UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Lesson 1.

Introduction

DATA SCIENCE AND AUTOMATION COURSE

MASTER DEGREE SMART TECHNOLOGY ENGINEERING

TEACHER
Mirko Mazzoleni

PLACE
University of Bergamo

Who I am

- **Name:** Mirko Mazzoleni

- **Studies:** Ph.D. Engineering and Applied Sciences at *University of Bergamo* (Control specialization) + Master degree Computer Engineering (CE) at *University of Bergamo*

- **Currently:** Assistant Professor @ *University of Bergamo*

  - ✓ System identification, machine learning, fault detection, condition monitoring
  - ✓ System identification and data analysis (Master Degree Computer Engineering)
  - ✓ Data science and automation (Master Degree Mechanical Engineering)

- **Contact details:**

  - ✓ mirko.mazzoleni@unibg.it
  - ✓ https://mirkomazzoleni.github.io/
  - ✓ http://cal.unibg.it/ **CAL research laboratory**
  - ✓ https://www.facebook.com/calunibg/

Course content

Part I: Data science

1. **Introduction to data science**
   - 1.1 The business perspective
   - 1.2 CRISP-DM process
   - 1.3 Supervised vs. unsupervised problems
2. **Linear regression**
3. **Feasibility of learning**
   - 3.1 Bias-Variance tradeoff
4. **Logistic regression**
5. **Overfitting and regularization**
   - 5.1 Validation and cross-validation
   - 5.2 Performance metrics
6. **Decision trees**
7. **Neural networks**
8. **Machine vision**
   - 8.1 Classic approaches
   - 8.2 Convolutional neural networks and deep learning
9. **Unsupervised learning**
   - 9.1 k-means clustering
   - 9.2 Principal Component Analysis
10. **Fault diagnosis**
    - 10.1 Model-based fault diagnosis
    - 10.2 Signal-based fault diagnosis
    - 10.3 Data-driven fault diagnosis

Course content

Part II: Automation

12. Introduction to industrial automation
13. Introduction to PLC
14. Ladder language
15. Structured text language
16. Automatic PLC code generation
17. Laboratory experience

Evaluation

- Written exam – 2 hours
- Theoretical open questions and exercises

Up to 25 points

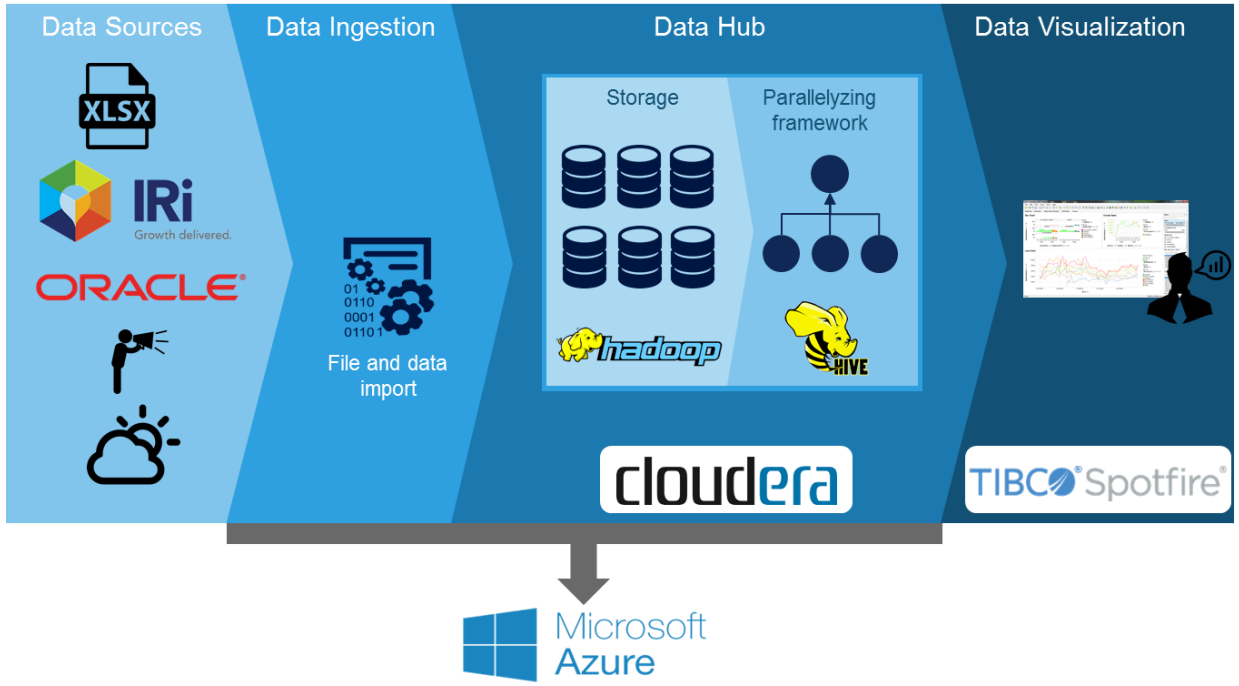
+

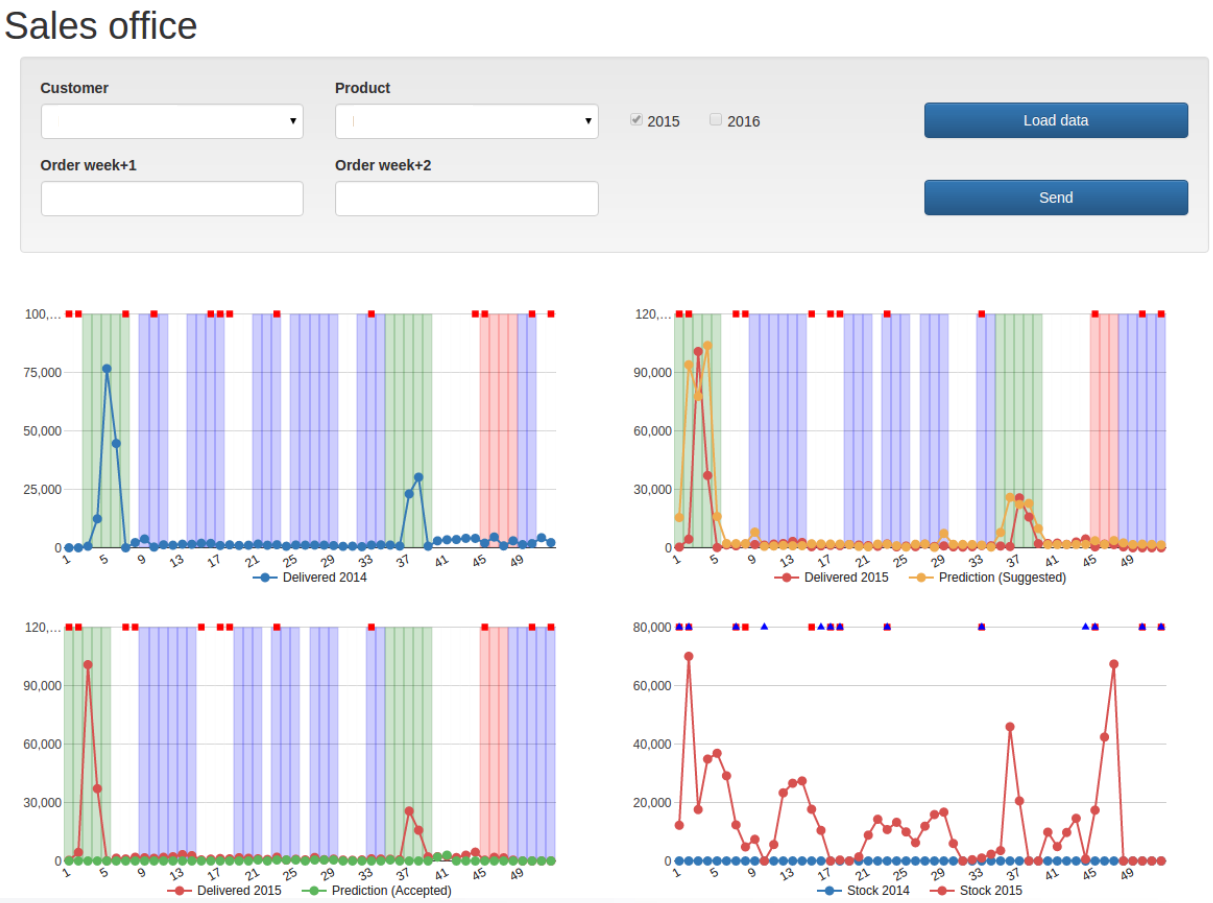
- **[OPTIONAL]** Small data analysis project (groups of max 3 people)

Up to 6 points

Data science projects in the CAL research group

1. **Forecasting of sales volume** (for food industry)



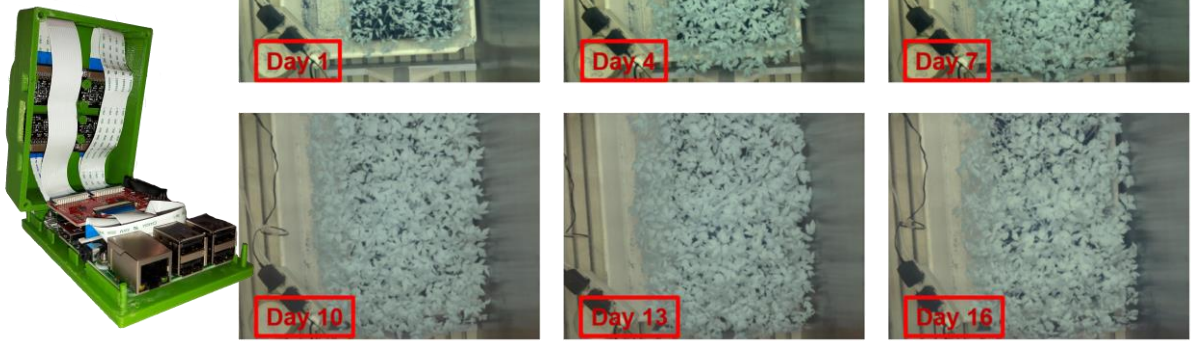


- Development of the data management platform
- Algorithm design
- Testing/validation

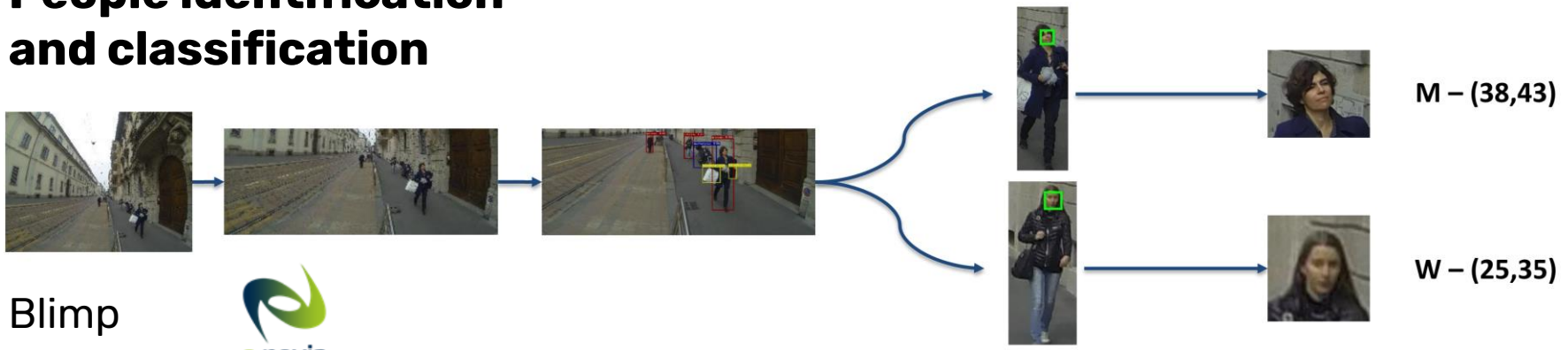
Data science projects in the CAL research group

2. Image processing

Plant disease classification



People identification and classification



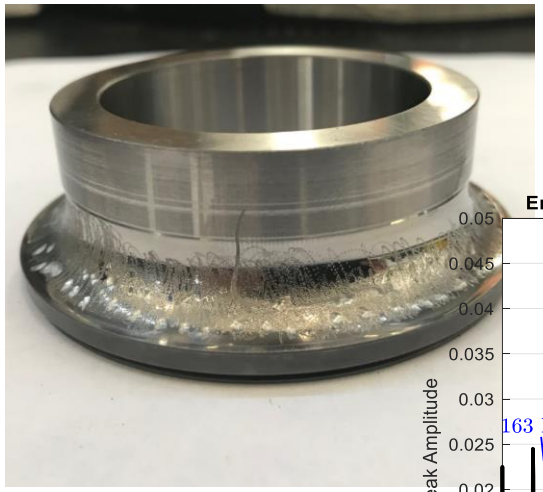
Blimp

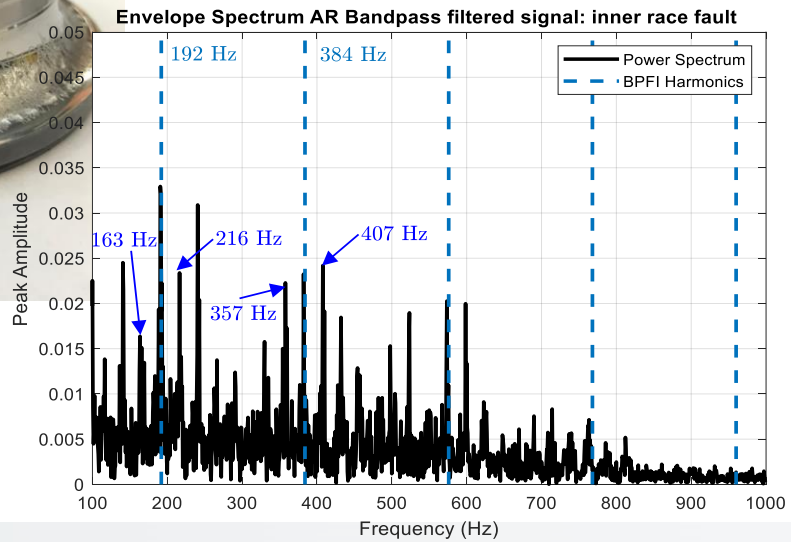
e-novia

Data science projects in the CAL research group

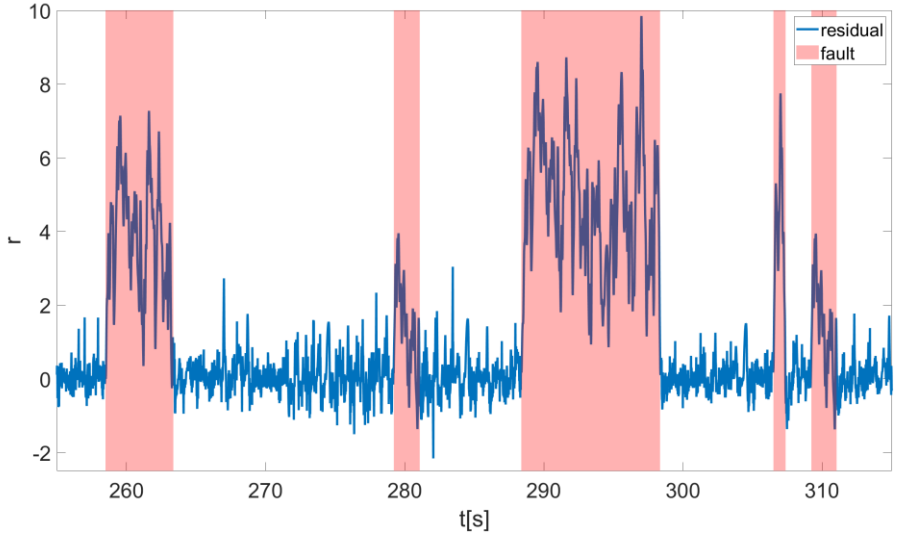
3. Fault diagnosis

Bearing inner race fault





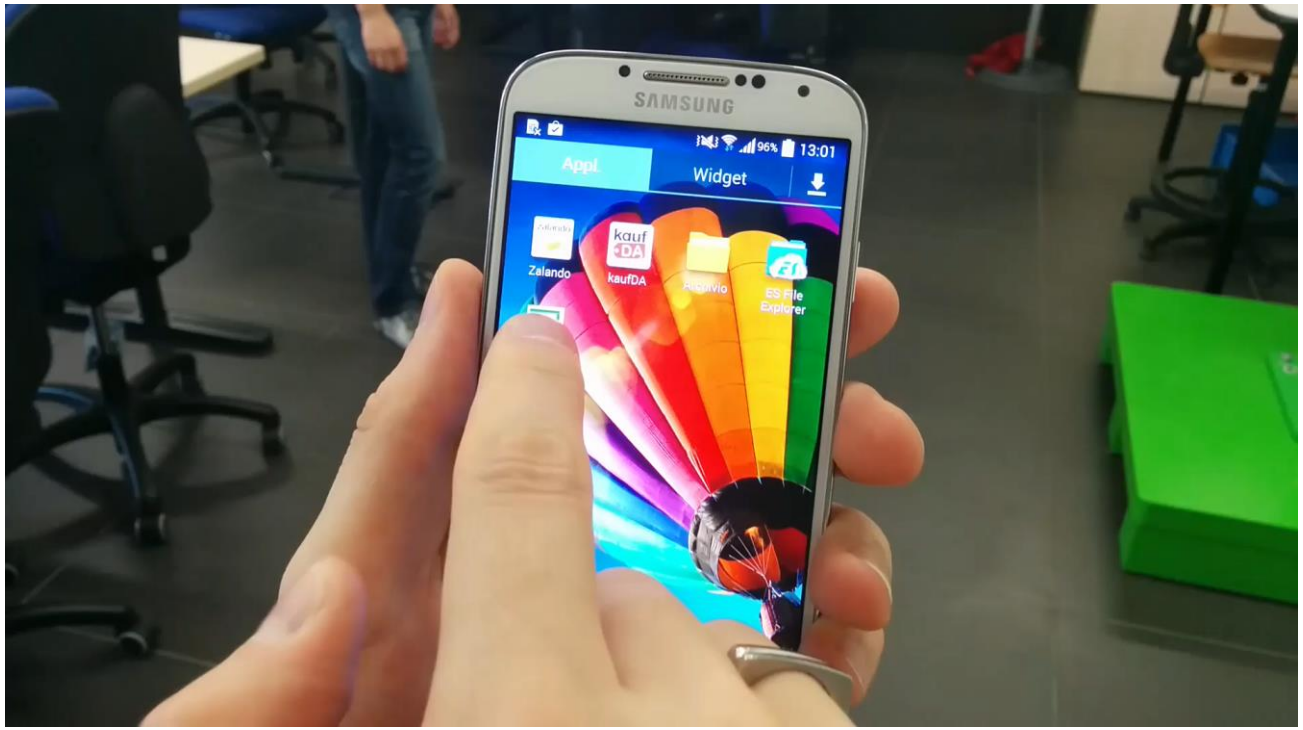
Ballscrew jam in EMA



Data science projects in the CAL research group

4. Industrial automation

ICT for remote mantainance



Automatic transplant machine

Outline

1. Introduction to data science

2. The business perspective and the CRISP-DM process

3. Supervised vs. unsupervised problems

Outline

1. **Introduction to data science**

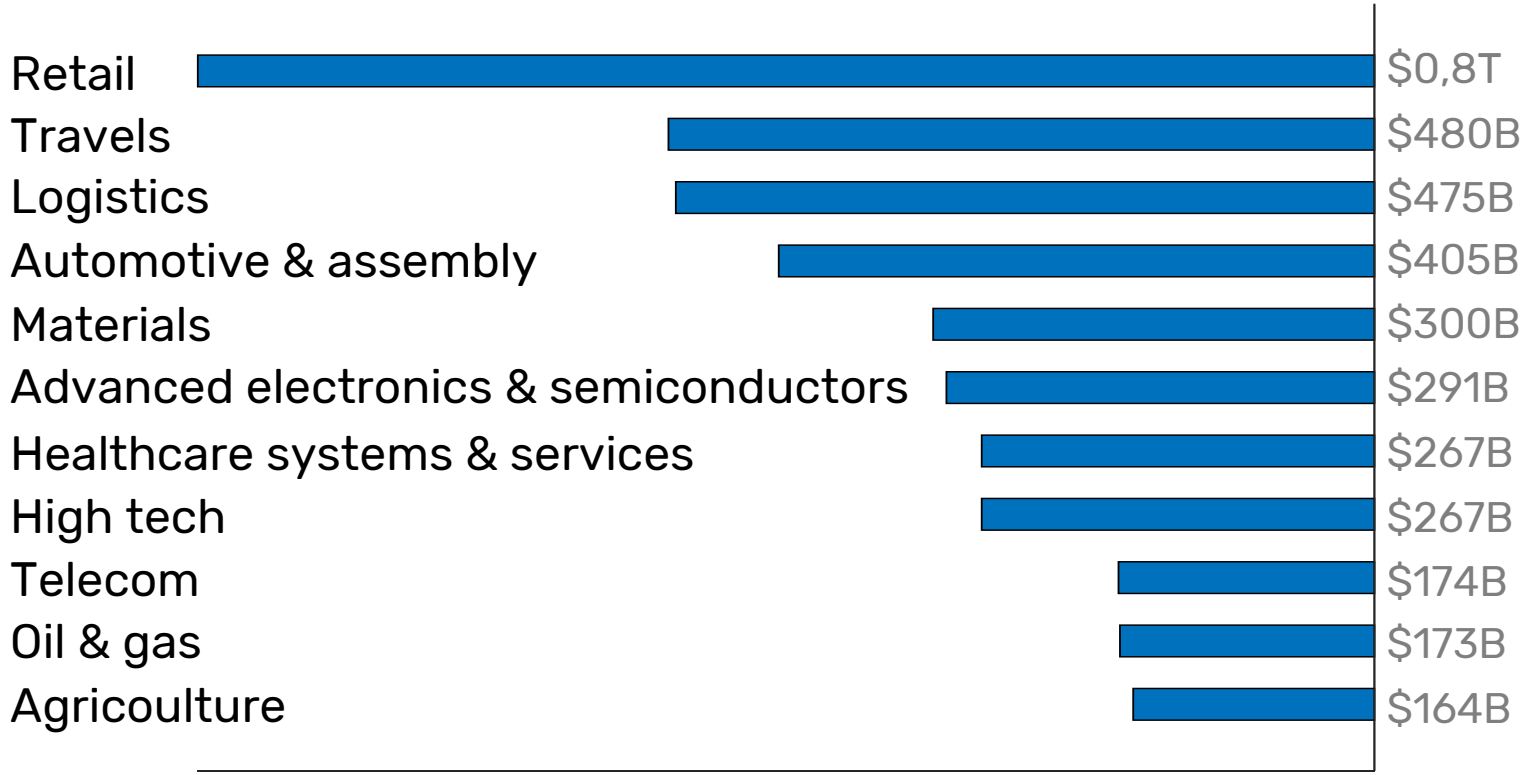
2. The business perspective and the CRISP-DM process

3. Supervised vs. unsupervised problems

Why

Business value created by the AI up to 2030 [1]

$13 Trillions



- It is **difficult** to find an industrial sector **that will not benefit** from AI in the near future

We will use the terms "machine learning", "data mining", "data science" quite interchangeably in this course

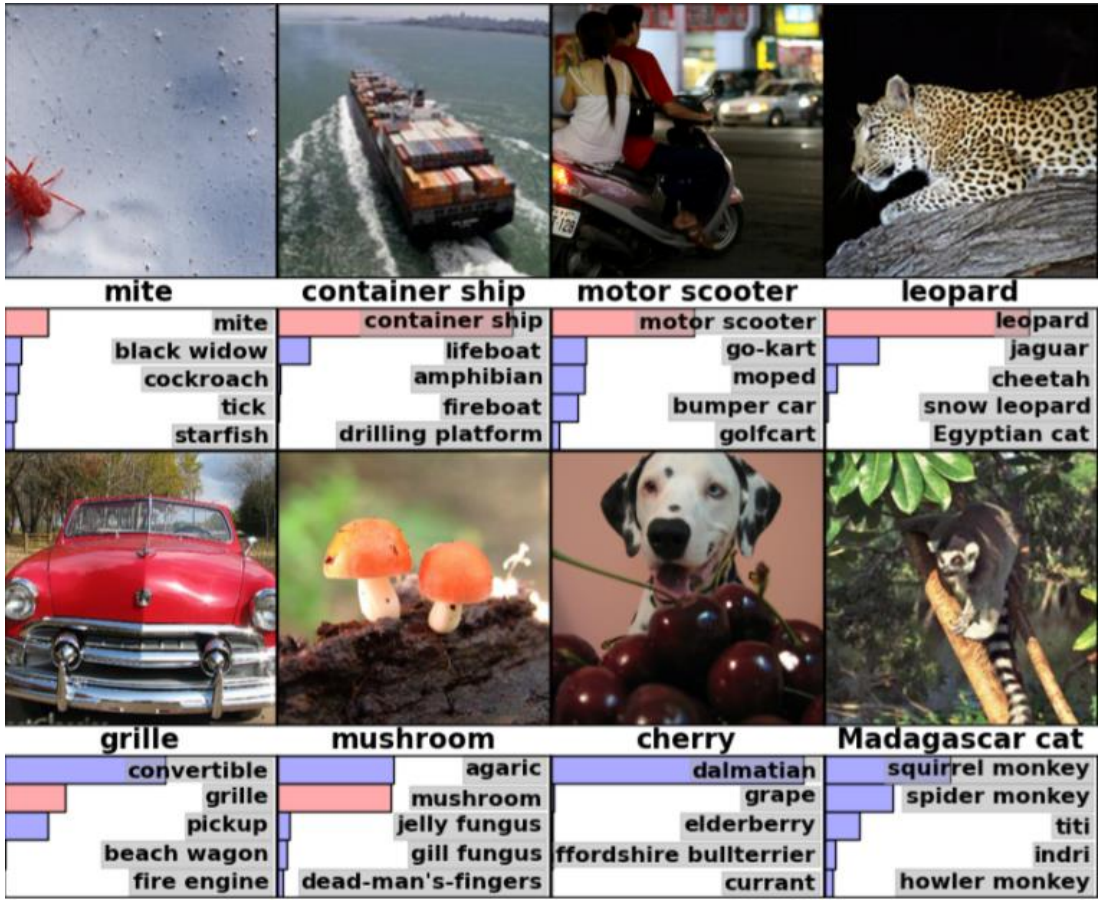
Why

Data science has been deemed as the **sexiest job** of the 21st century

- Virtually every aspect of business is now open to **data collection** (operations, manufacturing, supply-chain management, customer behaviour, marketing campaigns)
- Collected information need to be **analyzed properly** in order to get **actionable results**
- A huge amount of data requires **specific infrastructures** to be handled
- A huge amount of data requires **computational power** to be analyzed
- We can let computers perform decisions given **previous examples**
- Rising of **specific job** titles

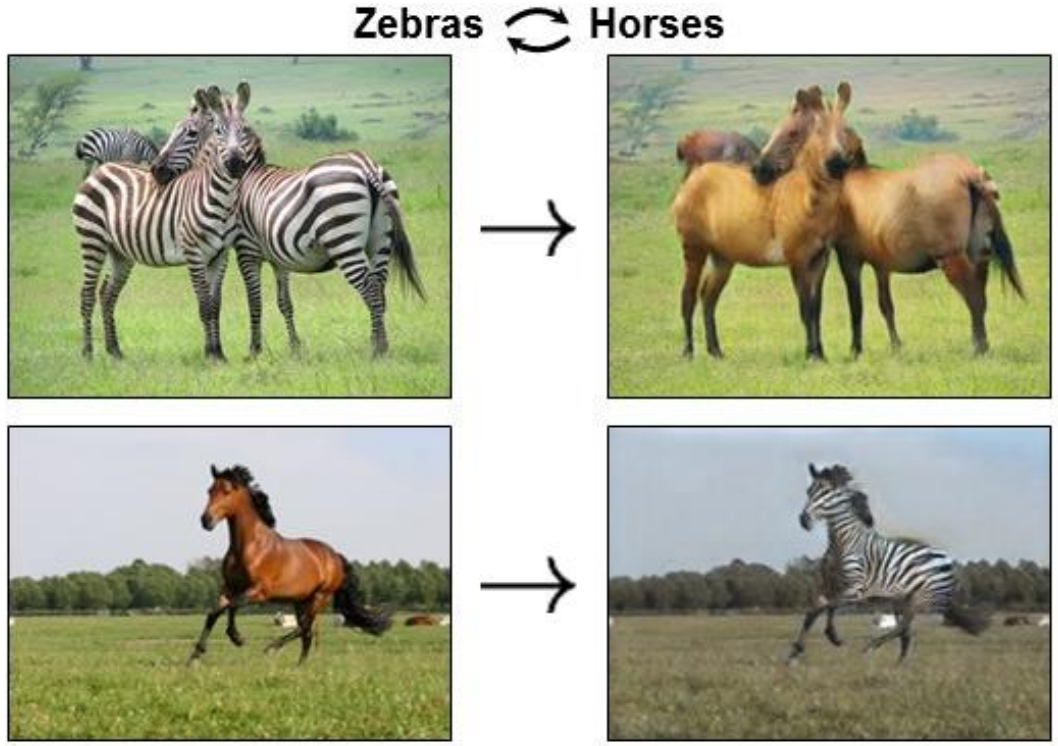
Learning examples

Recent years: stunning breakthroughs in **computer vision** applications



Learning examples

Recent years: stunning breakthroughs in **computer vision** applications



What learning is about

Machine learning and data science are meaningful to be applied if:

1. A pattern exists
2. We cannot pin it down mathematically (an analytical solutions does not exists)
3. **We have data on it**

Assumption 1. and 2. are not mandatory:

- If a pattern does not exist, I do not learn anything
- If I can describe the pattern mathematically, I will not presumably learn the best relation
- The real constraint is assumption 3

Data types

The data can have different formats. The most typical is that of a table

House area(feet^2)	**# bedrooms**	**Price** $(1000\$)$
523	1	115
645	1	150
708	2	210
1034	3	280
2290	4	355
2545	4	440

- **AIM:** predict house prices

Regression

- The data can come from a database or from .csv, Excel files…

A

A

B → Learn the relation **from** House area **to** Price

B → Learn the relation **from** House area **AND** #bedrooms **to** Price

Data types

Another type of data can be an **image**

Picture	Label
	Cat
	Not cat
	Cat
	Not cat

- **AIM:** recognize if there is a cat in the image

Classification

- Learn the relation **from** an image **to** a «class of belonging» (cat vs. not cat)

Data are dirty

Garbage IN, garbage OUT

Data problems:

- Missing values
- Not correct values

House area(feet^2)	**# bedrooms**	**Price** $(1000\$)$
523	1	115
645	1	0,001
708	unknown	210
1034	3	unknown
unknown	4	355
2545	unknown	440

Structured data

Different data types

Images, audio, text → **Not structured data**

Machine learning vs. data science

House area (feet^2)	**# bedrooms**	**# bathrooms**	**Recently renowed**	**Price** (1000$)
523	1	2	No	115
645	1	3	No	150
708	2	1	No	210
1034	3	3	Si	280
2290	4	4	No	355
2545	4	5	Si	440

A

B

Machine learning

- Predict B given A
- Running software (web site\ mobile app)

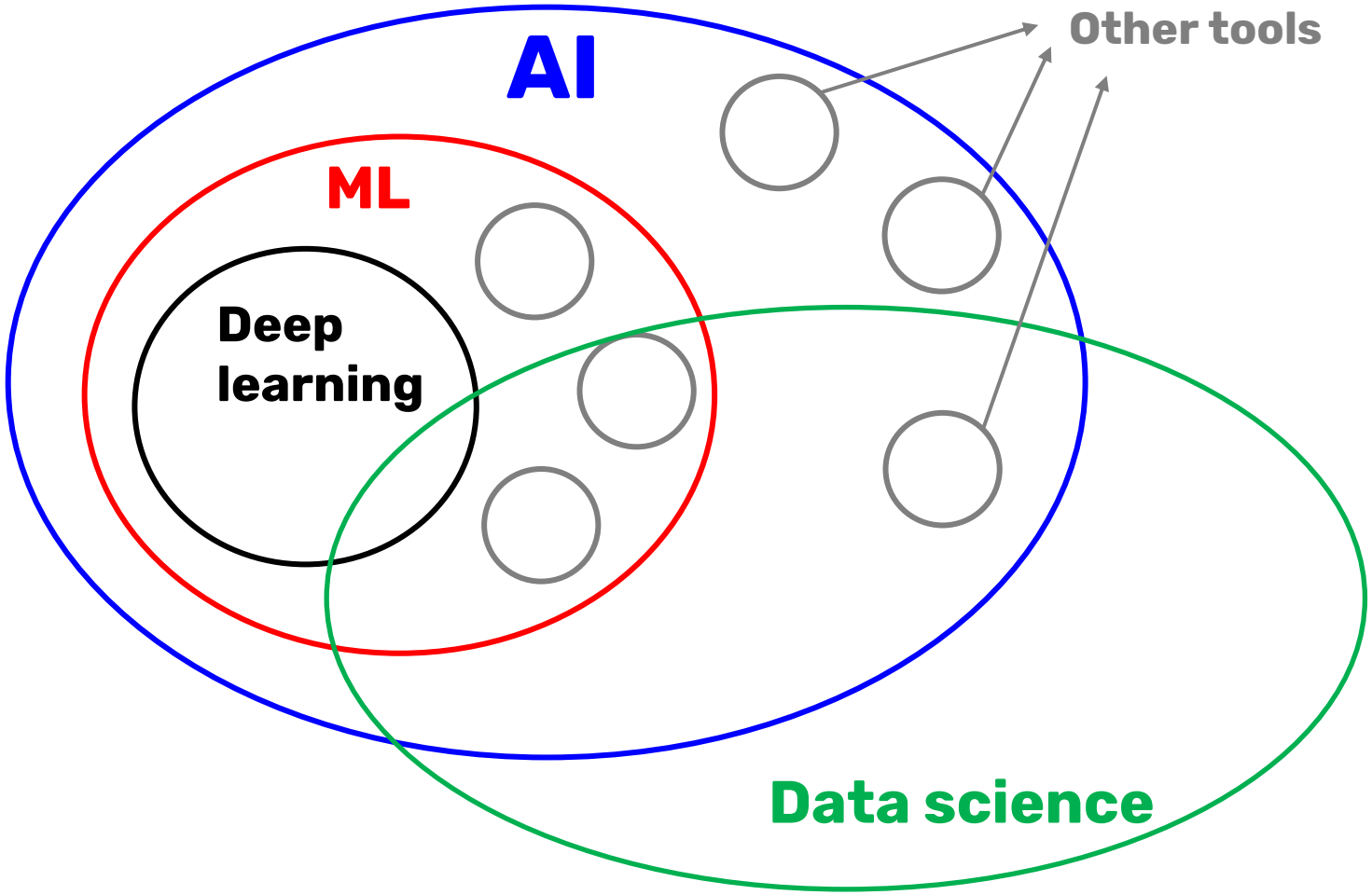
→ **Output:** Code and program

Data science

- Houses with 3 bathrooms are more expensive than those with 2 bathrooms of the same size
- Recently renovated houses cost 15% more

→ **Output:** Slide deck

Machine learning vs. data science



Outline

1. Introduction to data science

2. The business perspective and the CRISP-DM process

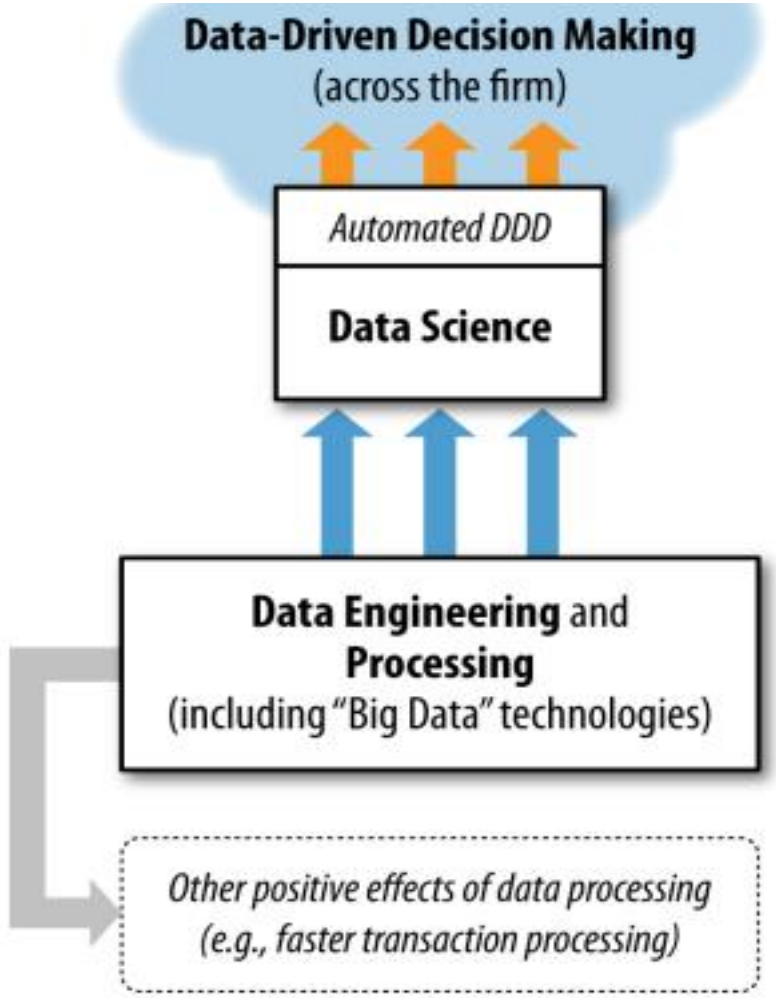
3. Supervised vs. unsupervised problems

Data-analytic thinking

Data-driven decision-making (DDD) refers to the practice of basing **decisions on the analysis of data**, rather than purely on intuition [2, 3]

- Some decisions can be made **automatically** (finance, recommendations)
- **Data engineering and processing** is a fundamental support to industrial analytics
- Data, and the capability to extract useful knowledge from data, should be regarded as **key strategic asset**
  - ✓ Need to invest to acquire the right data (even lose money)
  - ✓ Understand data science **even if you will not do it**

Picture taken from [1]: Provost, Foster, and Tom Fawcett. "Data Science for Business: What you need to know about data mining and data-analytic thinking". O'Reilly Media, Inc., 2013

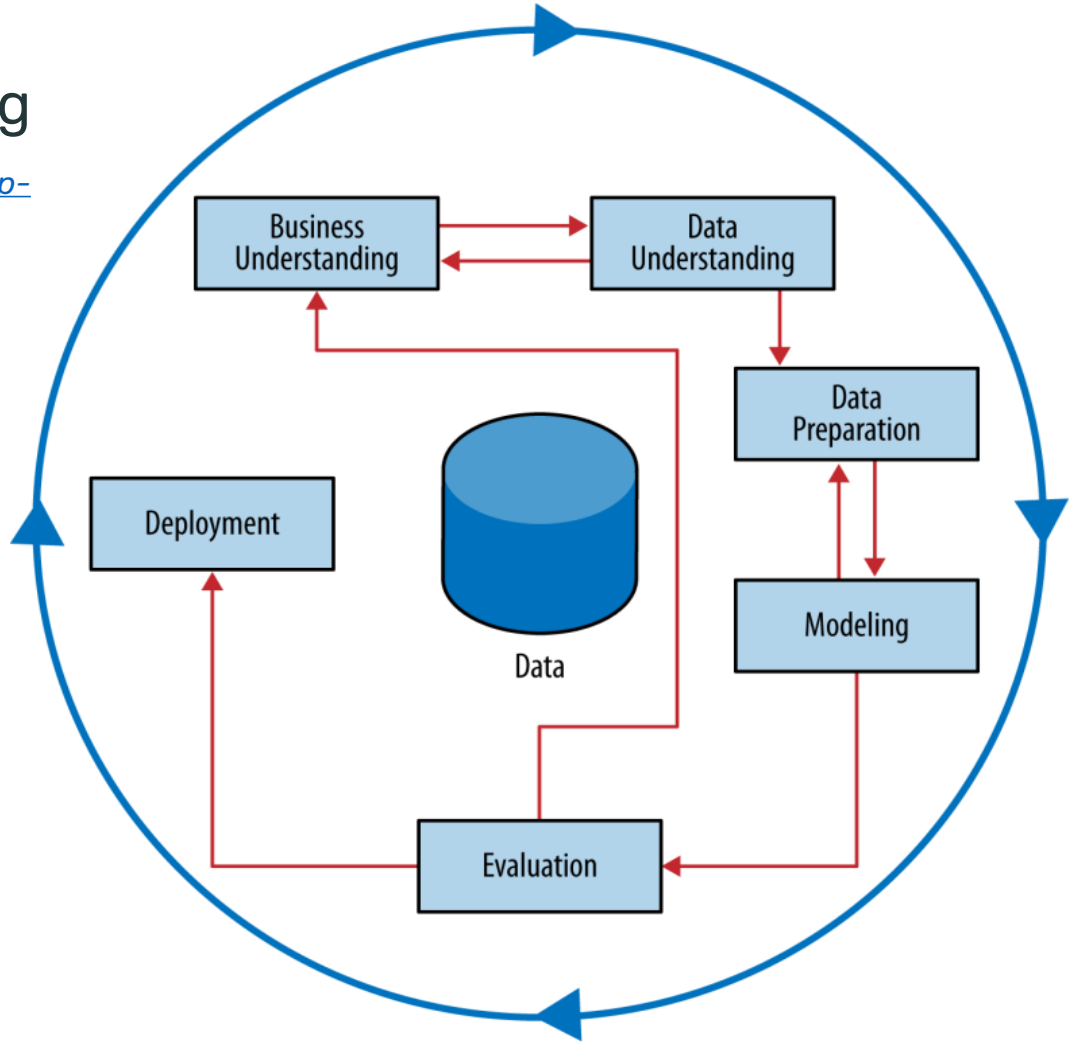


Approaching a data mining problem

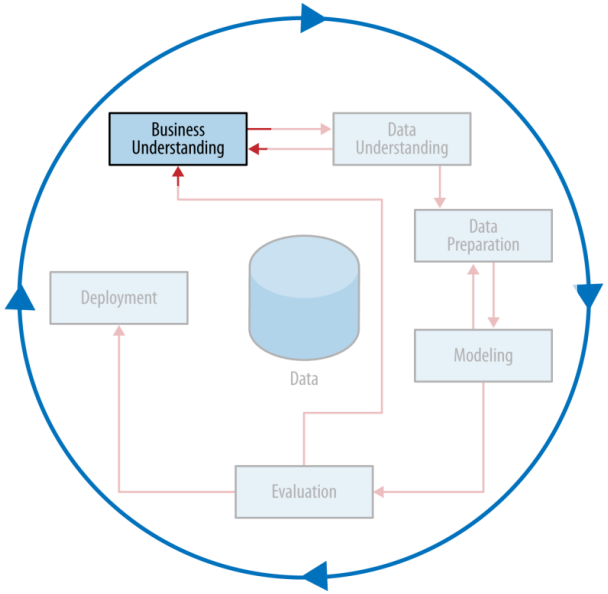
Cross Industry Standard Process for Data Mining (CRISP-DM) https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

Iteration is the rule rather the exception:

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment



CRISP-DM: Business understanding



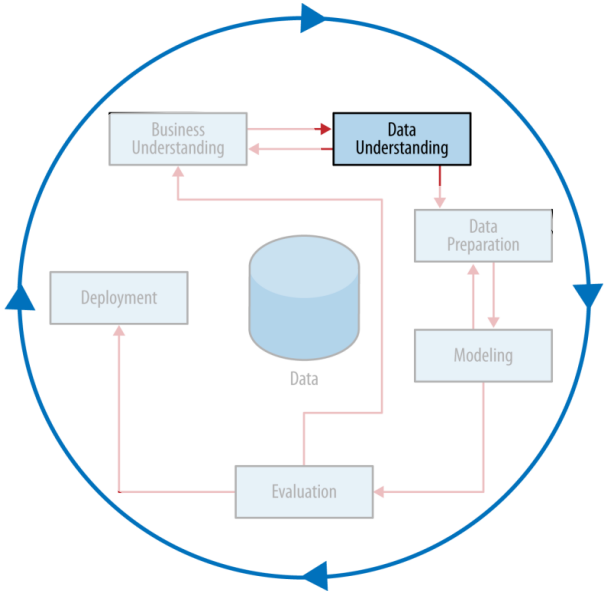
Cast the business problems into **one or more** data science problems

- Frame the problem such that one or more sub-problems involve building models for a **data mining task** *(classification, regression, probability estimation, and so on)*

- Think carefully about the **use scenario**
  - ✓ What exactly do we want to do?
  - ✓ How exactly would we do it?
  - ✓ What parts of this use scenario constitute possible data mining models?

CRISP-DM: Data understanding

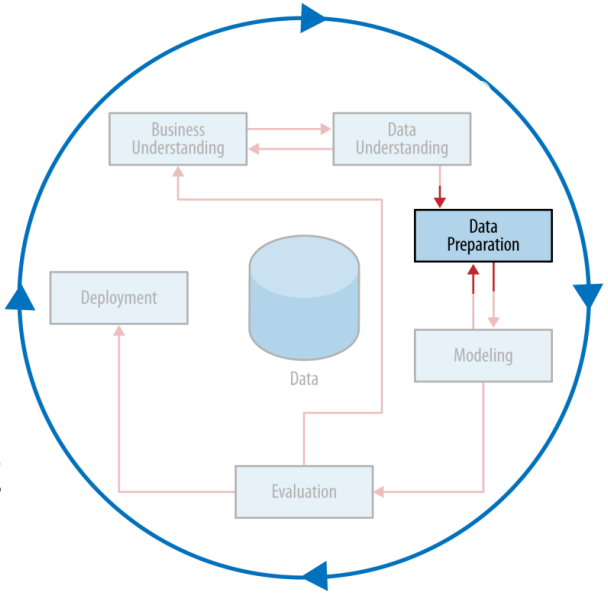
Identify the available and needed data



- Costs/benefits of acquiring each source of data
- Are the data at disposal related to the business problem?
- Can we use a proxy for data that we can not have?
- As data understanding progresses, the solution paths may differ

CRISP-DM: Data preparation

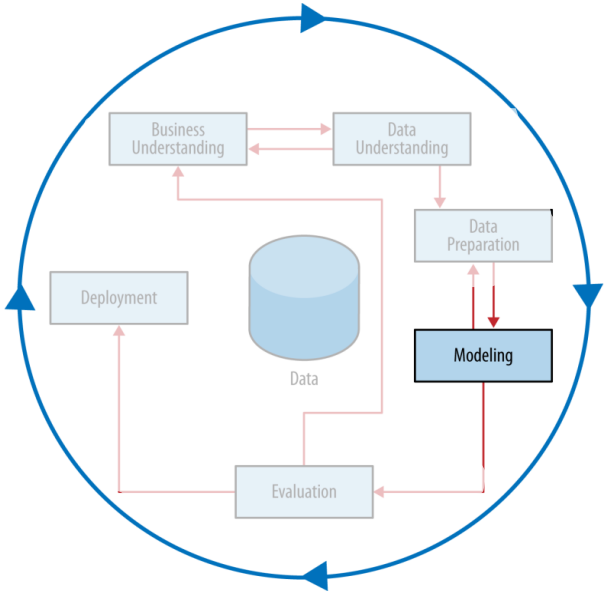
Clean and prepare data for use with algorithms



- Usually the algorithms we employ require data in a different format with respect to the available one
  - ✓ Convert string to numbers, infer missing data, import data from excel files, …
- Data preprocessing/cleaning/labeling [3] **(most of data science project time is spent here)**
- Pay attention to not use historical data that **will not be available** when your model will be used

CRISP-DM: Modeling

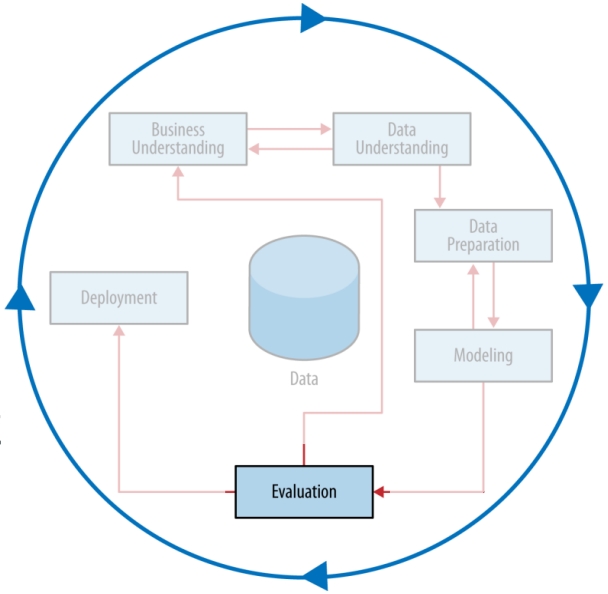
Estimate a mathematical model to extract pattern from data



- In most cases, **standard algorithms** can be directly applied on data
- The aim is to find a model in order **to use it on unseen data**
- The type of the model has to be chosen based on:
  - ✓ What data mining **task** we want to solve
  - ✓ Performance **measures**
  - ✓ Availability of **libraries** for deployement

CRISP-DM: Evaluation

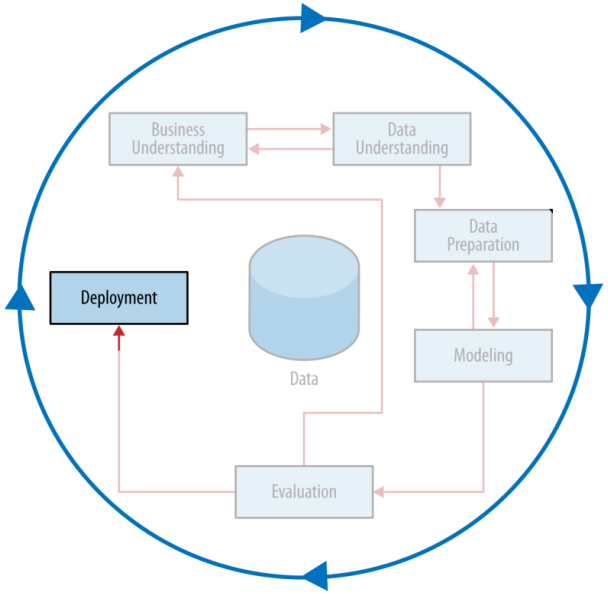
Assess the validity of the results



- We could find patterns that exist only in the particular dataset that we have at disposal **(overfitting)**
- The devised solution and the model's decisions should the **comprehensible** by the stakeholders
- Usually evaluation is performed **before deploying**. In this case, build environments that **closely mimic** the real use scenario
- Evaluation can be performed also **on-line (in production)** [4]

CRISP-DM: Deployment

Put the model (or the data mining steps) into production



- Usually requires to **re-code** the model, to make it compatible with the existing technology
- This step can require quite **investment in time**. Usually the data science team builds a propototype that is then passed to the development team
- For this reason, it is suggested to **involve a member of the development team** in the early phases of the data science project
- Deployment can involve not only the final model, but also **previous phases** (data collection, model building, evaluation)

From business problems to data mining tasks

Each data science project is **unique**. The aim is to **decompose** the business problem into subtasks for which a **common approach** exists.

There are many machine learning algorithms. However, they address a **handful** of tasks:

- Classification and class probability estimation
- Regression
- Symilarity matching
- Clustering
- Co-occurrence grouping
- Profiling
- Link prediction
- Data reduction
- Causal modeling

Outline

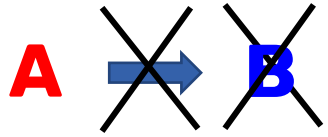
1. Introduction to data science

2. The business perspective and the CRISP-DM process

3. **Supervised vs. unsupervised problems**

Supervised vs unsupervised methods

A specific data science task can be tackled via a **supervised** or **unsupervised** approach

Unsupervised 

"Do our customers naturally fall into different groups?"

*There is no a specific target (or purpose) for the grouping. The aim is only to find **similarities** between individuals*

Supervised 

"Can we find groups of customers who have particularly high likelihoods of canceling their service soon after their contract expire?"

*There is a specific target: find people who will leave when contract expires. In this case, **there must be** data on the **target**.* The value of the target for an individual is called **label** or **class**. We need a dataset of people that **we know they left (labeled dataset)**

Supervised vs unsupervised methods

- Classification and class probability estimation
- Regression
- Causal modeling

Supervised

- Symilarity matching
- Link prediction
- Data reduction

Supervised or Unsupervised

- Clustering
- Co-occurrence grouping
- Profiling

Unsupervised

Business problems as data science examples

Supervised

- Spam e-mail detection system
- Credit approval
- Recognize objects in images
- Find the relation between house prices and house sizes
- Predict the stock market

Unsupervised

- Market segmentation
- Market basket analysis
- Language models (word2vec)
- Social network analysis
- Low-order data representations

- Movies recommendation

Supervised or unsupervised

Additional resources

MOOC

- Learning from data (*Yaser S. Abu-Mostafa - EDX*)
- Machine learning *(Andrew Ng - Coursera)*
- Deep learning *(Andrew Ng - Coursera)*
- The analytics edge *(Dimitris Bertsimas - EDX)*
- Statistical learning *(Trevor Hastie and Robert Tibshirani - Standford Lagunita)*

Books

- Data science for business *(Foster Provost, Tom Fawcett)*
- An Introduction to Statistical Learning, with application in R *(Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani)*
- Neural Networks and Deep Learning *(Michael Nielsen)*
- Pattern Recognition and Machine Learning *(Christopher Bishop)*

References

1. Notes from the AI frontier: Modeling the impact of AI on the world economy, 2018.
2. Provost, Foster, and Tom Fawcett. *"Data Science for Business: What you need to know about data mining and data-analytic thinking"*. O'Reilly Media, Inc., 2013.
3. Brynjolfsson, E., Hitt, L. M., and Kim, H. H. *"Strength in numbers: How does data driven decision making affect firm performance?"* Tech. rep., available at SSRN: http://ssrn.com/abstract=1819486, 2011.
4. Pyle, D. *"Data Preparation for Data Mining"*. Morgan Kaufmann, 1999.
5. Kohavi, R., and Longbotham, R. *"Online experiments: Lessons learned"*. Computer, 40 (9), 103–105, 2007.
6. Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. *"Learning from data"*. AMLBook, 2012.
7. Andrew Ng. *"Machine learning"*. Coursera MOOC. (https://www.coursera.org/learn/machine-learning)