Tomáš Horváth

INTRODUCTION TO DATA SCIENCE

Lecture 1

Introduction

Data Science and Engineering Department
Faculty of Informatics
ELTE University

The plan...

Lectures

1. Introduction
2. Clustering
3. Frequent Pattern Mining
4. Supervised Learning
5. Data pre-processing
6. Factorization techniques
7. Data types
   - Time-series
   - Text
   - Image
   - Social Networks
8. Factorization Techniques

Practicals

- Python
- tutored by my PhD students

Semestral project

- in last 2-3 weeks
- toy data
- research paper

The goal is

- basic concepts
- soft skills
  - presenting
  - scientific writing

The hard part...

Oral (on-line) **Exam** in front of a committee

- Theory and Practicals (**50 points**)
  - answering any question from lectures or practicals
  - no time for preparation (to avoid cheating)
- Semestral project (**50 points**)
  - presenting the scientific paper
  - answering questions

Final grading based upon points received

- $\langle 0, 60 \rangle$ = grade **1**
- $\langle 61, 70 \rangle$ = grade **2**
- $\langle 71, 80 \rangle$ = grade **3**
- $\langle 81, 90 \rangle$ = grade **4**
- $\langle 91, 100 \rangle$ = grade **5**

Good to know...

- **Consultations**, in case you need to talk
  - to me: Tuesdays from 9am to 12am
  - to my colleagues: based on arrangement with them
- Bureaucracy and administration
  - Every issue to be addressed to Mr. Ádám Horváth (secretary of our) department
- Adult behavior is expected
  - "I didn't know", "I couldn't find", "I had tests from other courses", ...
  - or, an email, 12 hours before the exam, asking what you should learn
- The "somehow I'll make it" isn't really working, thus, please, study!
  - Last years' statistics:
    - 96 students
    - 7 with grade 5, 13 with grade 4, 10 with grade 3, 12 with grade 2
    - 54 students failed :(

Brief Introduction

STEVE DEBENPORT VIA GETTY IMAGES

Why Data Science?

Mainly, because

- there are **(big) data** outside
- **intelligent systems** are demanded
- smart **decision support** is needed
- **data-intensive science**
  - T. Hey, S. Tansley and K. Tolle. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009.

and, also

- "data miner is the most sexiest job in the 21th Century" (HBR)

Relations between buzzwords

Data Science vs. **Machine Learning**

- ML, focused mostly to prediction is a part of DS
- DS covers also other areas such that data visualization

DS vs. **Artificial Intelligence**

- AI is often, but not correctly, associated with ML
- AI covers also other areas such that computer vision or planning

DS vs. **Business Analytics** and **Business Intelligence**

- BA is concerned with decision support based on extensive use of data analysis, thus DS is a methodology used in BA
- BI is reporting what was happened or where the problem is while BA is looking for why is this happening and what will happen next

DS vs. **Deep Learning** and **Big Data**

- DL, related to neural networks, is part of ML
- BD deals with techniques such that parallelization or distribution to efficiently process big data (a magic word managers like to say)

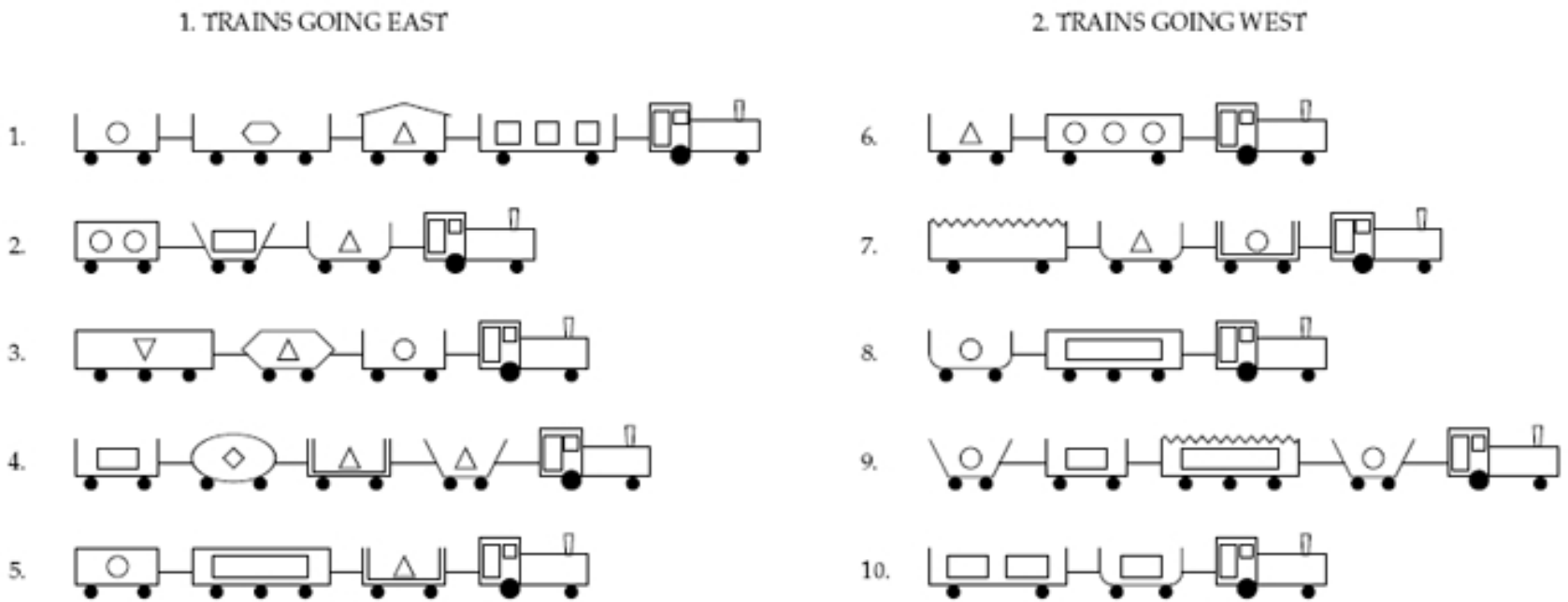
Big Data

Yes, that one with those **V** characteristics

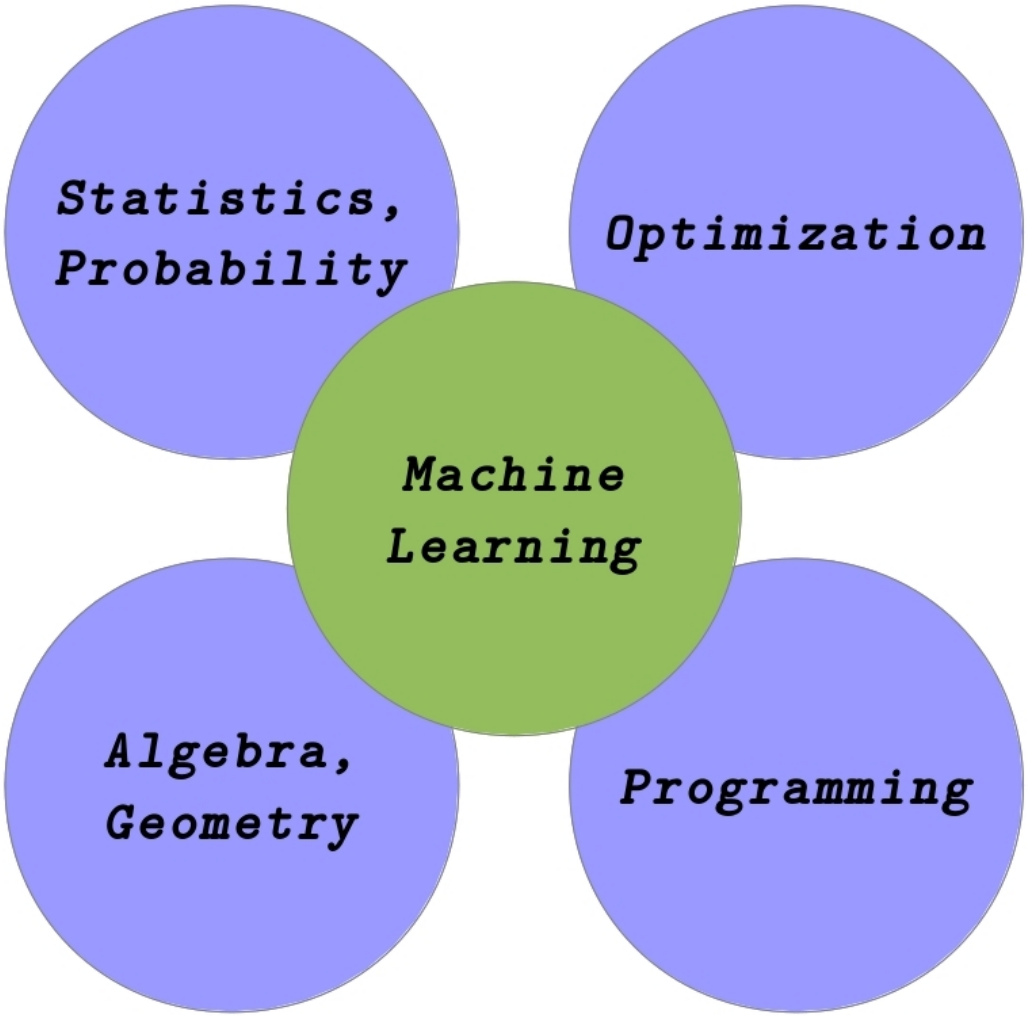
- Volume, Velocity, Variety, ...

But even in case of small data it is not so easy

- for example, how trains going East differ from those going West?
  - How would you represent such data within a computer program?



What skills is good to have?



Some Applications

Handwritten digit recognition

Image source: Subhransu Maji and Jitendra Malik: Fast and Accurate Digit Classification. Technical Report No. UCB/EECS-2009-159, Berkeley, 2009.

Some Applications

Spam filtering

Image source: Royce's spam collection, `http://xrl.us/rspam`

Some Applications

Robotics

Image source: `http://asimo.honda.com/`

Some Applications

fMRI (functional magnetic resonance imaging)

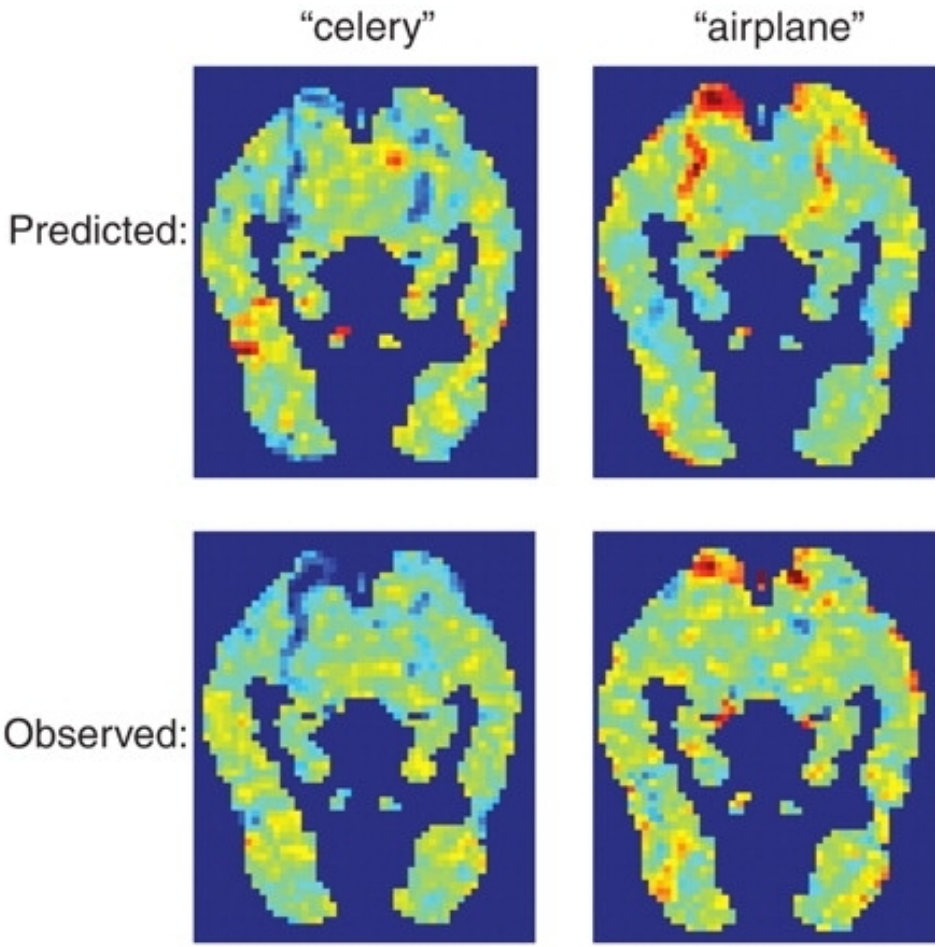
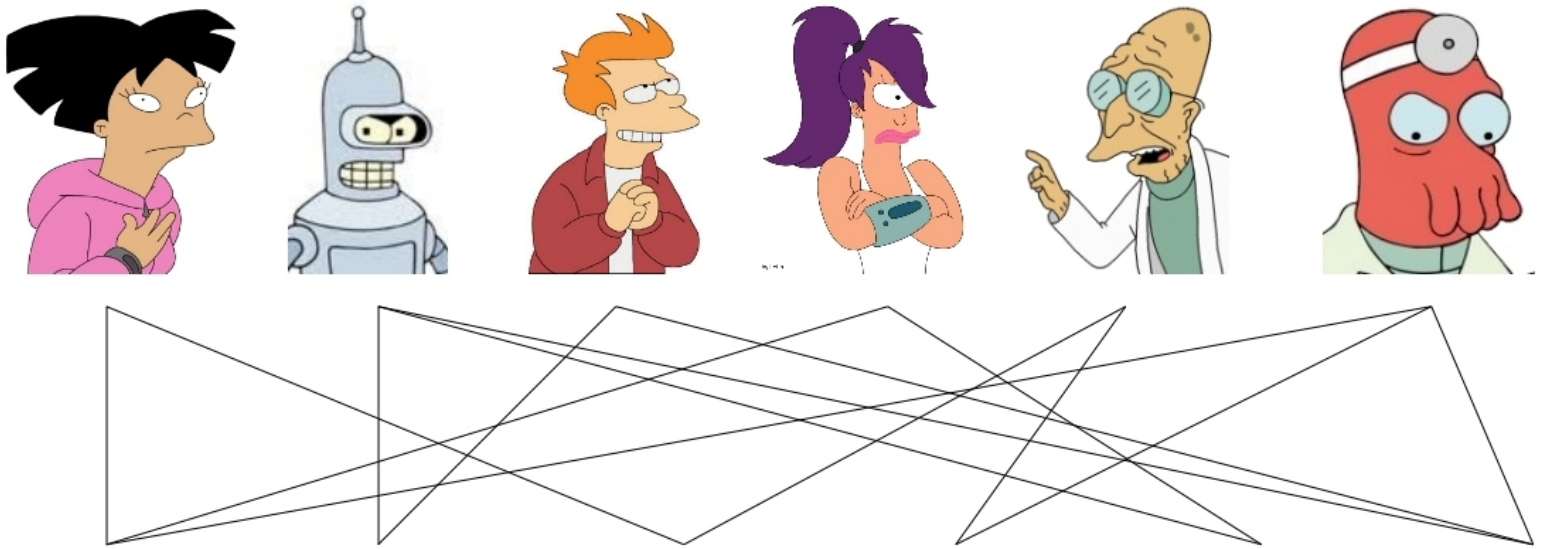


Image source: Tom M. Mitchell, et al. Predicting Human Brain Activity Associated with the Meanings of Nouns. Science 320, 1191 (2008).

Some Applications

Recommender systems

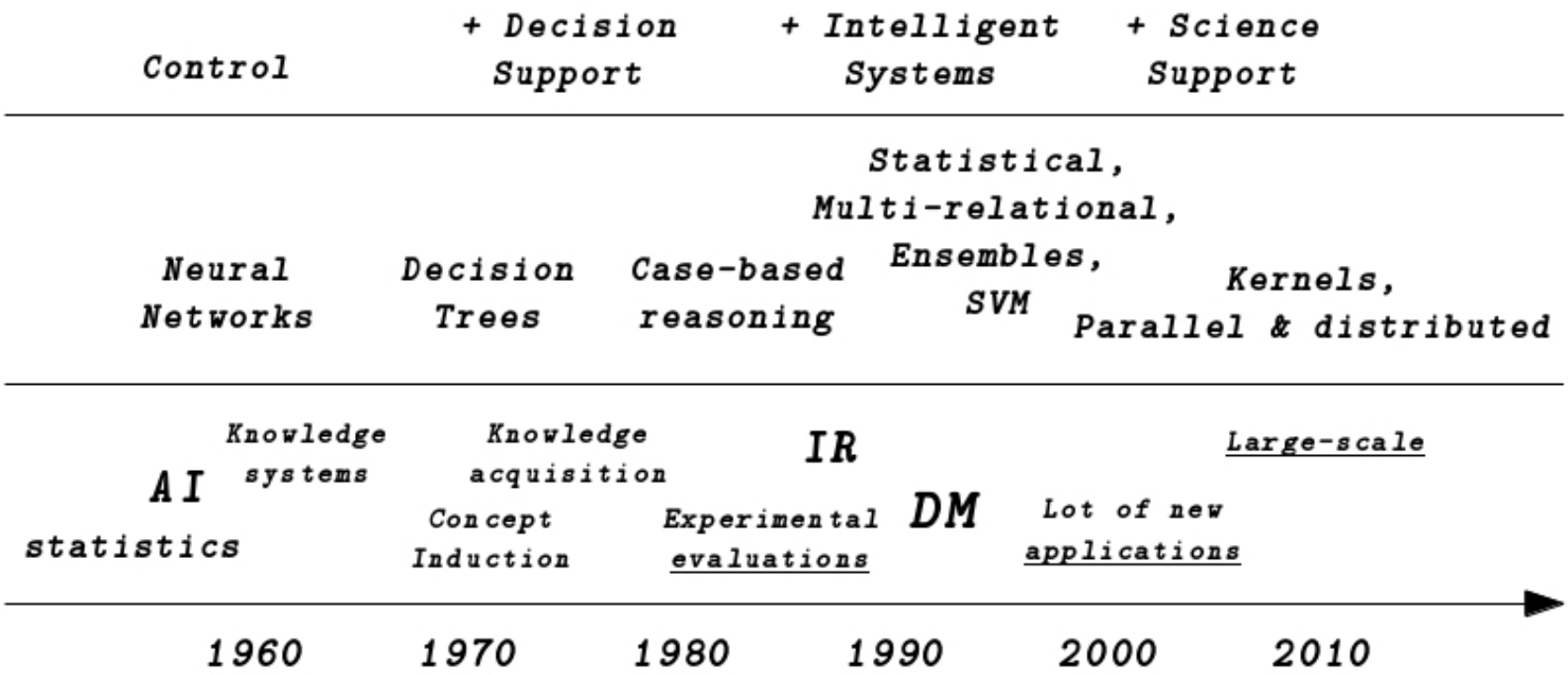


Some Applications

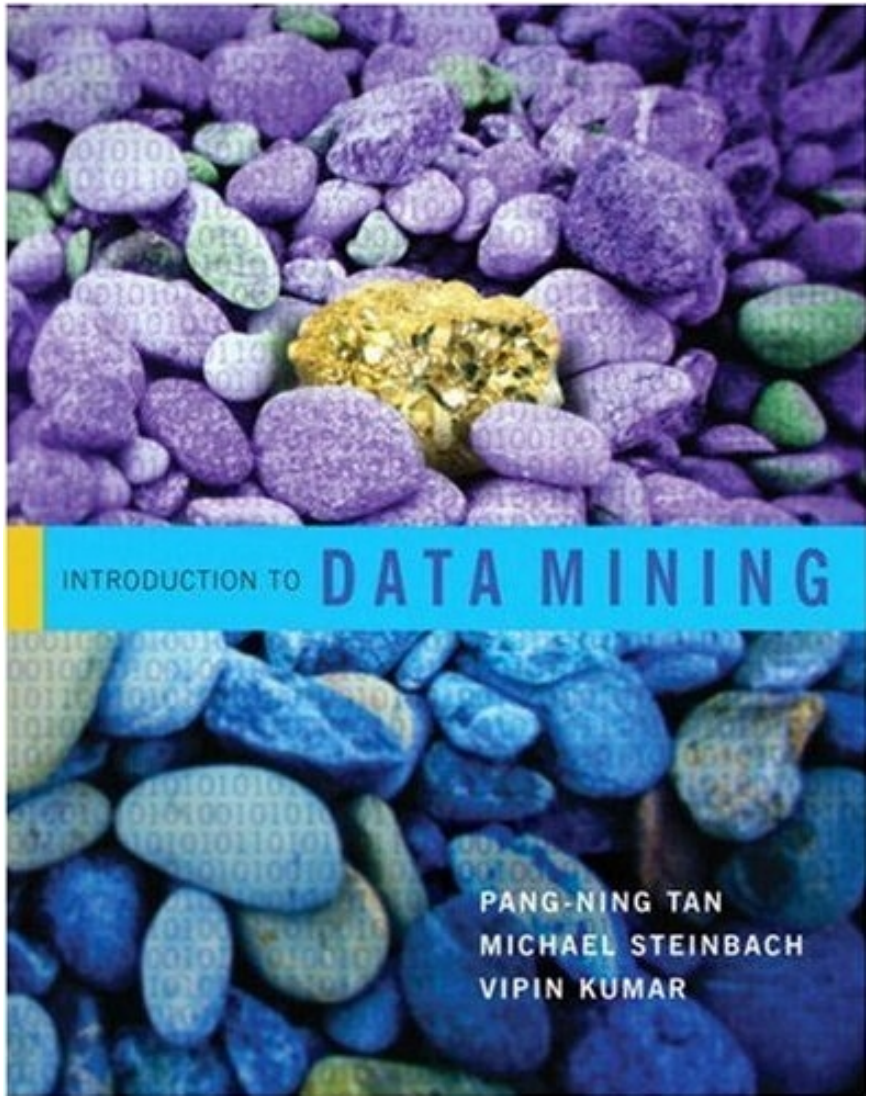
Fraud detection

Image source: `http://bdemarest.wordpress.com/`

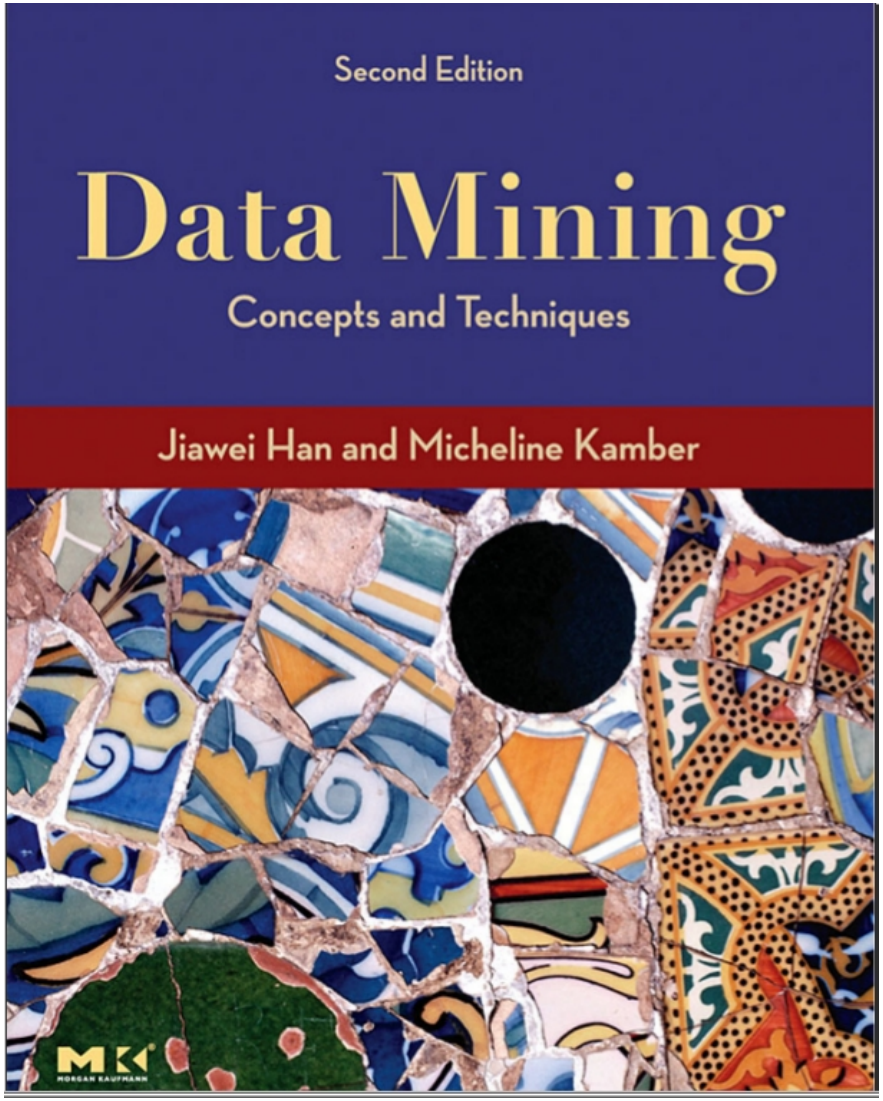
A Brief History



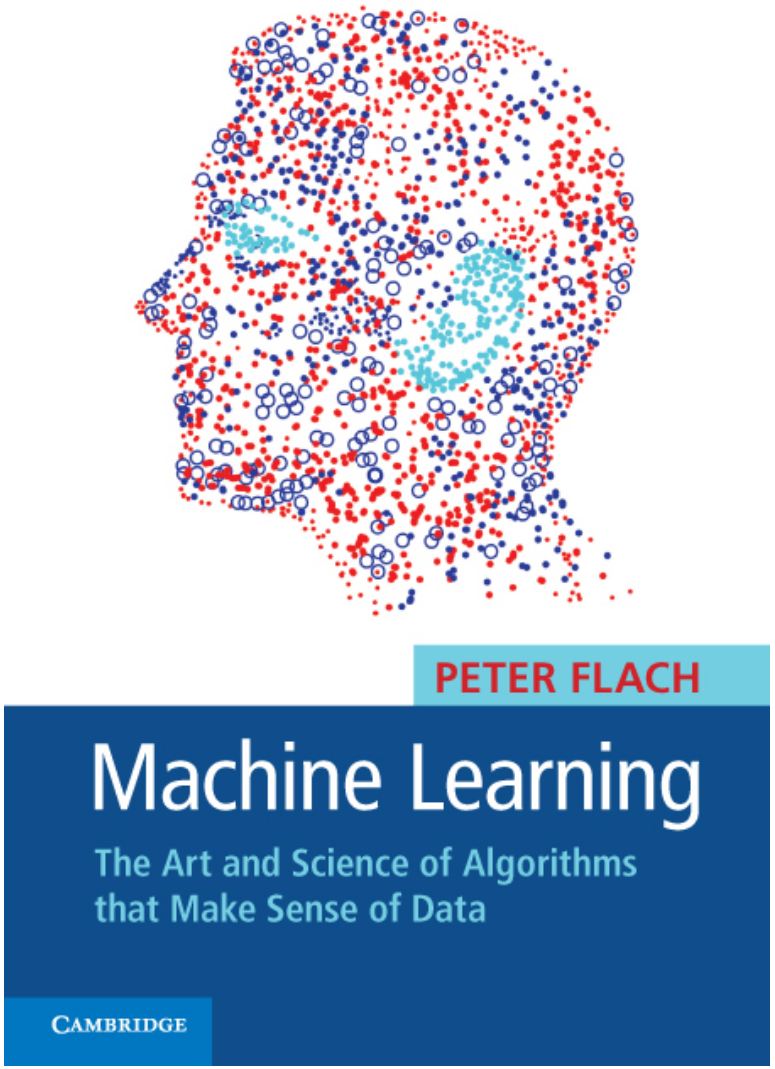
Textbook (1)


INTRODUCTION TO DATA MINING
PANG-NING TAN
MICHAEL STEINBACH
VIPIN KUMAR

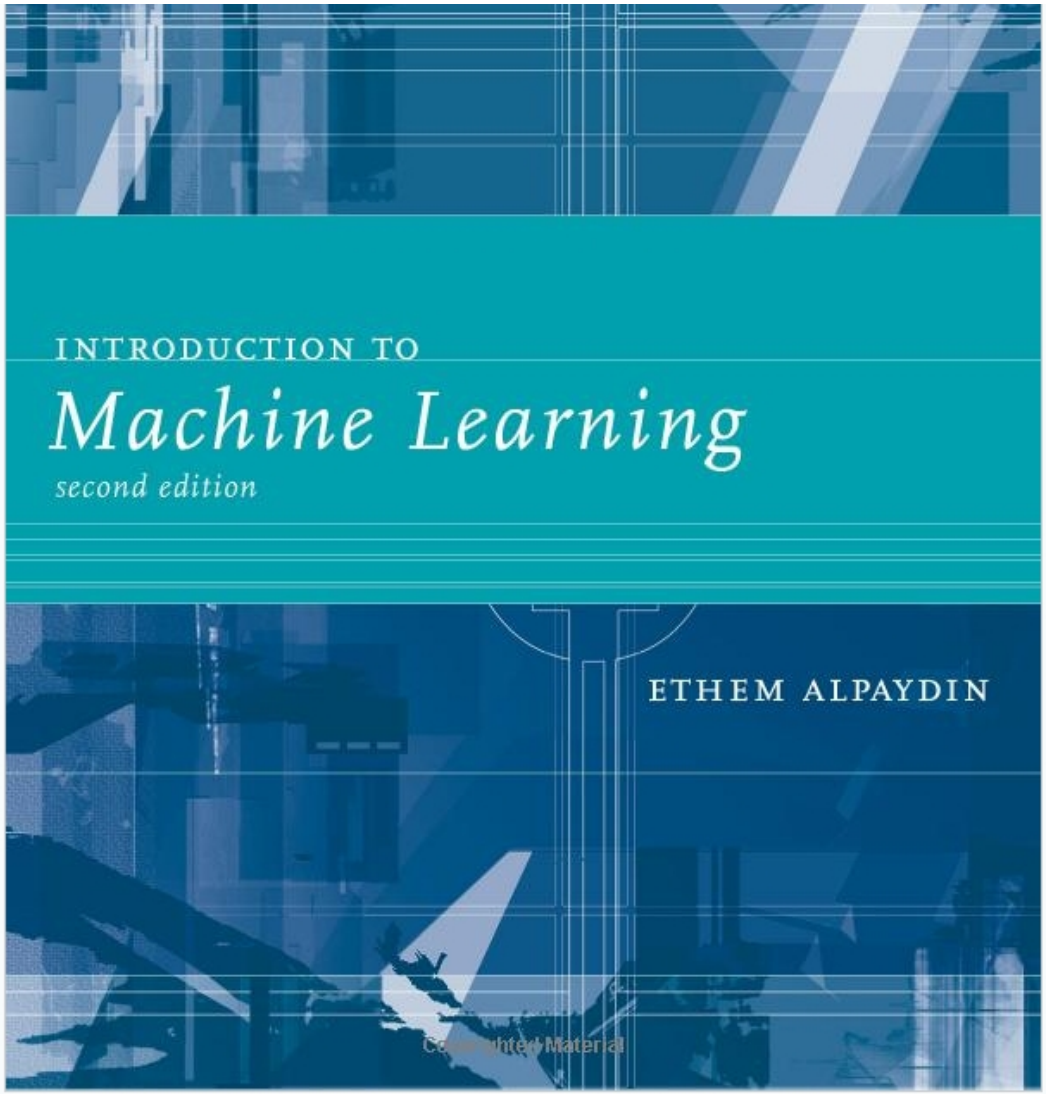
Textbook (2)


Second Edition
Data Mining
Concepts and Techniques
Jiawei Han and Micheline Kamber
MK
MORGAN KAUFMANN

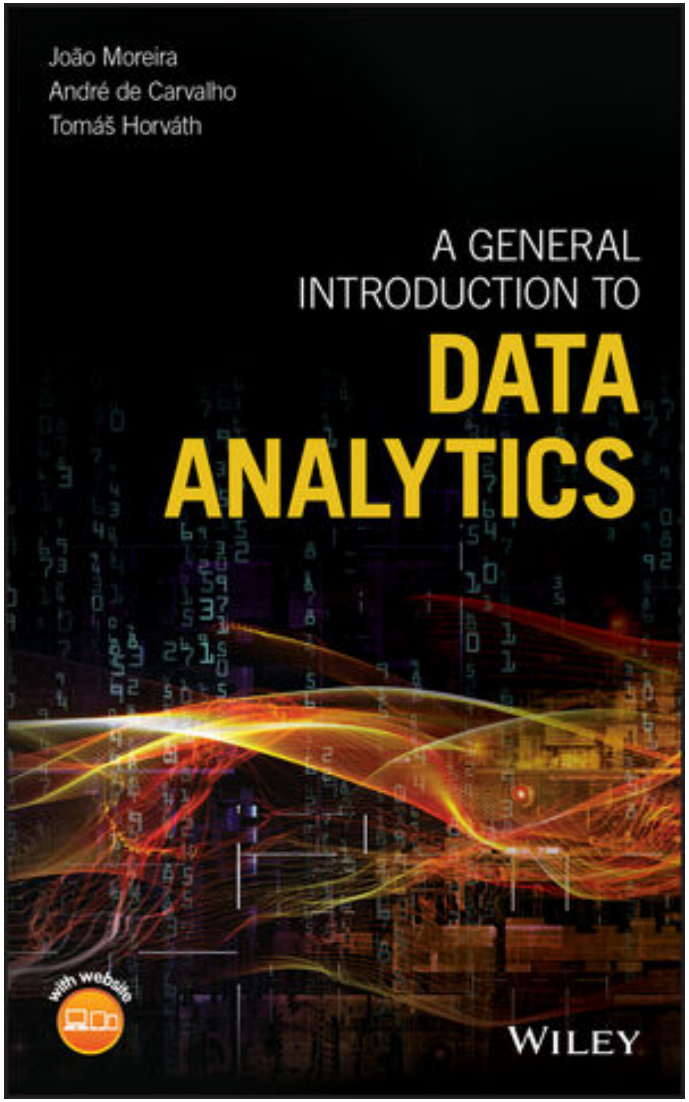
Textbook (3)


PETER FLACH
Machine Learning
The Art and Science of Algorithms
that Make Sense of Data
CAMBRIDGE

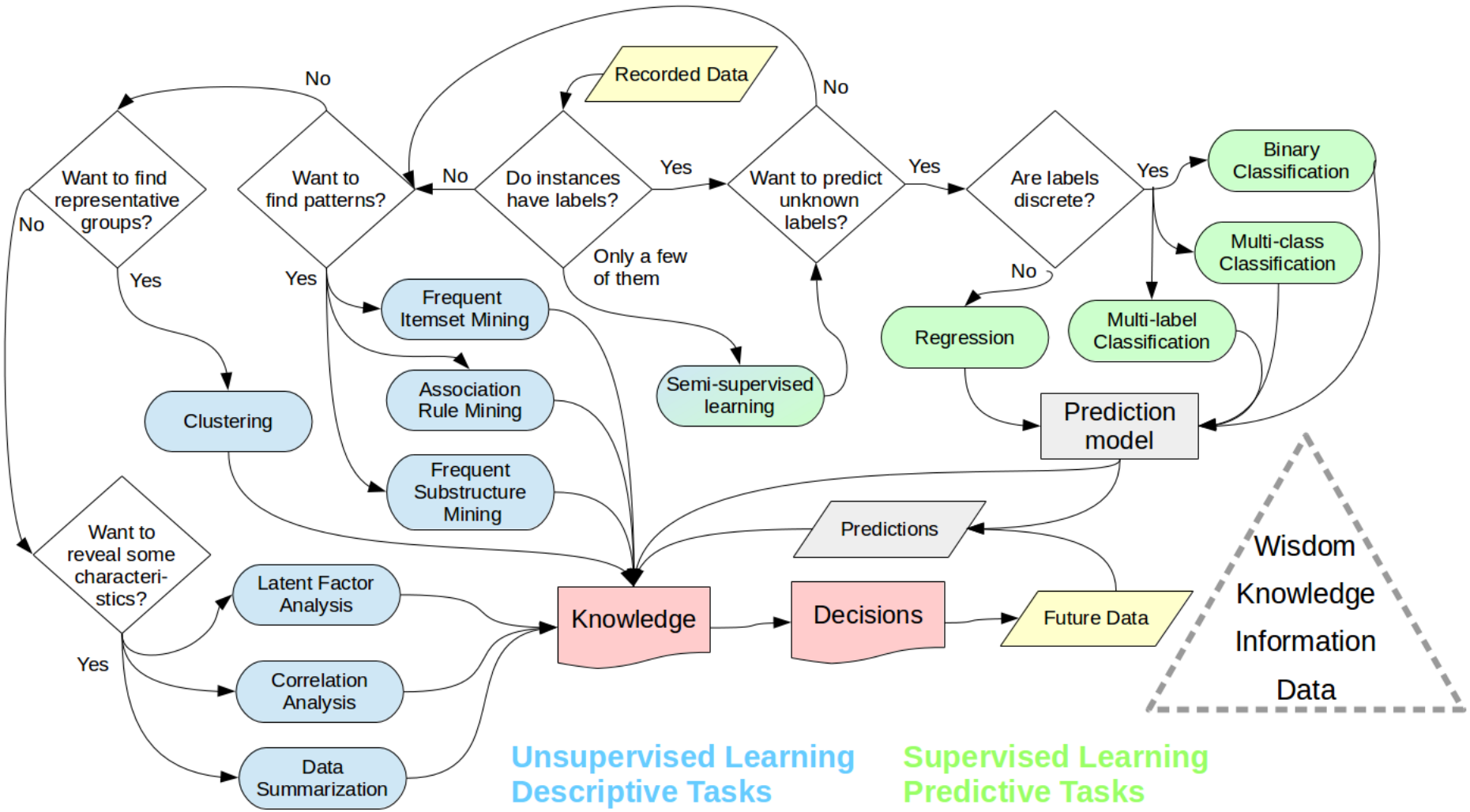
Textbook (4)


INTRODUCTION TO
Machine Learning
second edition
ETHEM ALPAYDIN
Copyrighted Material

Textbook (5)

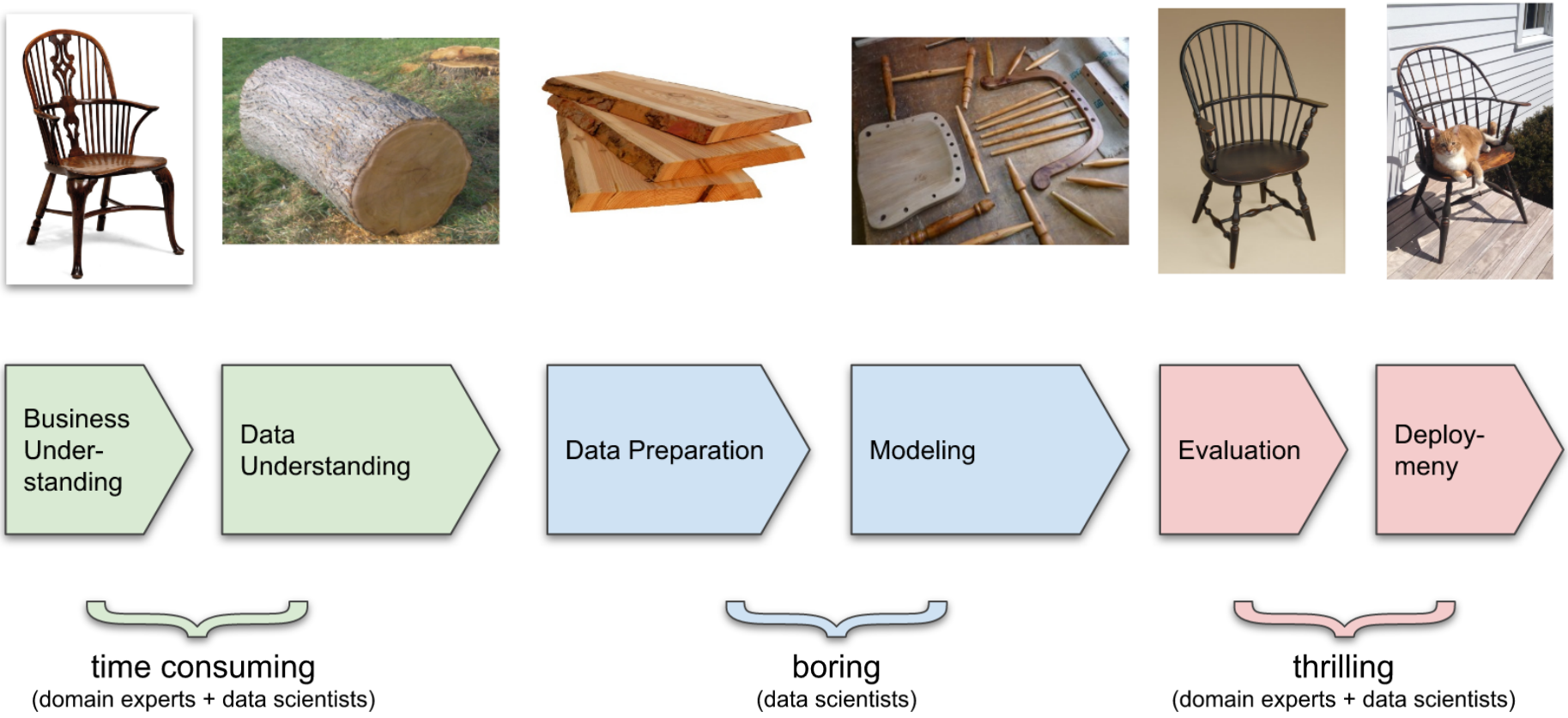

João Moreira
André de Carvalho
Tomáš Horváth
A GENERAL
INTRODUCTION TO
DATA
ANALYTICS
with website
WILEY

A rough overall picture



Various combinations of these approaches. . .

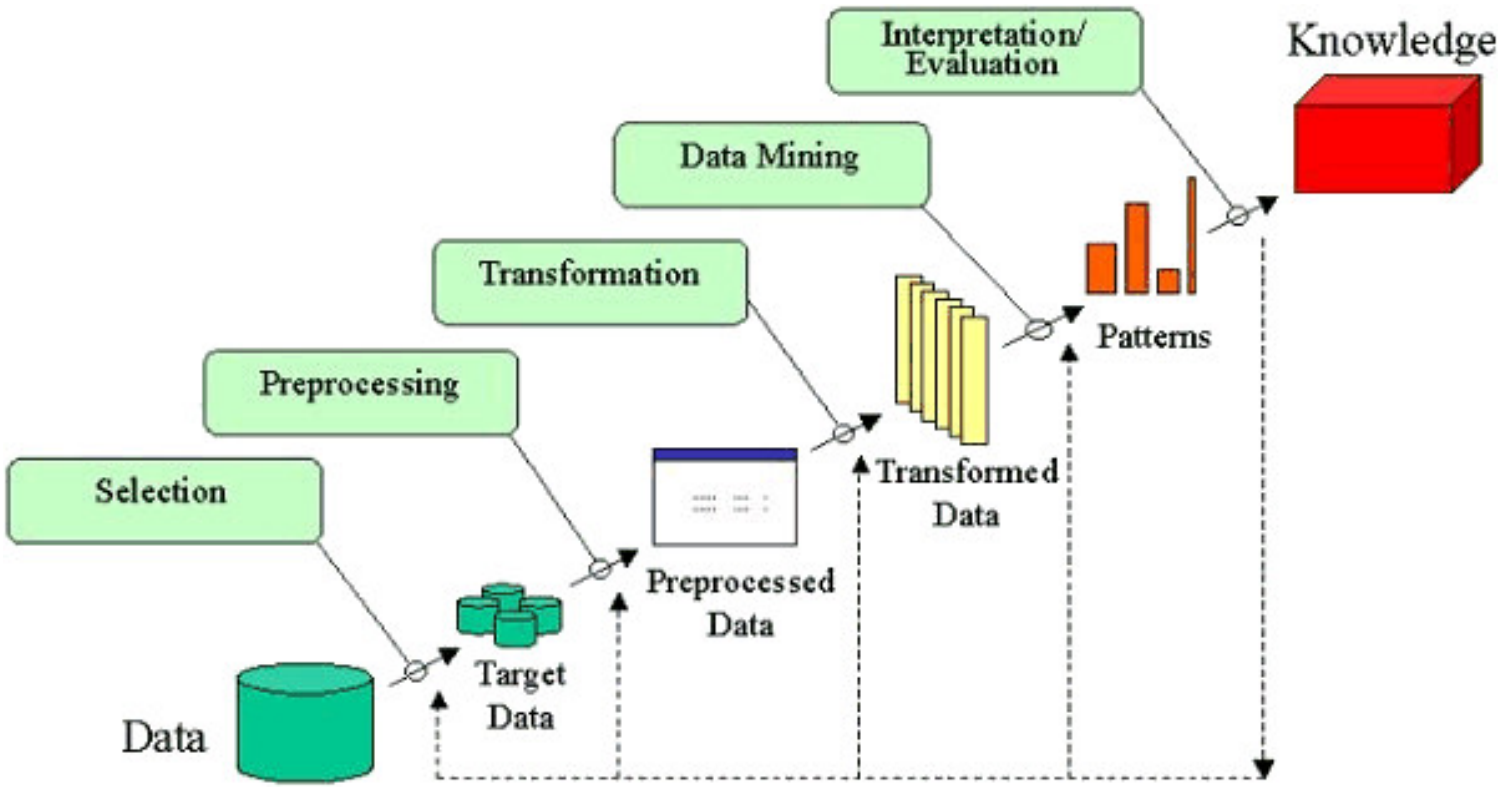
How to approach a DM task?



CRISP-DM methodology

- sometimes need to return to a previous step
  - e.g. $DU \rightarrow BU$, $M \rightarrow DP$, $E \rightarrow BU$
- continuation, improvement

A data mining project



About the CRISP-DM

A methodology developed in the project[1] (number 24.959), partially funded by the European Commission under the ESPRIT Program.

Project partners

- NCR Systems Engineering Copenhagen[2], USA and Denmark.
  - Data warehouse
- SPSS Inc.[3], USA.
  - Data mining solutions.
- DaimlerChrysler AG[4], Germany.
  - car industry
- OHRA Verzekering en Bankk Groep B.V.[5], Netherlands
  - insurance industry

[1] http://www.crisp-dm.org
[2] http://www.ncr.com
[3] http://www.spss.com
[4] http://www.daimlerchrysler.com
[5] http://www.ohra.nl

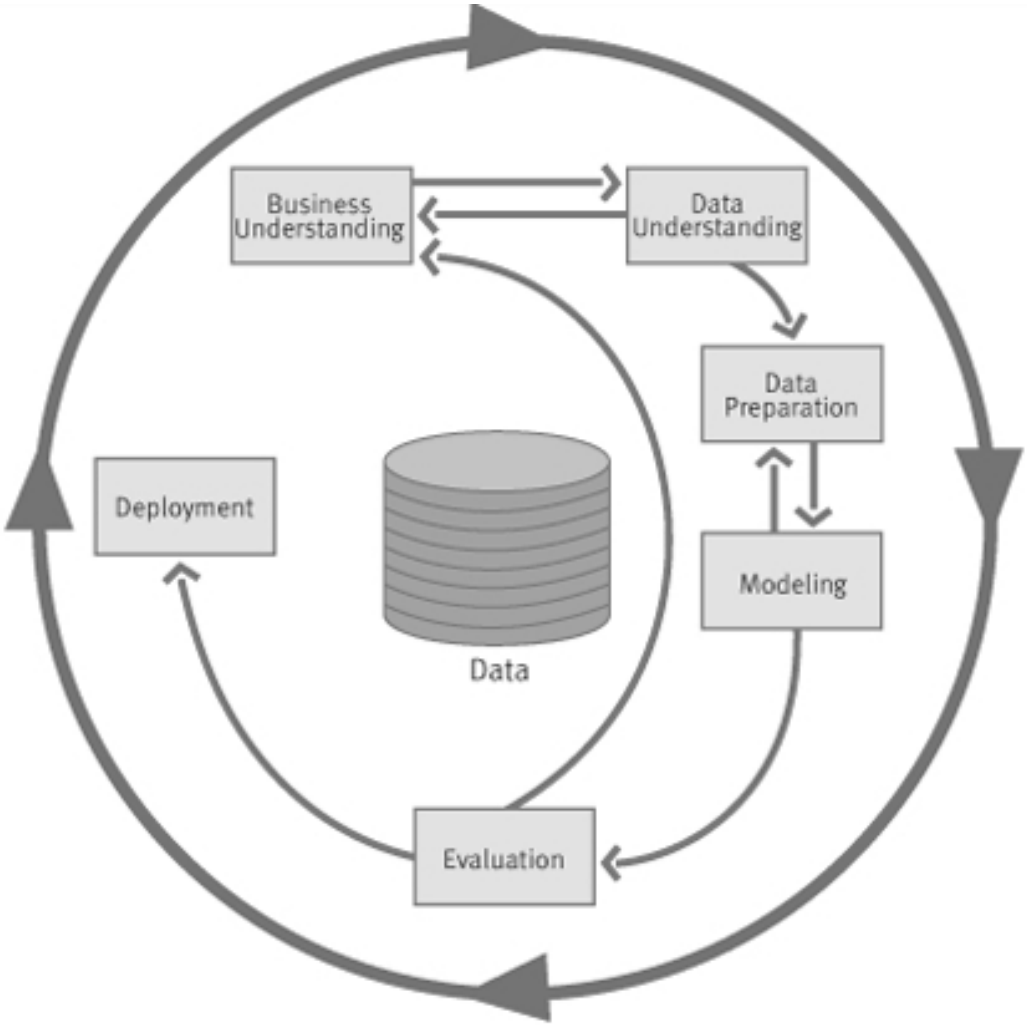
What is CRISP-DM?

CRoss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

Four levels of abstraction

- Phase
  - The data mining process is organized into several phases consisting of **tasks**.
- Generic task
  - The general level for tasks which should be **complete** (covering the whole data mining process as all possible applications) and **stable** (valid for yet unforeseen techniques).
- Specialized task
  - Description of tasks in certain specific situations, how they will be provided, etc.
- Process Instance
  - The record of actions. decisions and results of the actual data mining engagement.

The Process model



[1]

[1] Image source: http://www.crisp-dm.org

I. Business Understanding

The aim is to understand the needs of a client, the requirements and business objectives, convert the objectives to data mining goals, uncover important factors influencing these outcomes and prepare a preliminary plan for achieving the goals.

Generic tasks of this phase are

1. Determine business objectives
   - understanding the client's needs from the business perspective[1]
2. Assess situation
   - investigation of facts about the factors influencing the project
3. Determine data mining goals
   - determining the project objectives in technical terms[2]
4. Produce project plan
   - preparation of a detailed plan to reach the project objectives

[1] e.g. "Increase catalog sales to our customers."

[2] "Predict how many things customers will buy given information collected about them."

II. Data Understanding

The aim is to collect initial data, get familiar with data and identify the quality of data as well as detect subsets interesting to form some hypotheses.

Generic tasks of this phase are

1. Collect initial data
   - acquisition of data listed in resources and understanding them as well as initial data preparatin steps
2. Describe data
   - examination of the surface properties of acquired data
3. Explore data
   - querying, visualization and reporting data directly addressing the data minng goals
4. Verify data quality
   - examination of the quality of data

III. Data Preparation

The aim is to construct the final dataset from raw data which will be the input for the modeling tool.

Generic tasks of this phase are

1. Select data
   - decision on the data used for analysis according to their relevance to the specified objectives
2. Clean data
   - improve the quality of data as the selected analysis techniques require
3. Construct data
   - perform constructive data preparation operations
4. Integrate data
   - integrate data from multiple tables
5. Format data
   - mainly syntactic modifications of data to be suitable for the modeling tools

IV. Modeling

In this phase, modeling techniques are selected and their parameters are tuned to optimal values.

Generic tasks of this phase are

1. Select modeling technique
   - selection of the actual modeling technique
2. Generate test design
   - generation of a procedure to validate the model and test it's quality
3. Build model
   - run the modeling technique to build models
4. Assess model
   - interpretation, evaluation, comparison and ranking of models according to the evaluation criteria from a data mining perspective

V. Evaluation

In this phase, the model is thoroughly evaluated to be certain that it achieves the business objectives, the whole process is reviewed and next steps are determined.

Generic tasks of this phase are

1. Evaluate results
   - evaluation of the achievments of business objectives
2. Review of process
   - summarization of the whole process and detecting important factors which could be overlooked
3. Determine next steps
   - decision of the next steps to be made

VI. Deployment

In this phase, the knowledge gained during the process is organized, eventually, presented for the customers.

Generic tasks of this phase are

1. Plan deployment
   - creation of the strategy for deployment of the project results into the business
2. Plan monitoring and maintenance
   - preparation of the maintenance strategy
3. Produce final report
   - final documentation of the project
4. Review project
   - experience documentation

Basic Concepts

$$\frac{1}{F(p;\xi)} = 1+\frac{\alpha}{2\pi^2 p^2}\int \frac{d^3k}{q^2}\frac{F(k;\xi)}{k^2+\mathcal{M}^2(k;\xi)}\Bigg[\mathcal{G}(q)\Bigg\{\frac{a(k,p)}{2q^2}\left(-q^4+(k^2-p^2)^2\right)-$$

$$\left[\frac{1}{F(k;\xi)}-\frac{1}{F(p;\xi)}\right]\frac{\Omega(k,p)}{2}(k^2+p^2-q^2)-\left[\frac{b(k,p)(k^2+p^2)-c(k,p)\mathcal{M}(k;\xi)}{2q^2}\right]$$

$$\left(-q^4+2q^2(k^2+p^2)-(k^2-p^2)^2\right)\Bigg\}+\xi\Bigg\{\frac{a(k,p)}{2q^2}(q^2(k^2+p^2)-(k^2-p^2)^2)-\ b(k,p)$$

$$\frac{(k^2-p^2)^2}{2q^2}(k^2+p^2-q^2)+\frac{c(k,p)}{2q^2}\mathcal{M}(k;\xi)\left((k^2-p^2)^2-q^2(k^2-p^2)\right)\Bigg\}\Bigg]\ ,\ \frac{\mathcal{M}(p;\xi)}{F(p;\xi)}$$

$$=\frac{\alpha}{2\pi^2}\int\frac{d^3k}{q^2}\frac{F(k;\xi)}{k^2+\mathcal{M}^2(k;\xi)}\Bigg[\mathcal{G}(q)\Bigg\{2a(k,p)\mathcal{M}(k;\xi)-\mathcal{M}(k;\xi)\left[\frac{1}{F(k;\xi)}-\frac{1}{F(p;\xi)}\right]\Omega(k,p)$$

$$+\left[\frac{2b(k,p)\mathcal{M}(k;\xi)+c(k,p)}{2q^2}\right]\left(-q^4+2q^2(k^2+p^2)-(k^2-p^2)^2\right)\Bigg\}$$

$$+\xi\Bigg\{a(k,p)\mathcal{M}(k;\xi)+b(k,p)\mathcal{M}(k;\xi)\frac{(k^2-p^2)^2}{q^2}+\frac{c(k,p)}{2q^2}(k^2-p^2)(k^2-p^2-q^2)\Bigg\}\Bigg],$$

$$\frac{1}{\mathcal{G}(q)} = 1-\frac{N_f\alpha}{2\pi^2}\int d^3k\frac{F(k;\xi)}{k^2+\mathcal{M}^2(k;\xi)}\frac{F(q;\xi)}{q^2+\mathcal{M}^2(q;\xi)}\Bigg[a(k,q)[W_1(k,p)$$

$$+W_2(k,p)\mathcal{M}(k;\xi)\mathcal{M}(q;\xi)]+b(k,q)[W_3(k,p)+W_4(k,p)\mathcal{M}(k;\xi)\mathcal{M}(q;\xi)]$$

$$-\ c(k,q)[W_5(k,p)\mathcal{M}(q;\xi)+W_6(k,p)\mathcal{M}(k;\xi)]\Bigg\},$$

Data Types & Attributes

Data

- raw measurements

  symbols, signals, ...

- corresponding to some **attributes**

  height, grade, heartbeat, ...

Attribute domain

- expresses the **type** of an attribute

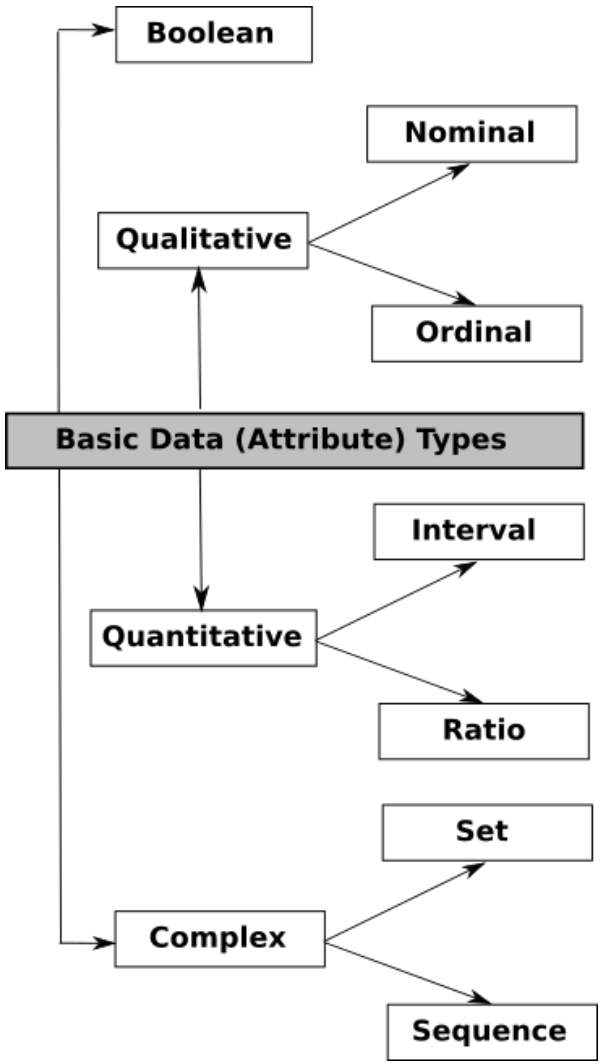
  number, string, sequence, ...

- by the set D of **admissible values**
  - called the **domain** of the attribute

    height up to 3 m, grade from A to F, ...

- and certain **operations** allowed on D

  $1 < 3$, "A" $\geq$ "C", "Jon" $\neq$ "John", ...



Objects, Records, Observations

Object

- A collection of **record**ed measurements (attributes) representing an **entity of observation** (context, meaning)

  e.g a student represented by ID (nominal), age (quantitative), sex (boolean), English proficiency (ordinal), list of absolved courses (set), yearly scores from IQ tests (time-series), ...

- $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in D_1 \times D_2 \times \cdots \times D_m$

- Objects with **mixed types of attributes can be transformed** to objects having boolean or/and quantitative attribute types
  - Be aware of the possible loss of information!
  - *Can you propose some approaches to such transformation?*

    BTW: we'll talk about it in the next lecture...

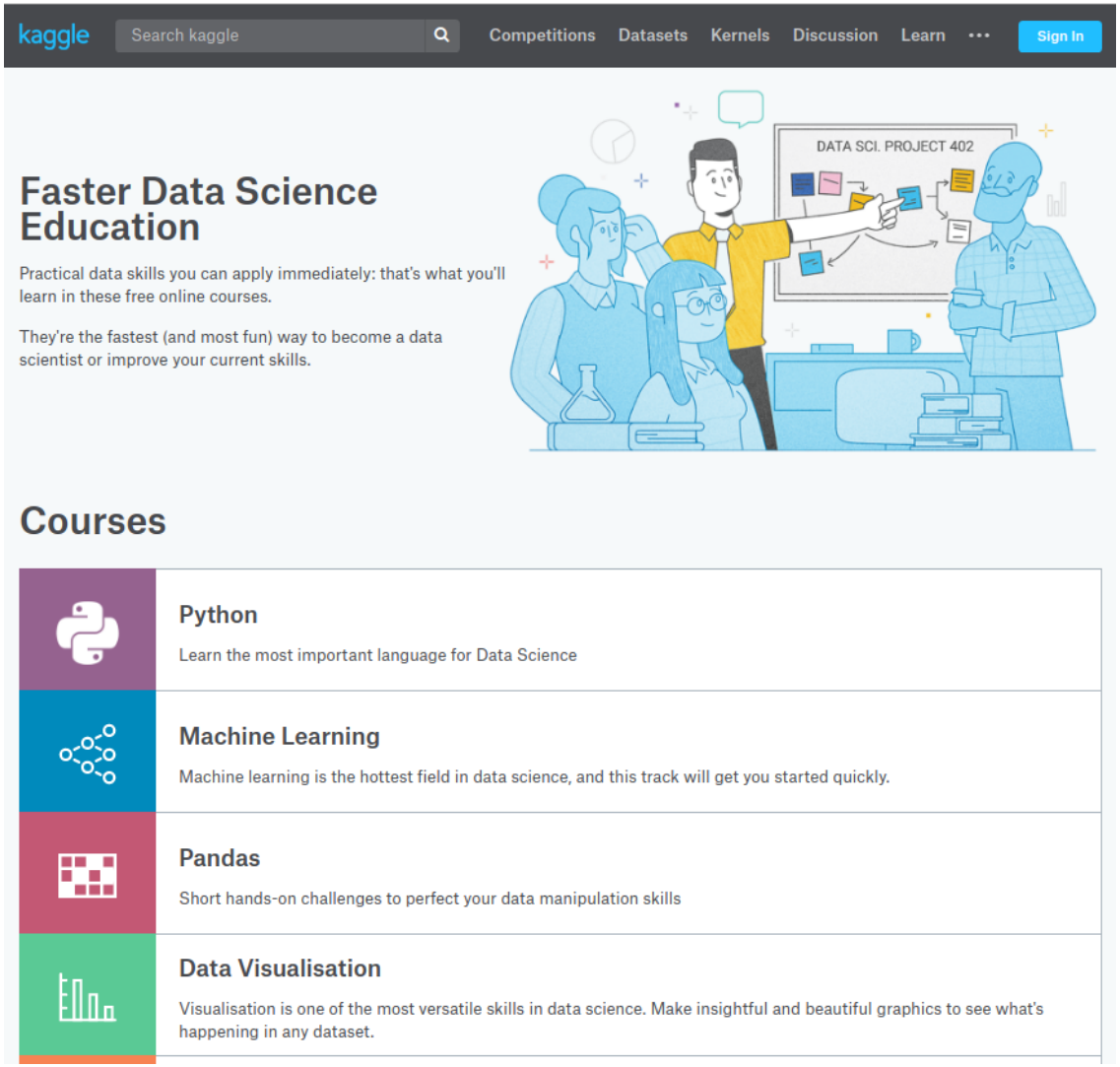
Final Remarks

- **Domain knowledge is important**
  - need to be an understanding between data scientists and domain experts and all the stakeholders
- **Choose the right tool for the job**
  - often the right tool depends on the data you have (to deal with)
  - just because you have/know/like a hammer, not everything is a nail
  - important is the project and not the tool, thus, know the requirements
- **Talking about big data is rather a jibber-jabber**
  - try to better specify things, i.e. talk about sensor data, blog data, tweets, video streaming, etc.
  - leave buzzwords for the managers and sales people

Thanks for your attention

Homework

Learn (basics of) Python



Questions?

tomas.horvath@inf.elte.hu