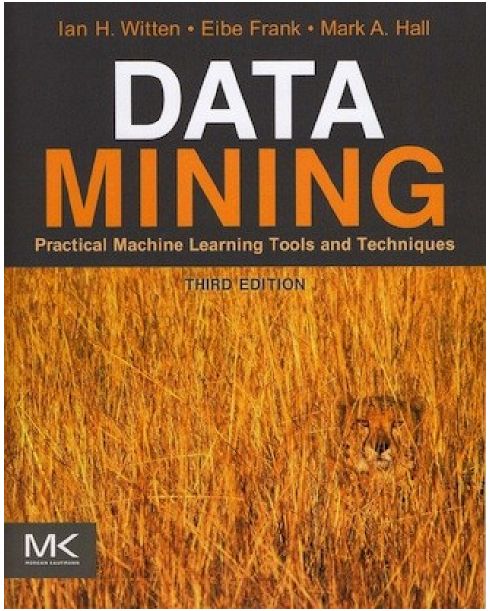
Data Mining & Knowledge Discovery

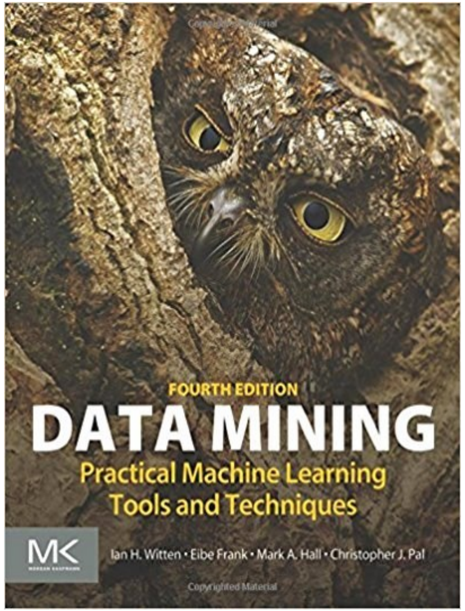
Class 1a – Introduction & Overview
2025

Data mining → Knowledge discovery

The purpose of D.M. is to find new, useful, and relevant knowledge hidden in large amounts of data


Copyrighted Material
ELSEVIER
Ian H. Witten & Eibe Frank
DATA
MINING
Practical Machine Learning Tools and Techniques
SECOND EDITION
MK


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA
MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

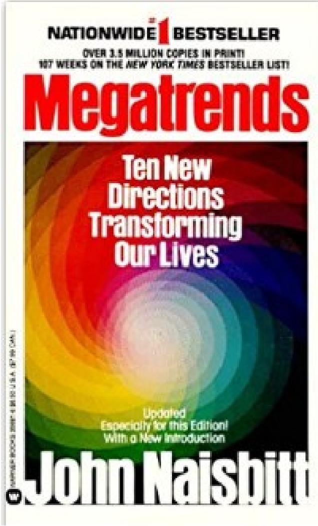

FOURTH EDITION
DATA MINING
Practical Machine Learning
Tools and Techniques
MK
Ian H. Witten • Eibe Frank • Mark A. Hall • Christopher J. Pal
Copyrighted Material

The Multidisciplinarity of Data Mining

- Data mining uses concepts and methods from many areas:
  - Machine Learning
  - Databases
  - Computational Intelligence (EC, NN, FS)
  - Mathematics / Statistics
  - Programming languages

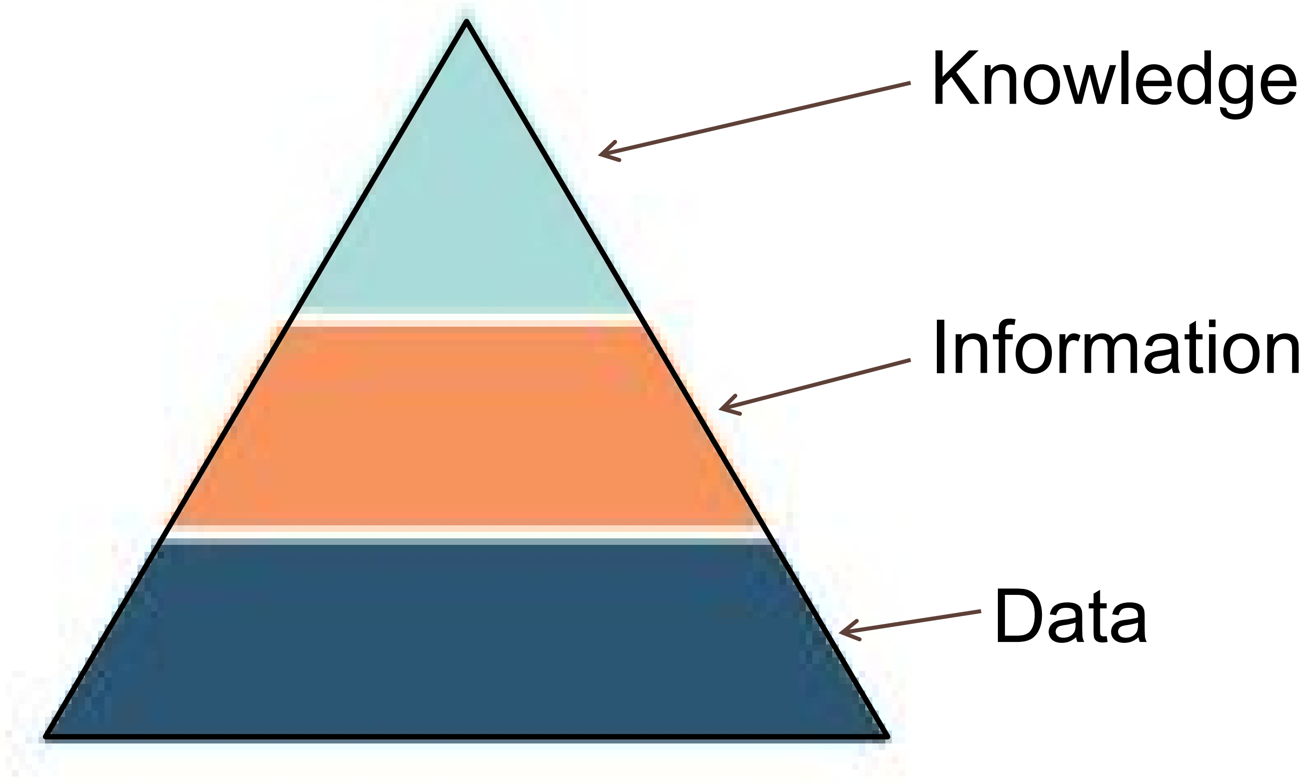
Data x Information X Knowledge

- Data:
    - Instances (objects, people, timestamps, etc)
    - Describe individual, not collective, properties, and they are:
        - Easy to collect
        - Available in large amounts and forms
        - Few useful for predictions or decision-making
- Information:
    - Classes (groups) of instances
    - Describe generic patterns, structures, principles, etc
        - Hard to obtain
        - Few abundant
        - Allow generalizations and predictions
- Knowledge
    - Regards the comprehension of something (including facts, habilities and informations)
    - Obtained by means of human perceptions or learning


NATIONWIDE #1 BESTSELLER
OVER 3.5 MILLION COPIES IN PRINT!
107 WEEKS ON THE NEW YORK TIMES BESTSELLER LIST!
Megatrends
Ten New Directions Transforming Our Lives
Updated Especially for this Edition! With a New Introduction
John Naisbitt

We are drowning in <u>information</u>,
but starving for <u>knowledge</u>.
John Naisbitt (**1982**)

Data x Information X Knowledge



Some important definitions of Data Mining

- Automatic/semi-automatic discovery of structural patterns in data (Witten et al., 2000)

- Extraction of structured knowledge which is useful, previously unknown, non-trivial, humanly comprehensible, from large amounts of data (Fayyad et al., 1996)

- Desirable features of discovered knowledge:
  - Correctness
  - Generality
  - Utility
  - Comprehensibility
  - Novelty

Examples of rules discovered using data mining

- Case 1: consider a dataset of patient records from a maternity hospital. A data-mining procedure found this rule:

IF (patient.age >) 15 **AND** (patient.age < 50) **AND** (sector = “surgical clinic”) **AND** (surgery.type = “cesarean”) **THEN** (patient.sex = “female”)

Correctness ☺
Generality ☺
Utility ☹
Comprehensibility ☺
Novelty ☹

- Case 2: consider a dataset of pediatric oncological medical records*. A data-mining procedure found this rule:

IF (histology.type = carcinoma) **AND** (patient.age < 3) **AND** (oncological.stage = 1) **AND** (metastasis=“no”) **THEN** (years.survival > 5)

Correctness ☺
Generality ☺
Utility ☺ ☺
Comprehensibility ☺
Novelty ☺ ☺ ☺

* Bojarczuk, C.C., Lopes, H.S., Freitas, A.A. A constrained-syntax genetic programming system for discovering classification rules: application to medical data sets. ***Artificial Intelligence in Medicine***, v. 30, n. 1, p. 27-48, 2004.

Life-cycle of Data Mining projects

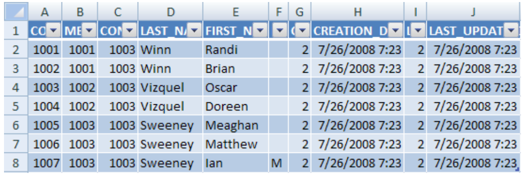
Hard work !

Raw data

Collection, selection, data integration

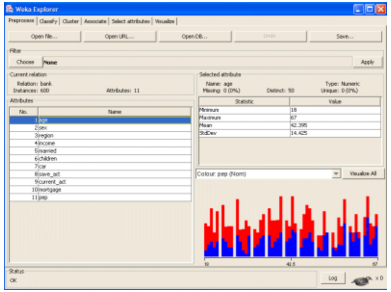
Data warehouse

Pre-processing: formatting, cleaning, data reduction



	A	B	C	D	E	F	G	H	I	J
1	CO	ME	CON	LAST_N/	FIRST_N		(	CREATION_D	L	LAST_UPDAT
2	1001	1001	1003	Winn	Randi		2	7/26/2008 7:23	2	7/26/2008 7:23
3	1002	1001	1003	Winn	Brian		2	7/26/2008 7:23	2	7/26/2008 7:23
4	1003	1002	1003	Vizquel	Oscar		2	7/26/2008 7:23	2	7/26/2008 7:23
5	1004	1002	1003	Vizquel	Doreen		2	7/26/2008 7:23	2	7/26/2008 7:23
6	1005	1003	1003	Sweeney	Meaghan		2	7/26/2008 7:23	2	7/26/2008 7:23
7	1006	1003	1003	Sweeney	Matthew		2	7/26/2008 7:23	2	7/26/2008 7:23
8	1007	1003	1003	Sweeney	Ian	M	2	7/26/2008 7:23	2	7/26/2008 7:23

Filtered/ cleaned data



Filter: None

Current relation

Relation: bank

Instances: 600 Attributes: 11

Selected attribute

Name: age Type: Numeric

Missing: 0 (0%) Distinct: 50 Unique: 0 (0%)

Statistic	Value
Minimum	18
Maximum	67
Mean	42.395
StdDev	14.425

Attributes

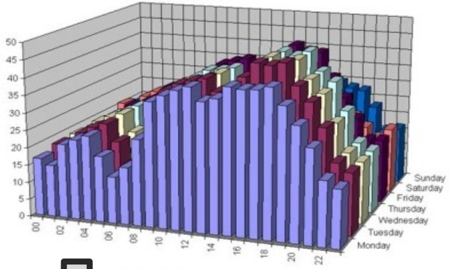
No.	Name
1	age
2	sex
3	region
4	income
5	married
6	children
7	car
8	save_act
9	current_act
10	mortgage
11	pep

Colour: pep (Nom)

Status

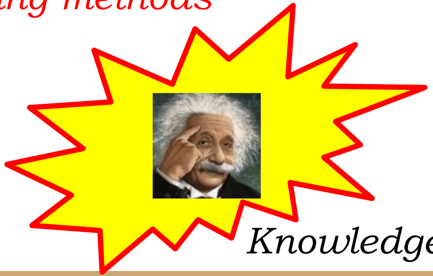
OK

Data mining methods



Pattern discovery

Pattern analysis and interpretation

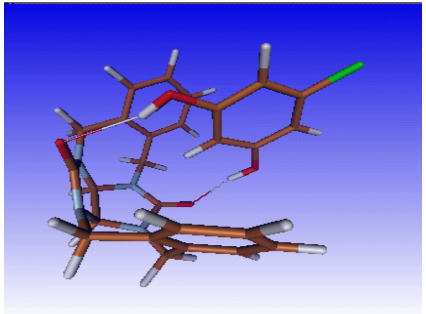


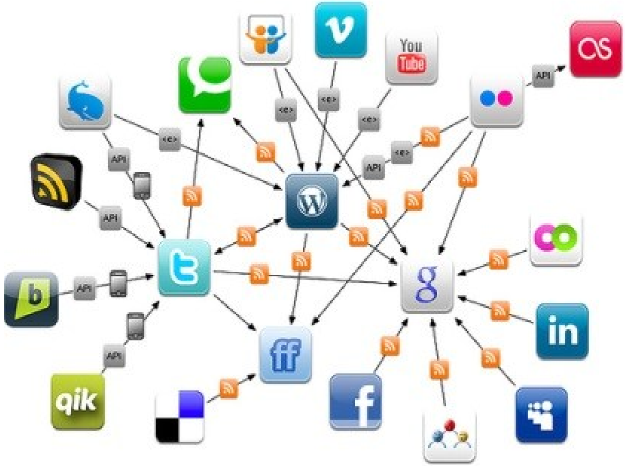
Knowledge !!

Motivations for Data Mining

1) VERY LARGE amount of data freely available in the internet

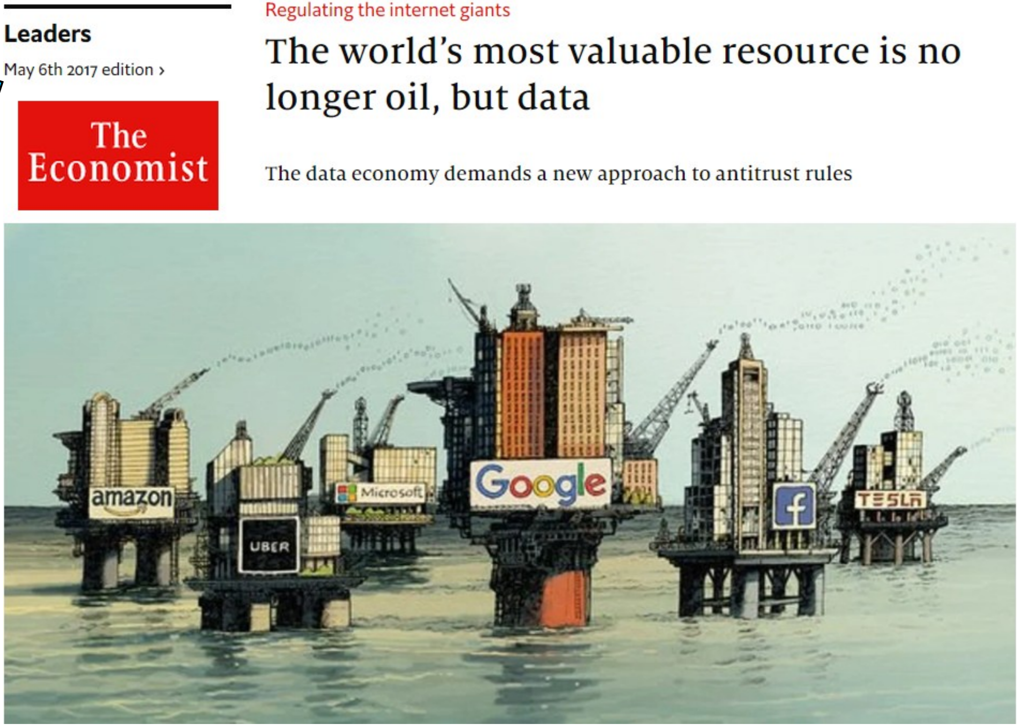
- E-mails and social networks
- Business and bank transactions
- Web page searches (Webscrapping!)
- Medical and biological data
- Scientific and astronomical data





Motivations for Data Mining

2) Business/commercial interest (**$$$**)



Leaders

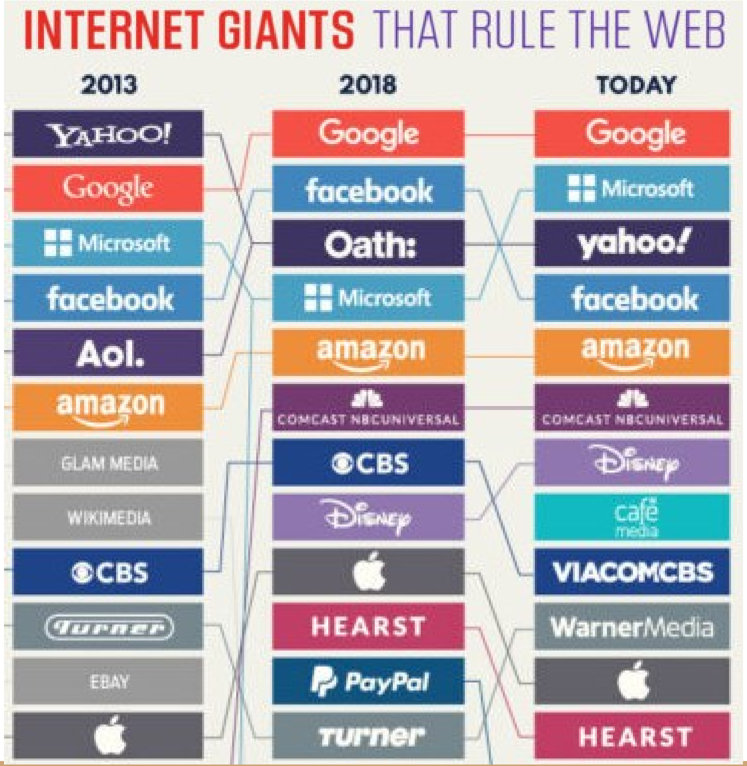
May 6th 2017 edition ›

The Economist

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules



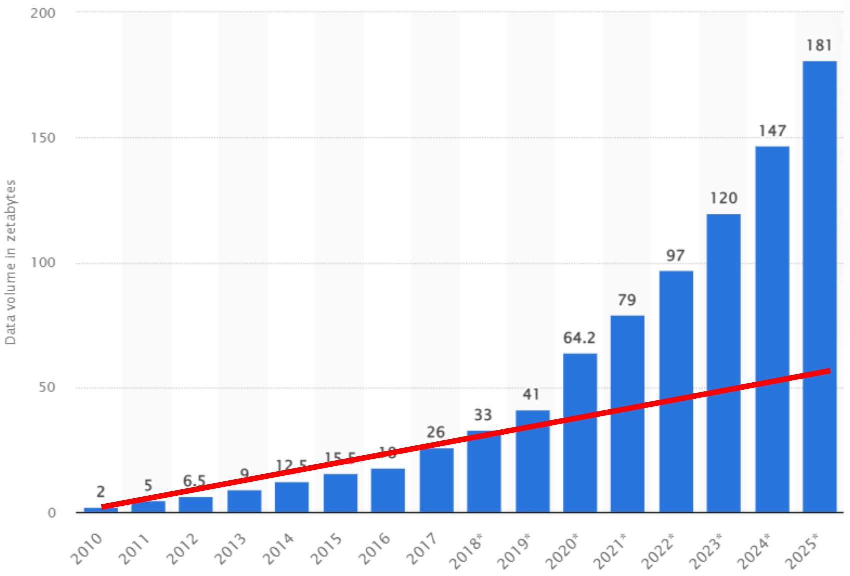
INTERNET GIANTS THAT RULE THE WEB

2013	2018	TODAY
YAHOO!	Google	Google
Google	facebook	Microsoft
Microsoft	Oath:	yahoo!
facebook	Microsoft	facebook
Aol.	amazon	amazon
amazon	COMCAST NBCUNIVERSAL	COMCAST NBCUNIVERSAL
GLAM MEDIA	CBS	Disney
WIKIMEDIA	Disney	café media
CBS		VIACOMCBS
Turner	HEARST	WarnerMedia
EBAY	PayPal	
	turner	HEARST

Critical Dilema in Data Mining

- The amount of data generated, created, stored, etc, grows *exponentially*
- The ability to mine, understand, and effectively use these data grows *linearly* (best case!)

- Data mining may help us to understand large amounts of data by extracting useful knowledge



* https://explodingtopics.com/blog/data-generated-per-day

Tasks x Methods in Data Mining

Tasks	Methods
Classification	Decision trees (C4.5), Cassification rules, k-nearest-neighboors, Random forest, Support vector machine, Bayesian classifier, Neural network, Adaboost
Association Rules	Apriori, FP-growth, Eclat, Zigzag
Regression	Linear Regression, Polynomial regression, Logistic regression
Feature Selection & Dimensionality Reduction	Principal component analysis (PCA), Chi-square, Entropy, Information gain
Clustering	K-means, Kohonen's self-organized map, Density-based scan, Hierarchical grouping, t-SNE
Data visualization *	Silhouette plot, scatter plot, heatmap, box plot, clusters, t-SNE

Tasks x Methods in Data Mining

- Types of data:
  - Numerical
  - Categorical
  - Text
  - Image/video
  - Time-series/signals

- Some data types require diferent tasks, for instance:
  - Image, time-series/signals can be clustered or classified
  - Text can be classified, but may require other specific tasks (e.g. sentiment analysis)

Some open-source softwares for Data Mining

- Orange (Python): developed and maintained by the University of Ljubljana (SL) https://orangedatamining.com/
  - Easy-to-use windows interface (visual programming), add-ons for specific tasks, allows integration with Python code.

- Weka (Java): created and maintained by the Waikato University (NZ) https://www.cs.waikato.ac.nz/ml/weka
  - Very large library of methods, community support
  - Not-so-user-friendly interface, Poor documentation

- Knime (Java): developed and maintained by the Konztanz Universitaet (GE) https://www.knime.com/

- Further information: https://www.datamation.com/big-data/open-source-data-mining-tools/