UNIVERSITÀ DEGLI STUDI DI BERGAMO

Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione

DATA SCIENCE AND AUTOMATION

Master degree in MECHATRONICS AND SMART TECHNOLOGY ENGINEERING

Lecture 01: Introduction to data science

SPEAKER
Davide Previtali

PLACE
University of Bergamo

Who am I

Davide Previtali

- **Currently**: Fixed-term Assistant Professor (RTD-A) *@ Control systems and Automation Laboratory (University of Bergamo)*
- **Studies**: Ph.D. in Engineering and Applied Sciences *@ University of Bergamo*. MSc in Computer Science Engineering *@ University of Bergamo*
- **Research topics**: lithium-ion batteries, control systems, black-box and preference-based optimization, machine learning
- davide.previtali@unibg.it

Control systems and Automation Laboratory (CAL)

- @ *University of Bergamo*
- **Members***: 7* professors, 5 Ph.D. students
- **Research topics**: control systems, advanced control, optimization, fault diagnosis, system identification, machine learning
- http://cal.unibg.it/
- https://www.linkedin.com/school/cal-unibg/

Outline

1. Course introduction
2. Data science and the data-driven company
3. Data and its types
4. What we are going to do with data (supervised and unsupervised learning)
5. Static and dynamical models in supervised learning
6. From business problems to data science tasks
7. The data mining life cycle (CRISP-DM)

Outline

1. **Course introduction**
2. Data science and the data-driven company
3. Data and its types
4. What we are going to do with data (supervised and unsupervised learning)
5. Static and dynamical models in supervised learning
6. From business problems to data science tasks
7. The data mining life cycle (CRISP-DM)

Accessing the course materials

All the course materials are provided on **Moodle** at the following link:

DATA SCIENCE AND AUTOMATION

39170-ENG | 2025-26

Password: linear_regression2526

The course materials will be uploaded throughout the semester. Please monitor the Moodle page. Therein, you will also find the **exam results** and **news** on the course

Course prerequisites

It is **strongly suggested** to have a good knowledge of the following topics:

- Linear algebra
- Calculus 1 and Calculus 2
- Statistics

Please fill out the following **questionnaire to assess your knowledge on the prerequisites** (on Moodle):

Prerequisites assessment questionnaire

Course prerequisites

How to «refresh» the prerequisites?

- Linear algebra

  ⇨ UniBg course, course by Gilbert Strang @MIT on YouTube, Addendum provided among the materials for this course

- Calculus 1 and Calculus 2

  ⇨ UniBg course

- Statistics

  ⇨ Brief review at the beginning of this course, UniBg course

Educational objectives

At the end of the course, you will be able to:

- **Formulate** a business problem as a data science problem
- **Formulate** and **solve** regression and classification problems
- **Formulate** and **solve** image analysis and object recognition problems
- **Apply clustering** and **dimensionality reduction** techniques
- **Evaluate** the goodness of a model estimated from data
- **Visualize** and **present** the results of a data science project

Teaching materials

Provided materials

- Lessons' slides
- MATLAB code ⟹ **Focus**: learn the methods, not the programming language

Structure of the course

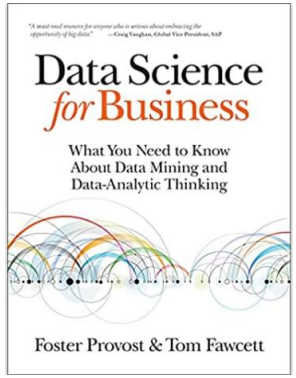
The course will be divided in **theory** and **practice** lessons

- ✓ Theory lessons will be (mostly) of 2 hours on Mondays
- ✓ Practice lessons will be (mostly) of 3 hours on Fridays

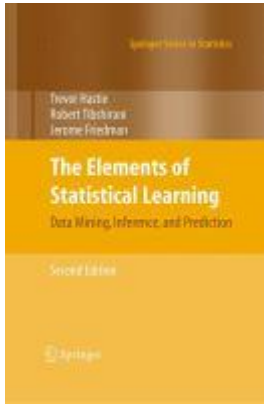
Teaching materials

Suggested books

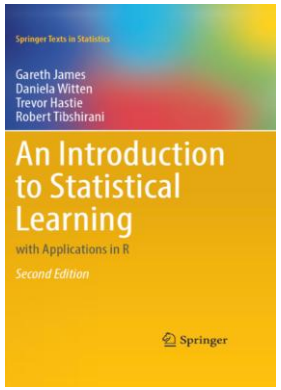
- Foster Provost, Tom Fawcett. **Data Science for Business: What you need to know about data mining and data-analytic thinking**, O'Reilly Media, Inc. (2013)



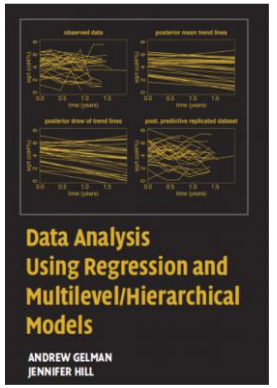
- T. Hastie, R. Tibshirani, J. Friedman. **The elements of statistical learning: data mining, inference, and prediction**, 2° Edition, Springer (2009)



- G. James, D. Witten, T. Hastie, R. Tibshirani. **An Introduction to Statistical Learning**, 2° Edition, Springer (2021)



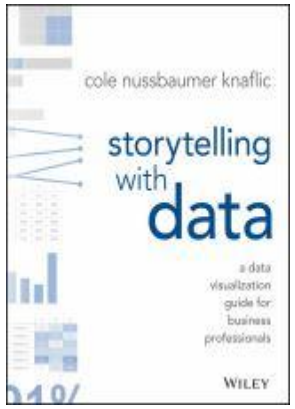
- Andrew Gelman, Jennifer Hill. **Data Analysis Using Regression and Multilevel/Hierarchical Models**, Cambridge University Press (2006)



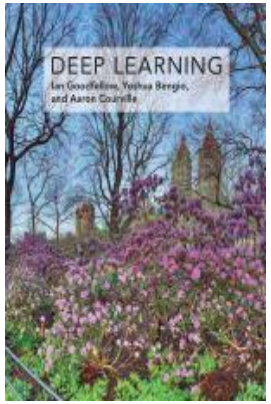
Teaching materials

Suggested books

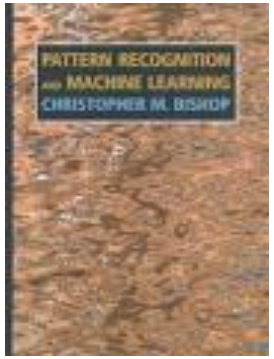
- Cole Nussbaumer Knaflic. **Storytelling with data: a data visualization guide for business professionals**, Wiley (2015)



- I. Goodfellow, Y. Bengio, A. Courville. **Deep Learning**, The MIT Press (2016)



- Christopher Bishop. **Pattern recognition and machine learning**, Springer (2006)



Interactions and feedback

To get a grasp of your **level of understanding during the course**:

1. We will do (together) brief **Quizzes** at the end of each set of slides. These are aimed at testing your overall understanding of the theory
2. I will give you some **Homeworks** to do after each practice lesson
    - Each homework involves analyzing (in MATLAB) a given dataset following what we have seen during the corresponding practice lesson
    - You will be asked to answer a set of multiple-choice questions **on Moodle**
    - These homeworks are optional (**although strongly encouraged**). In addition, they will give a bonus of (at most) **+3 points** to the final grade

Evaluation

- **Written** exam, **1:30 hour**
- True-false and multiple-choice questions
- Open questions

} **Up to 20 points**

+

- Data science **project** with **discussion**
- You will receive more information on the project during the course

} **Up to 10 points**

+

- Bonus points from the **homeworks**

} **Up to 3 points**

You need to complete **both** to pass the exam (the total must be ≥ 18)

How to prepare for the exam?

Written exam:

- Attend the theory and practice lessons
- Study the theory
- Solve the **Exam examples** provided among the course materials
- Solve the **Quizzes** at the end of each set of slides (also come up with your own quizzes!)
- Answer the provided **possible exam open questions** (also come up with your own questions!)

How to prepare for the exam?

Project:

- Attend the theory and practice lessons
- Re-code the practice lessons on your own (try to carry out additional tasks compared to those seen during the practice lessons)
- Do the **Homeworks** provided at the end of each practice lesson

Syllabus

1. Introduction to data science
2. Exploratory data analysis
3. Recap of statistics
4. Maximum likelihood estimation
5. Linear regression
6. Logistic regression
7. Bias-variance trade-off
8. Overfitting and regularization
9. Validation and cross-validation
10. Decision trees
11. Neural networks
12. Convolutional neural networks
13. Clustering methods
14. Principal component analysis

Outline

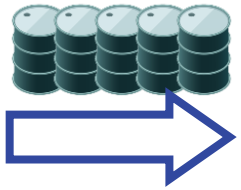
1. Course introduction
2. **Data science and the data-driven company**
3. Data and its types
4. What we are going to do with data (supervised and unsupervised learning)
5. Static and dynamical models in supervised learning
6. From business problems to data science tasks
7. The data mining life cycle (CRISP-DM)

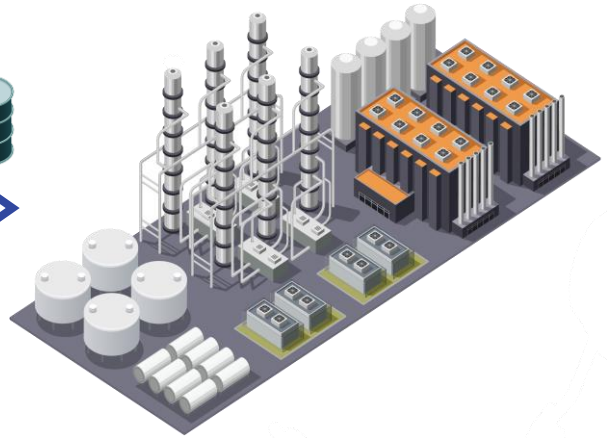
«Data is the new oil»

«Data is the new oil»

Crude oil extraction

Barrels of oil



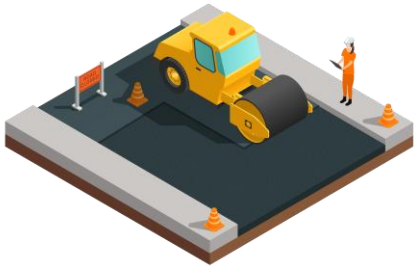


Refinement process

- Fuels
- Oils
- …

- Automobiles,
- Planes,
- Generators,
- Engines,
- …

Asphalt



- Infrastructures,
- Streets,
- …

Plastic

- Bottles,
- Containers,
- Films,
- …

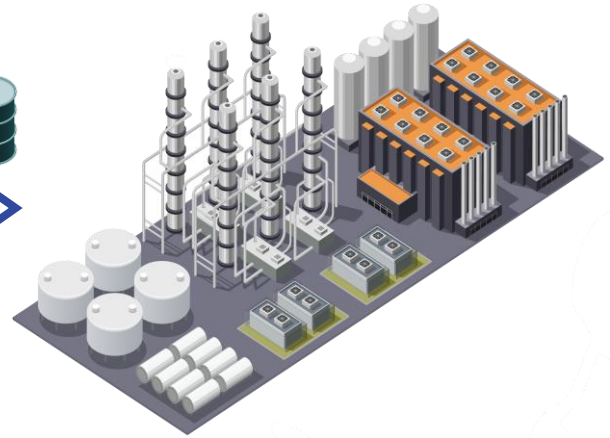
«Data is the new oil»

Crude oil extraction

DATA COLLECTION

Barrels of oil

DATA

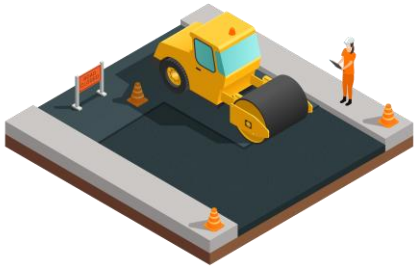


Refinement process

- Fuels
- Oils
- …

- Automobiles,
- Planes,
- Generators,
- Engines,
- …

Asphalt



- Infrastructures,
- Streets,
- …

Plastic

- Bottles,
- Containers,
- Films,
- …

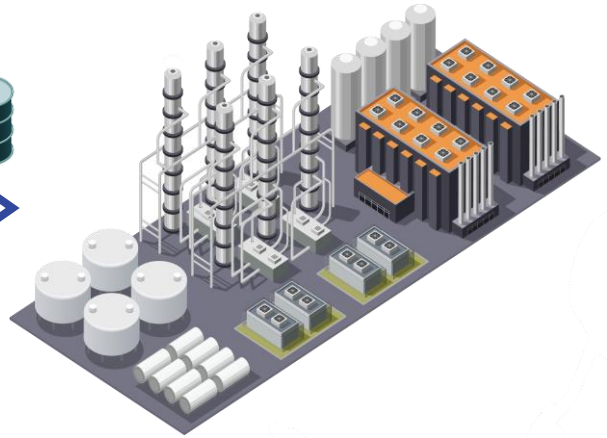
«Data is the new oil»

Crude oil extraction

DATA COLLECTION

Barrels of oil

DATA

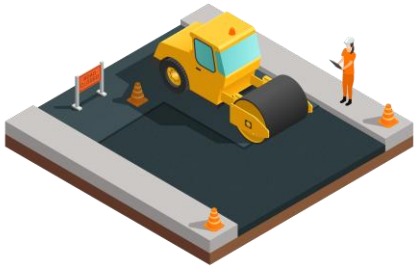


Refinement process

PRODUCT QUALITY

- Automobiles,
- Planes,
- Generators,
- Engines,
- …

GOODS FORECAST



- Infrastructures,
- Streets,
- …

PROCESS OPTIMIZATION

- Bottles,
- Containers,
- Films,
- …

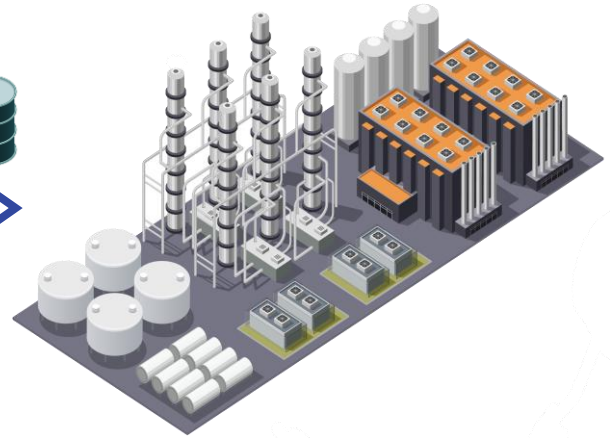
«Data is the new oil»

Crude oil extraction

DATA COLLECTION

Barrels of oil

DATA

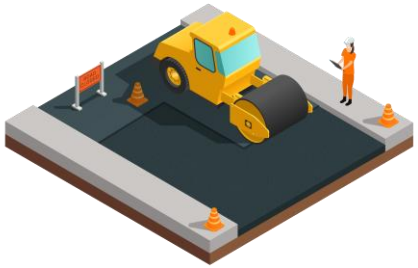


Refinement process

PRODUCT QUALITY

- **Machine parameters optimization**
- **…**

GOODS FORECAST

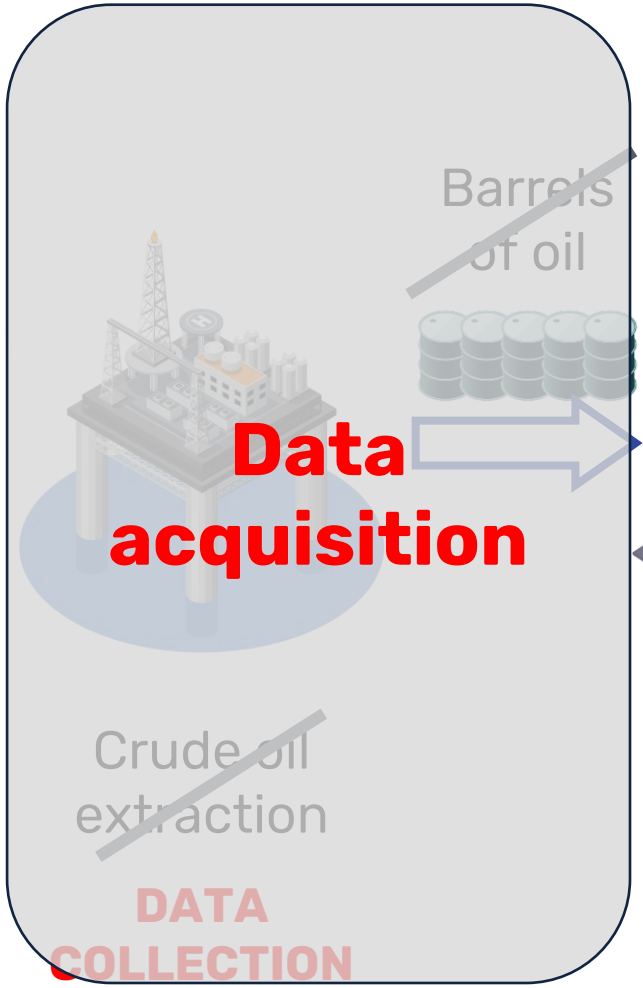


- **Production and purchasing management**
- **…**

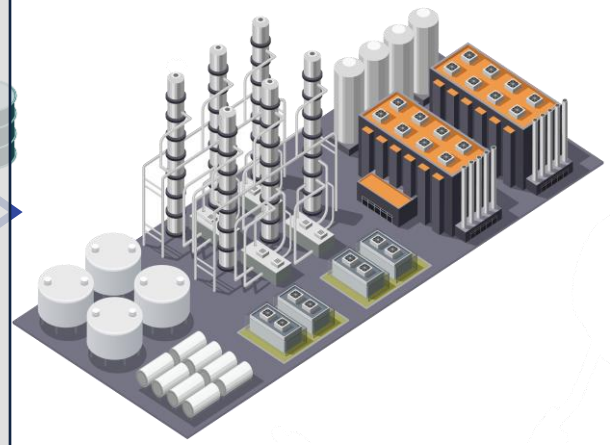
PROCESS OPTIMIZATION

- **Reduction of materials used**
- **…**

«Data is the new oil»



DATA



Refinement process

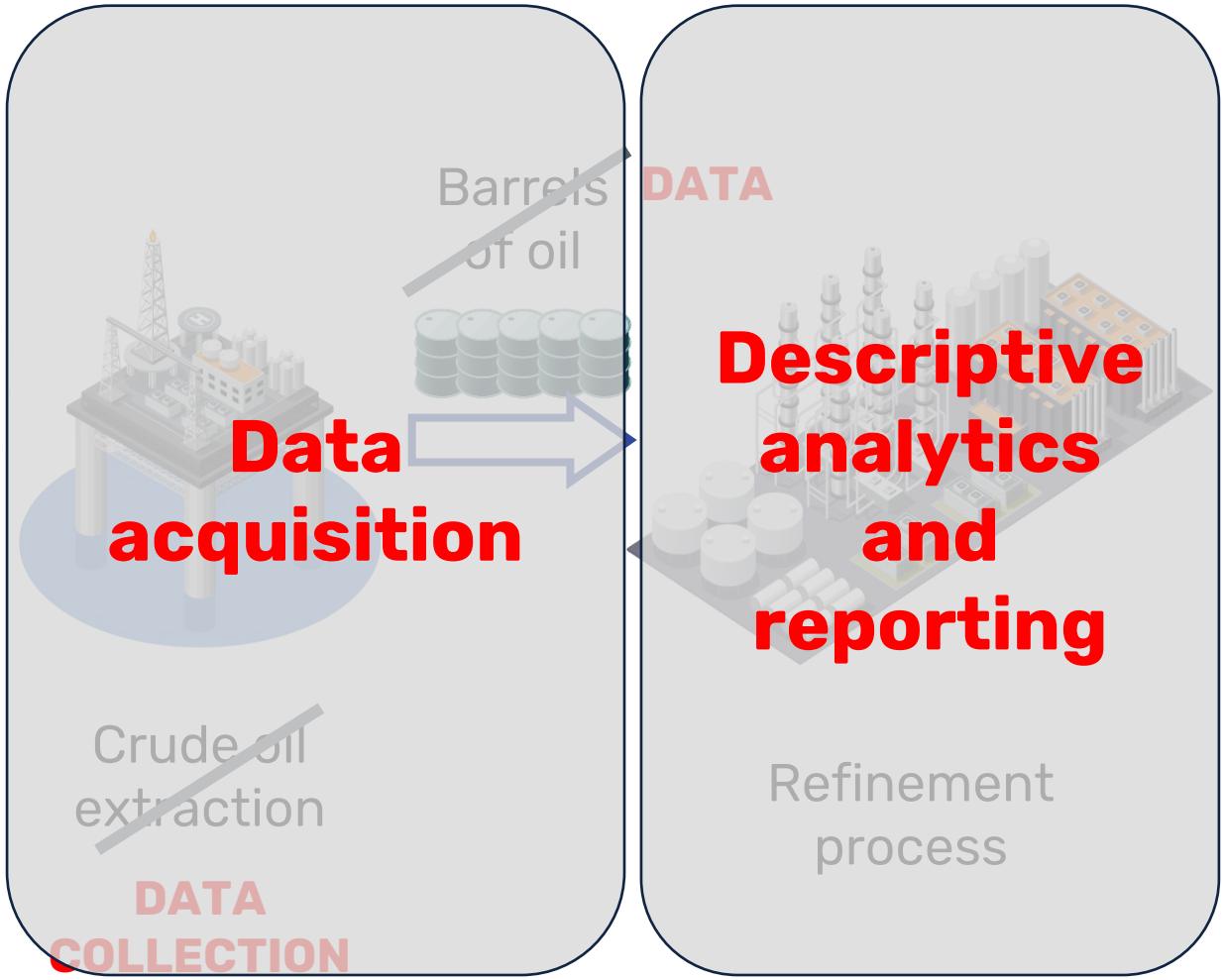
PRODUCT QUALITY

GOODS FORECAST

PROCESS OPTIMIZATION

- **Machine parameters optimization**
- **…**
- **Production and purchasing management**
- **…**
- **Reduction of materials used**
- **…**

«Data is the new oil»



PRODUCT QUALITY

- **Machine parameters optimization**
- **…**

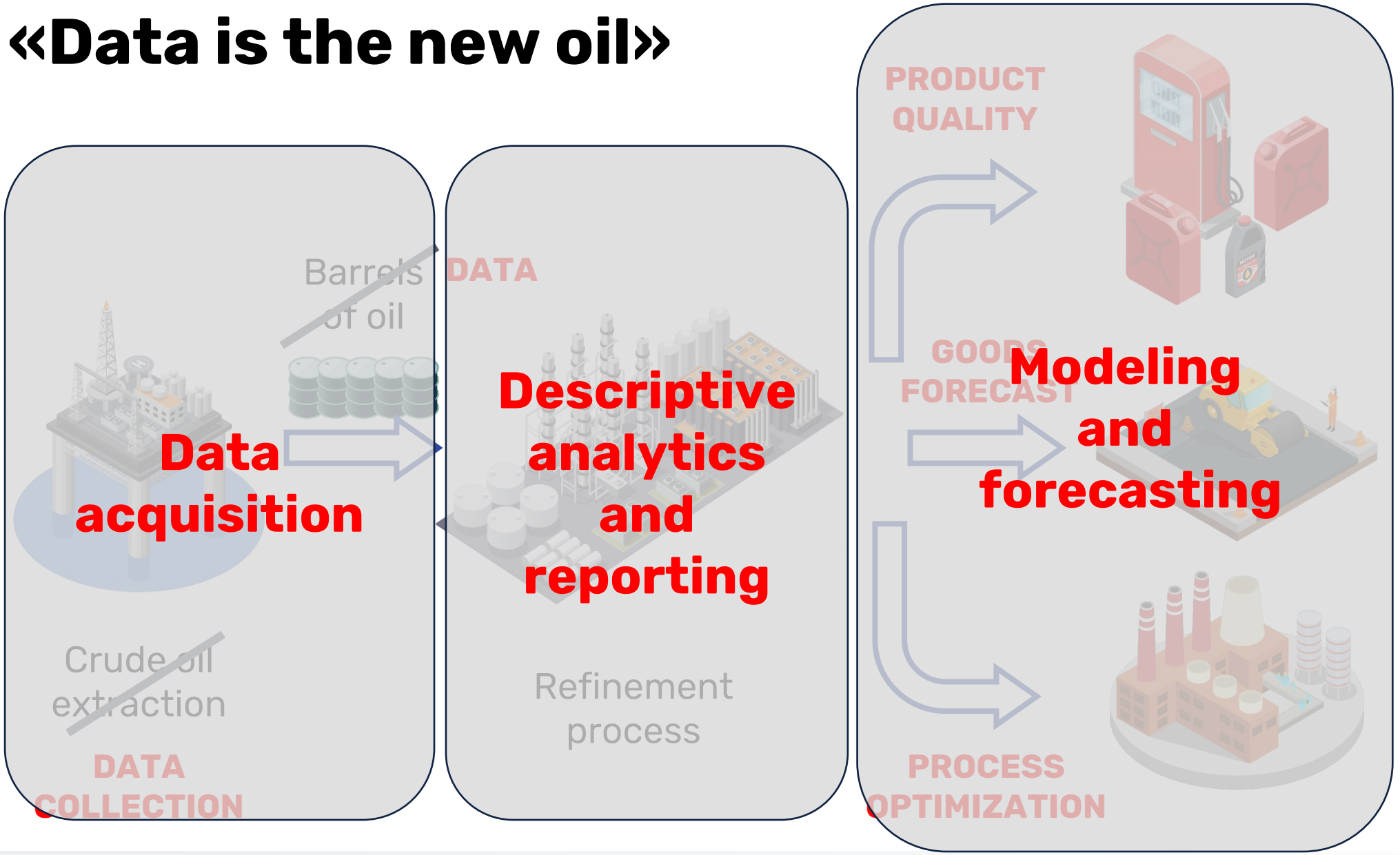
GOODS FORECAST

- **Production and purchasing management**
- **…**

PROCESS OPTIMIZATION

- **Reduction of materials used**
- **…**

«Data is the new oil»

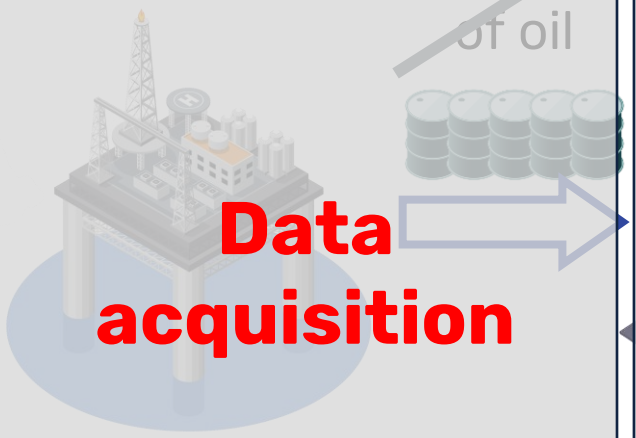


- **Machine parameters optimization**
- **…**

- **Production and purchasing management**
- **…**

- **Reduction of materials used**
- **…**

«Data is the new oil»



Barrels of oil

Crude oil extraction

Data acquisition

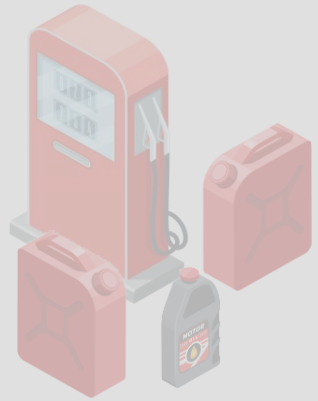
DATA COLLECTION

DATA

Descriptive analytics and reporting

Refinement process

PRODUCT QUALITY



GOODS FORECAST

Modeling and forecasting

PROCESS OPTIMIZATION

- Machine parameters optimization
- …

- Production and [illegible] management
- …

- Reduction of materials used
- …

Actions

Data is the new oil and data science is «sexy»

The data scientist role has been deemed the **sexiest job** of the 21st century [7]

- Virtually every aspect of a 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 **past data**
- Rising of **specific job** titles

Job positions that involve data

Data analyst	Data scientist	Data engineer	Machine learning engineer
• Data retrieval (database queries) • Spot trends and patterns in the data • Visualize the data and produce reports to present information to third parties • …	• Use different machine learning techniques to derive insights from data to guide business decisions • Make predictions on products, assets and consumer behavior based on past data • …	• Design and maintain data management systems • Data collection and management • Make data accessible to the other members of the data science team • …	• Design and implementation of machine learning methods • Extend existing machine learning frameworks and libraries • …

And many more…

Often, career opportunities require a **good mix** of all the aforementioned skills

What is data science?

Data science is a set of fundamental principles, processes and techniques that guide the extraction of knowledge from data with the goal of **improving decision-making**

It is an **interdisciplinary** academic field that is based on:

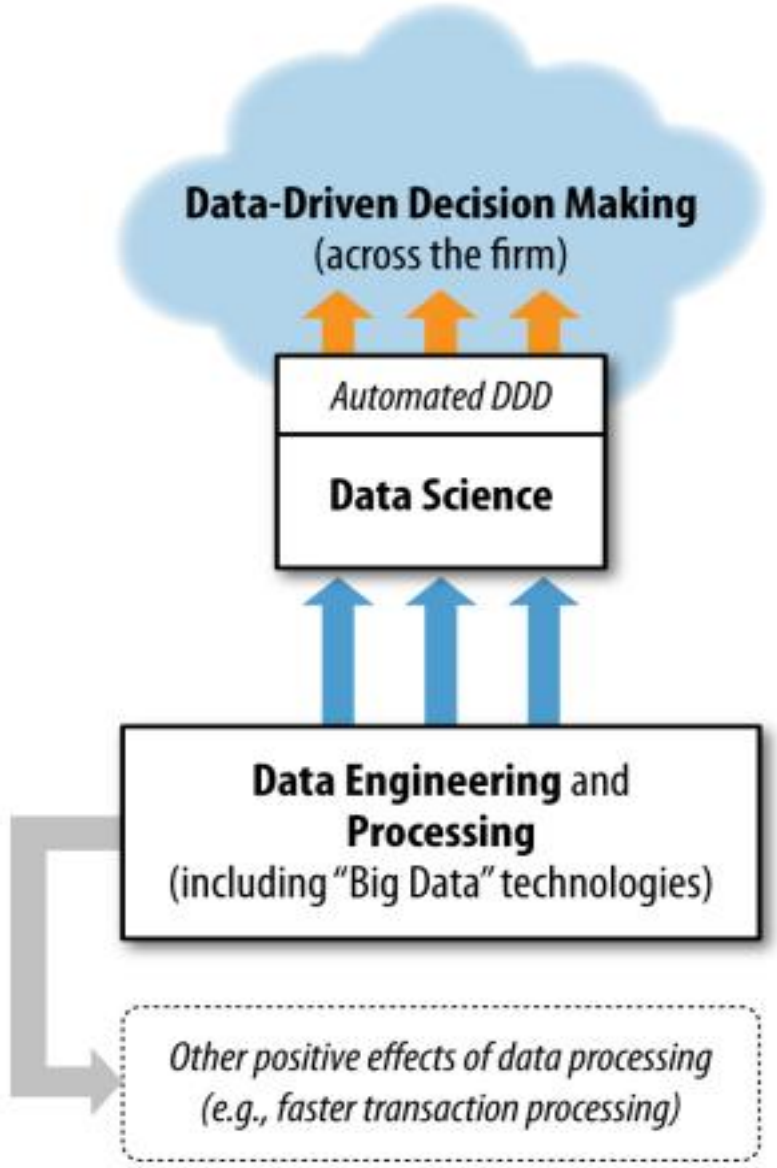
- Mathematics
- Statistics
- Machine learning and artificial intelligence
- Specialized programming

Data mining is the extraction of knowledge from data, via technologies that incorporate data science principles

The data-driven company

Data-Driven Decision-making (DDD) refers to the practice of basing decisions on the analysis of data, rather than purely on intuition [1, 2]

- Some decisions can be made **automatically** (finance, recommendations)
- **Data engineering and processing** support many data-oriented business tasks but do not necessarily involve extracting knowledge or data-driven decision making
- 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]

Anti-hippo culture

Hippos are among the most dangerous animals in Africa. Conference rooms too.

—Jonathan Rosenberg

The road to becoming data-driven

1

Data Denial

Data are not used and are viewed with distrust

2

Data Indifference

There is no interest to acquire or use data

3

Data Aware

Data are collected and used for monitoring, but no decisions are made based on them

4

Data Informed

Data are mainly used by managers in decision-making

5

Data-Driven

Data play a central role in the most disparate decisions that are made in the various business sectors

Why become data-driven?

Data-driven companies are

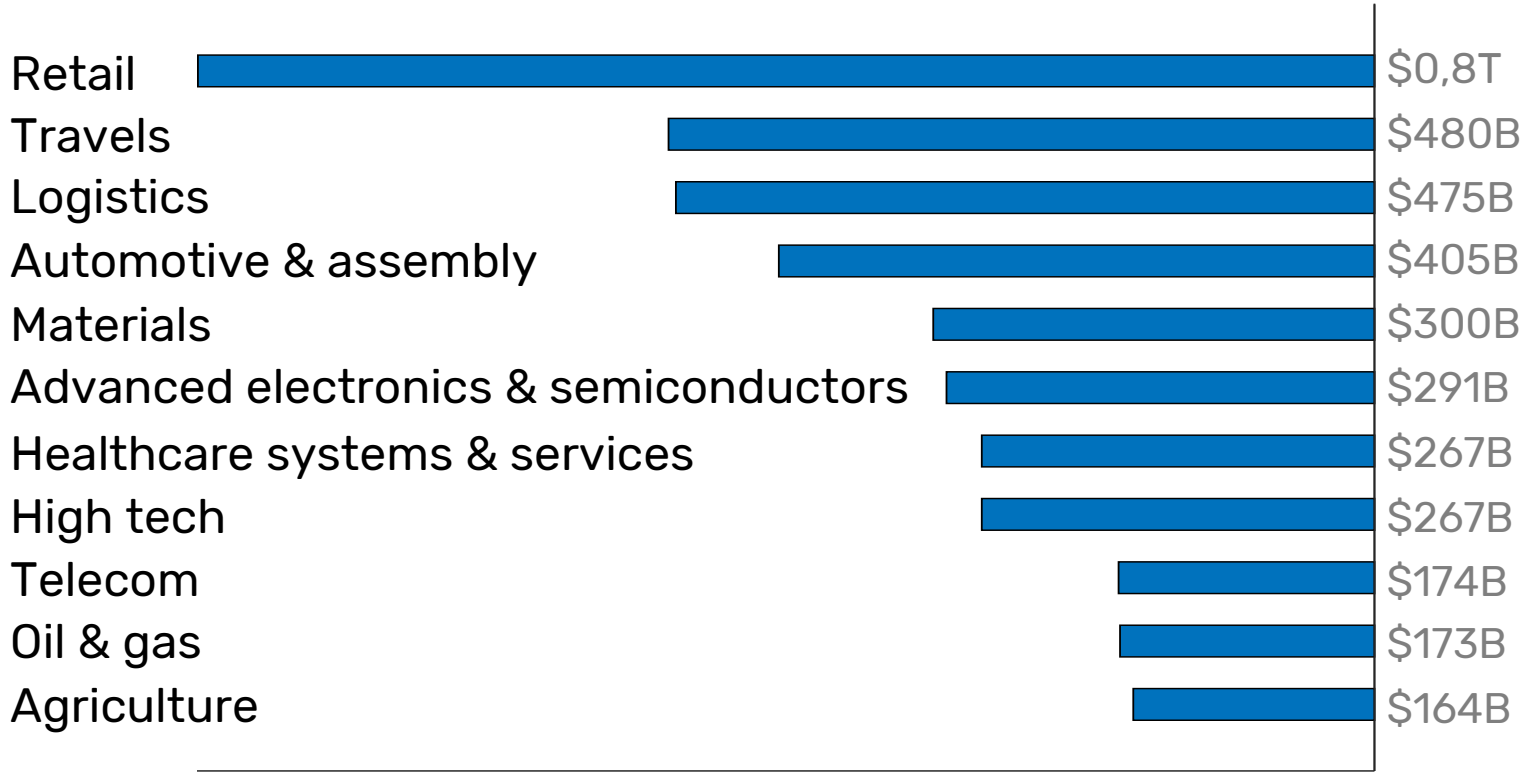
5% more productive [2]

1 $
invested in analytics pays back **13 $** [3]

Why become data-driven?

Business value created by Artificial Intelligence by 2030 [4]

$13 Trillions



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

Outline

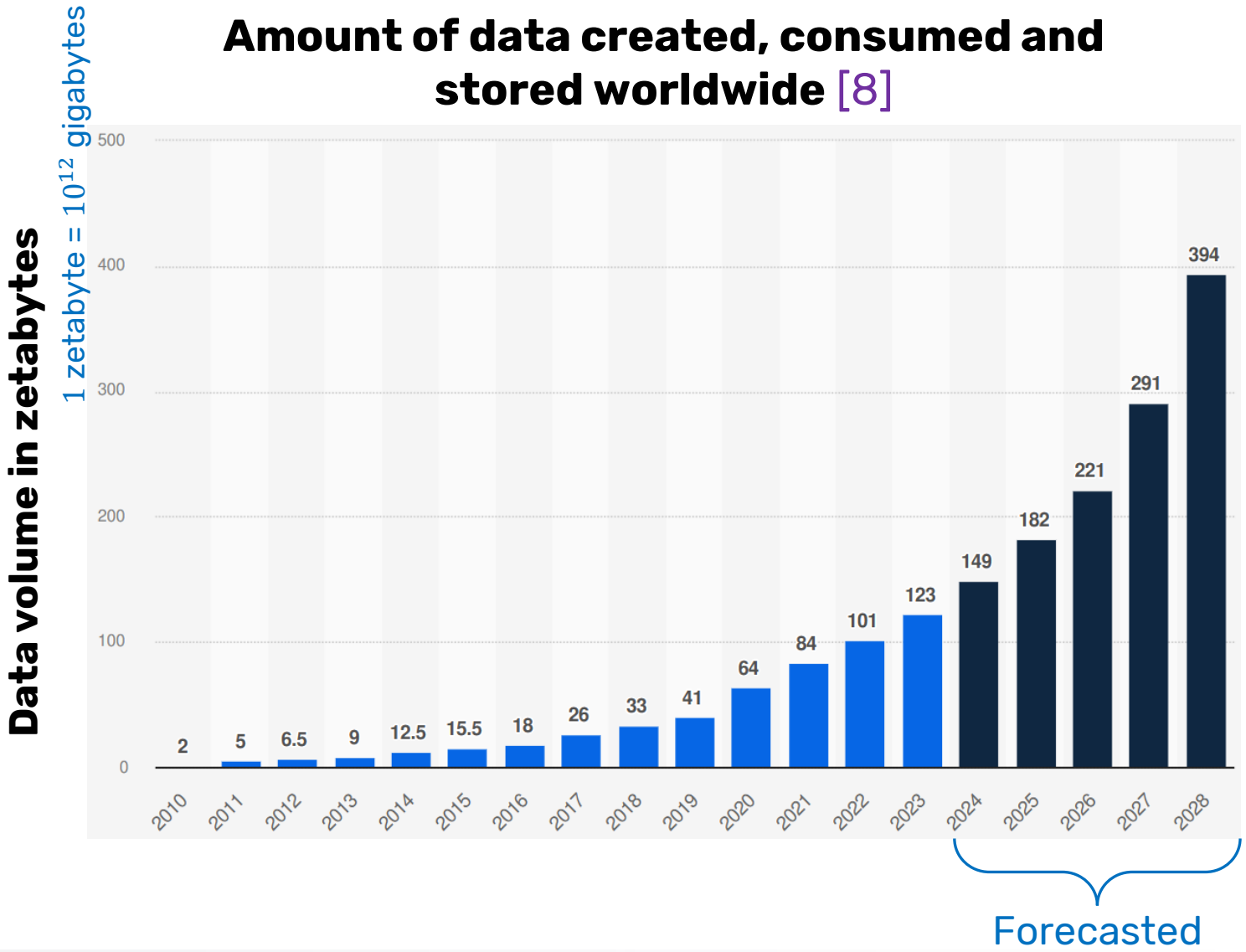
1. Course introduction
2. Data science and the data-driven company
3. **Data and its types**
4. What we are going to do with data (supervised and unsupervised learning)
5. Static and dynamical models in supervised learning
6. From business problems to data science tasks
7. The data mining life cycle (CRISP-DM)

What are data?

We refer to **data** as any piece of information that has been collected and stored in a computer

Examples:

- Sensor measurements
- Customer information
- Transaction history
- Social media posts
- …



Types of data: structured vs unstructured

Structured data

Data that are organized following a predefined scheme and stored in tabular formats (excel sheets, SQL databases…)

House area [feet^2]	**# bedrooms**	**Price** [k$]
523	1	115
645	1	150
708	2	210
⋮	⋮	⋮

Unstructured data

Data that can have an internal structure but do not follow a predefined data model or scheme

Audio files

Text files

Video files

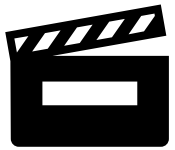


Image files

Types of data: quantitative vs qualitative

Nominal qualitative data
cannot be ordered

Ordinal qualitative data
can be ordered. Other examples: low/high income, age ranges…

Runner name	Sex	Placement	Time [seconds]
Orlando Dillon	M	First	14.75
Izabella Kent	F	Second	15.01
Sophia Sanders	F	Third	15.33
⋮	⋮	⋮	⋮

Qualitative (or categorical) data
assume non-numerical values, typically belonging to pre-defined categories

Quantitative (or continuous) data
assume numerical values

Data are dirty

Common data problems:

- Missing values
- Unlikely values (outliers)
- Inconsistent formats
- …

House area [feet^2]	# bedrooms	Completion date	Price [k$]
523	1	23/06/1998	115
645	1	01/07/2000	0.001
708	unknown	19/01/1980	210
1034	3	31-Jan-2001	unknown
unknown	4	17/12/2005	355
2545	unknown	14/02/1999	440
⋮	⋮	⋮	⋮

Typically, data must be cleaned before usage (**data cleaning**)

Outline

1. Course introduction
2. Data science and the data-driven company
3. Data and its types
4. **What we are going to do with data (supervised and unsupervised learning)**
5. Static and dynamical models in supervised learning
6. From business problems to data science tasks
7. The data mining life cycle (CRISP-DM)

What are we going to do with data?

In this course, we will use data for:

- **Descriptive analysis** and **visualization**

  ⇨ **LECTURE 02**

- **Supervised learning** (in particular, regression and classification)

  ⇨ **LECTURE 05-12**

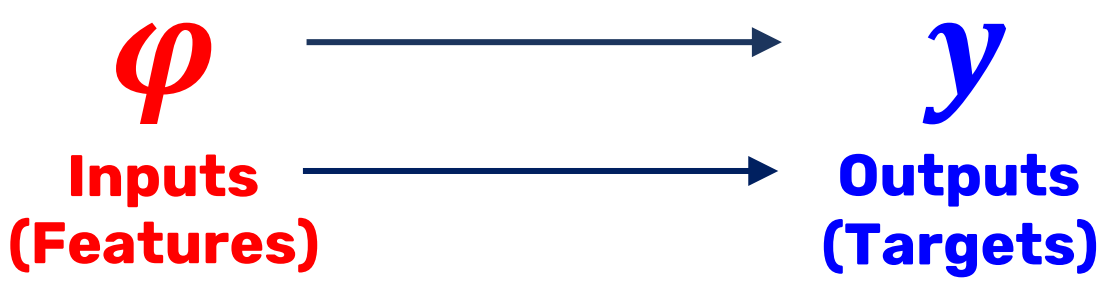
- **Unsupervised learning** (in particular, clustering and dimensionality reduction)

  ⇨ **LECTURE 13,14**

Supervised vs unsupervised learning

Many data science tasks can be tackled either by supervised or unsupervised learning methods

- **Supervised learning**: predict the values of one or more **dependent variables** (**output(s)**) based on the values of one or more **independent variables** (**input(s)**)



Typically, we will focus on supervised learning problems with **only one output**

- **Unsupervised learning**: there are **no outputs**! The goal may be to discover groups of similar entities within the data or to project the data from a high-dimensional space (#**inputs** > 3) down to two or three dimensions for the purpose of visualization

Data science tasks

- **Regression***: predict the values assumed by the **continuous** **output(s)** from the **input(s)**

Example:
- Predict the **prices** of houses based on their **area**
- Predict the **prices** of houses based on their **area** and **number of bedrooms**

House area [feet^2]	**# bedrooms**	**Price** [k$]
523	1	115
645	1	150
708	2	210
⋮	⋮	⋮

$\varphi \in \mathbb{R}$

$y \in \mathbb{R}$

$\boldsymbol{\varphi} \in \mathbb{R}^{2\times 1}$

*: covered in this course : supervised : unsupervised

Data science tasks

- **Classification***: predict the values assumed by the **categorical** **output(s)** from the **input(s)**

Example: ➢ Develop an application that recognizes cats in **images**

Image	Label
	Cat
	Not cat
	Cat
	Not cat

Input: an image

- W = width
- H = height
- D = depth

$$\varphi = \text{  } \in \mathbb{N}^{W \times H \times D}$$

Images are basically matrices of numbers that describe color intensity

Output: the class label

$$y \in \{\text{Cat}, \text{Not cat}\}$$

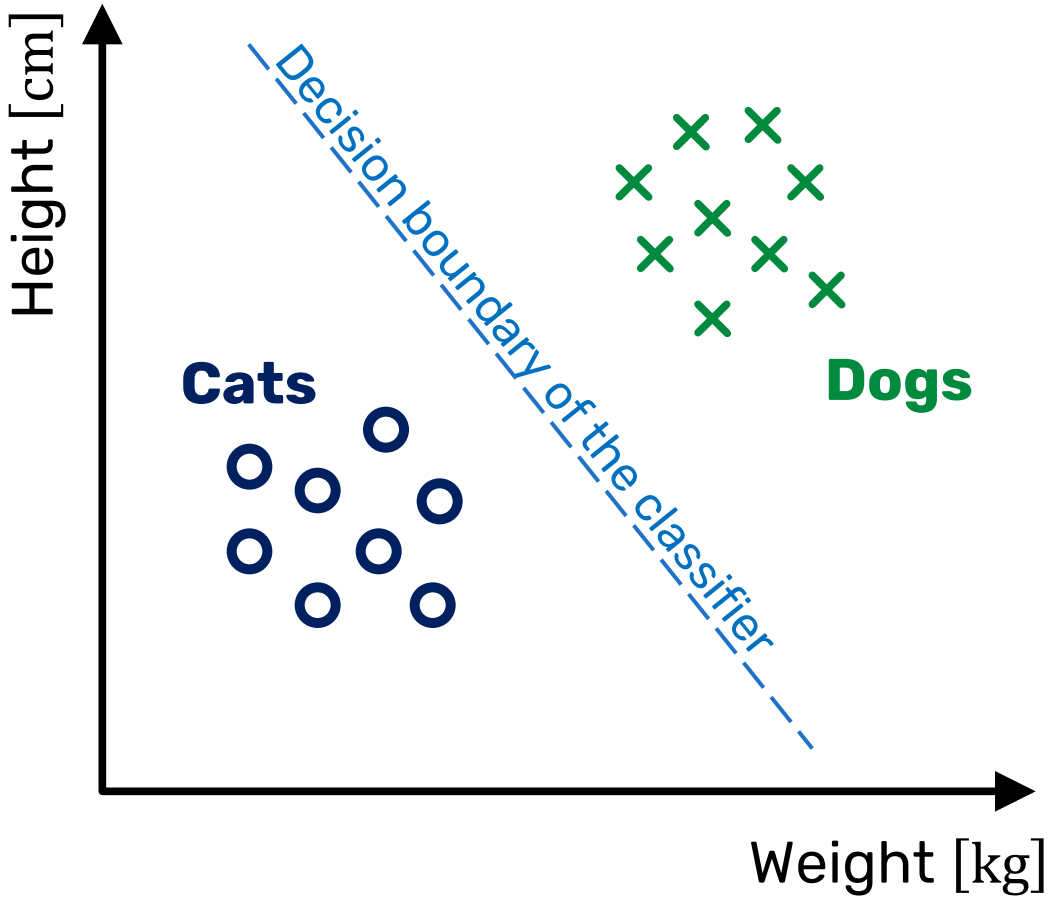
(single output)

*: covered in this course : supervised : unsupervised

Data science tasks

- **Classification***: predict the values assumed by the **categorical** **output(s)** from the **input(s)**

Example: ➢ Distinguish cats from dogs based on their **height** and **weight**



$$\boldsymbol{\varphi} \in \mathbb{R}^{2\times 1}$$

(height and weight of the animal)

Output: the class label

$$y \in \{\text{cat}, \text{dog}\}$$

(single output)

*: covered in this course : supervised : unsupervised

Data science tasks

- **Causal modeling**: identify which **inputs** (**causes**) actually influence the **outputs** (**effects**) and, possibly, to what extent

  Example: ➢ Did a particular marketing campaign influence the consumers to purchase our product?

  Causal modeling typically involves substantial investments in data, such as randomized controlled experiments (**A/B tests**) and sophisticated methods for drawing causal observation data (**"counterfactual" analysis**)

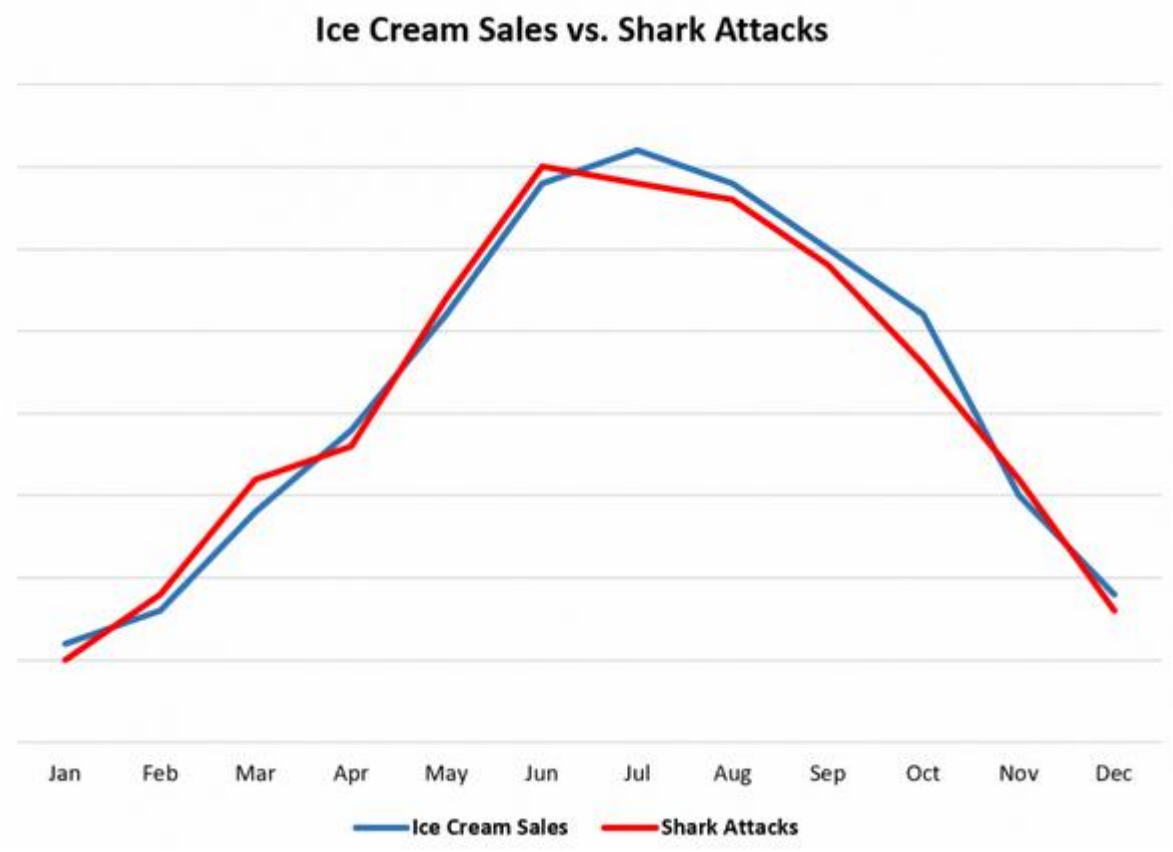
  What would be the difference in sales if we used an advertisement instead of another?

  Technical note: regression and classification are based on correlation, causal modeling is based on causality

*: covered in this course : supervised : unsupervised

Data science tasks

- **Causal modeling**: identify which **inputs** (**causes**) actually influence the **outputs** (**effects**) and, possibly, to what extent



Picture taken from [9]

Correlation does not imply causation!

If we take a look at the data representing monthly ice cream sales and monthly shark attacks around the United States each year, we can see that the two variables are highly correlated

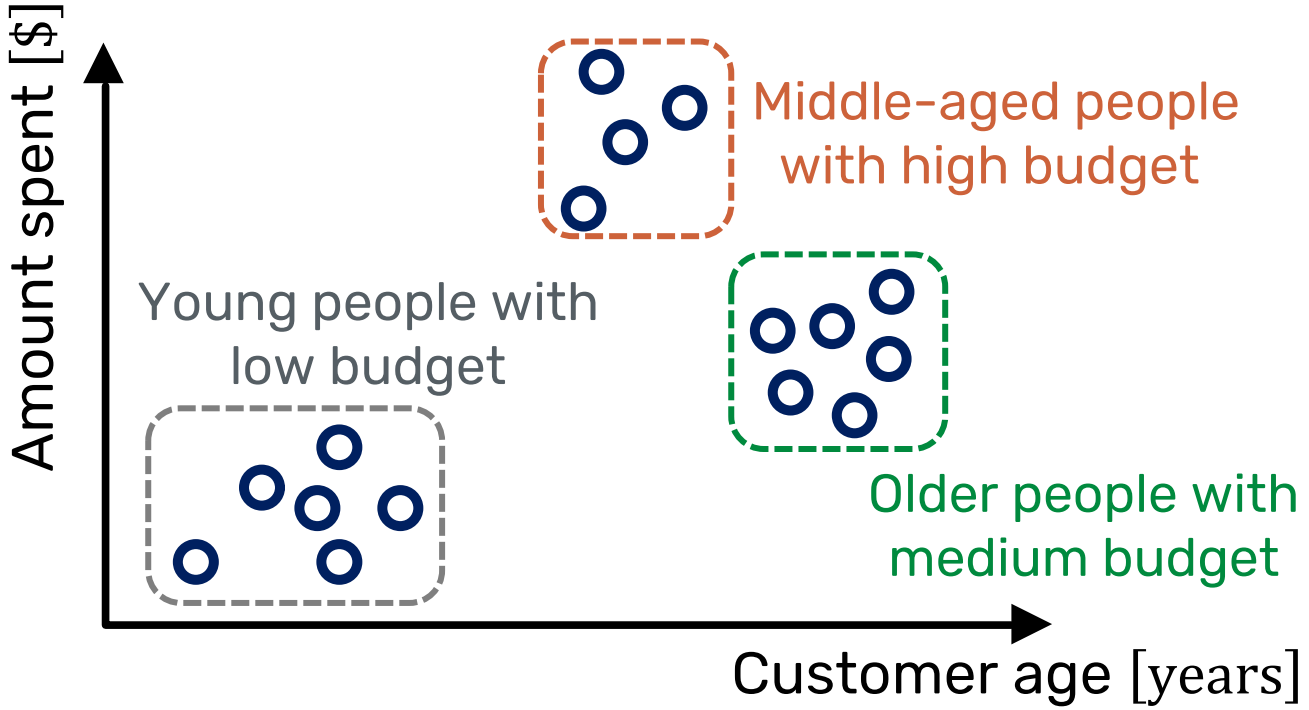
Does this mean that consuming ice cream causes shark attacks? No! The more likely explanation is that more people consume ice cream and get in the ocean when it's warmer outside, explaining the high correlation

*: covered in this course : supervised : unsupervised

Data science tasks

- **Clustering***: organize the data into different groups based on their similarity

Example: ➢ Understand which types of customers are similar to each other by grouping individuals according to several **characteristics** → personalized marketing campaigns



$\varphi \in \mathbb{R}^{2\times 1}$
(customer age and amount spent)

Output: none

*: covered in this course : supervised : unsupervised

Data science tasks

- **Co-occurrence grouping**: find associations between different entities (characterized by a set of **features**) based on transactions involving them

  Example: ➢ What items are commonly purchased together? (**market basket analysis**)

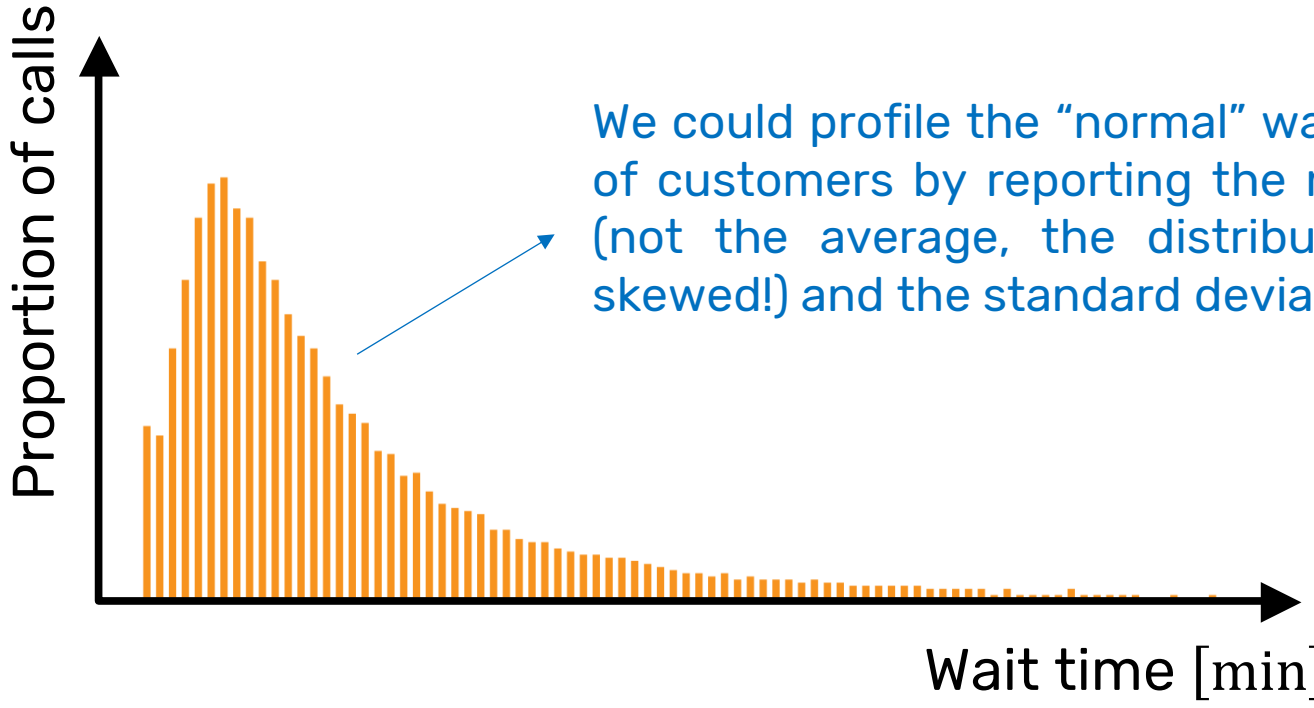
Clustering looks at the similarity between entities based on their features, co-occurrence grouping considers the similarity of entities based on their appearing together in transactions (e.g., "a keyboard is not similar to a mouse, although they are typically bought together")

*: covered in this course : supervised : unsupervised

Data science tasks

- **Profiling**: find the typical behavior of an individual, group or population

Example:
- What is the typical credit card usage of a customer segment?
- Profile the typical wait time of customers who call into a call center



We could profile the "normal" wait time of customers by reporting the median (not the average, the distribution is skewed!) and the standard deviation

$\varphi \in \mathbb{R}$
(wait time)

Output: none

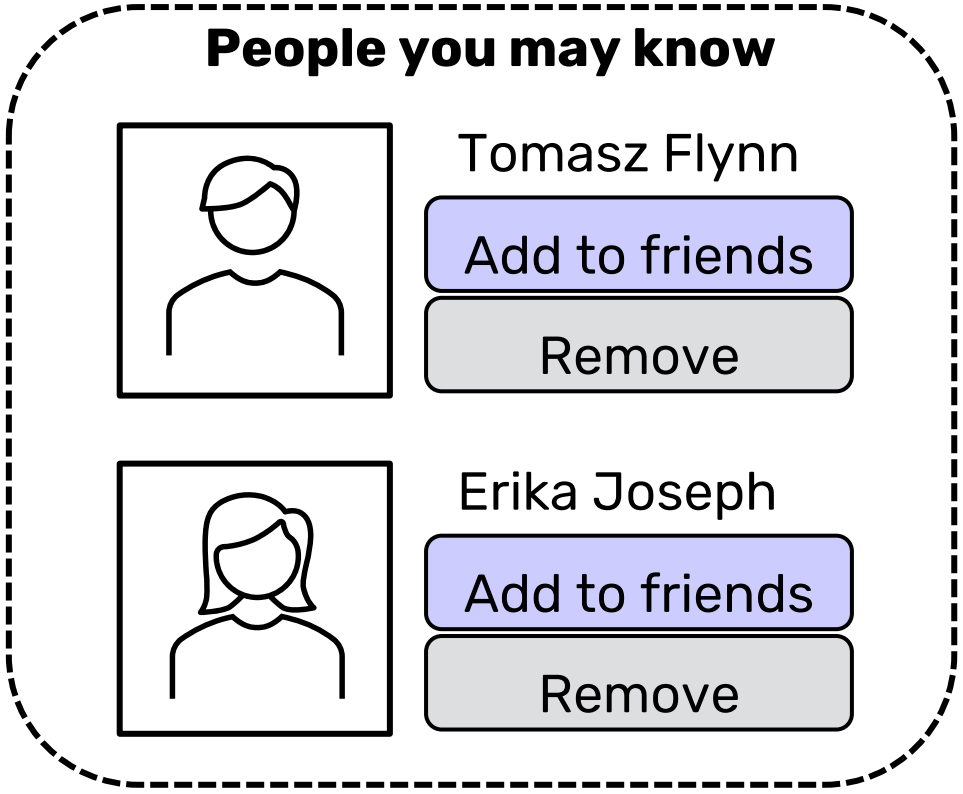
Picture taken from [1]

*: covered in this course : supervised : unsupervised

Data science tasks

- **Link prediction**: predict connections between entities in a network, usually by suggesting that a link should exist, and possibly also estimating the strength of the link

Example: ➢ Friend recommendations in social networks

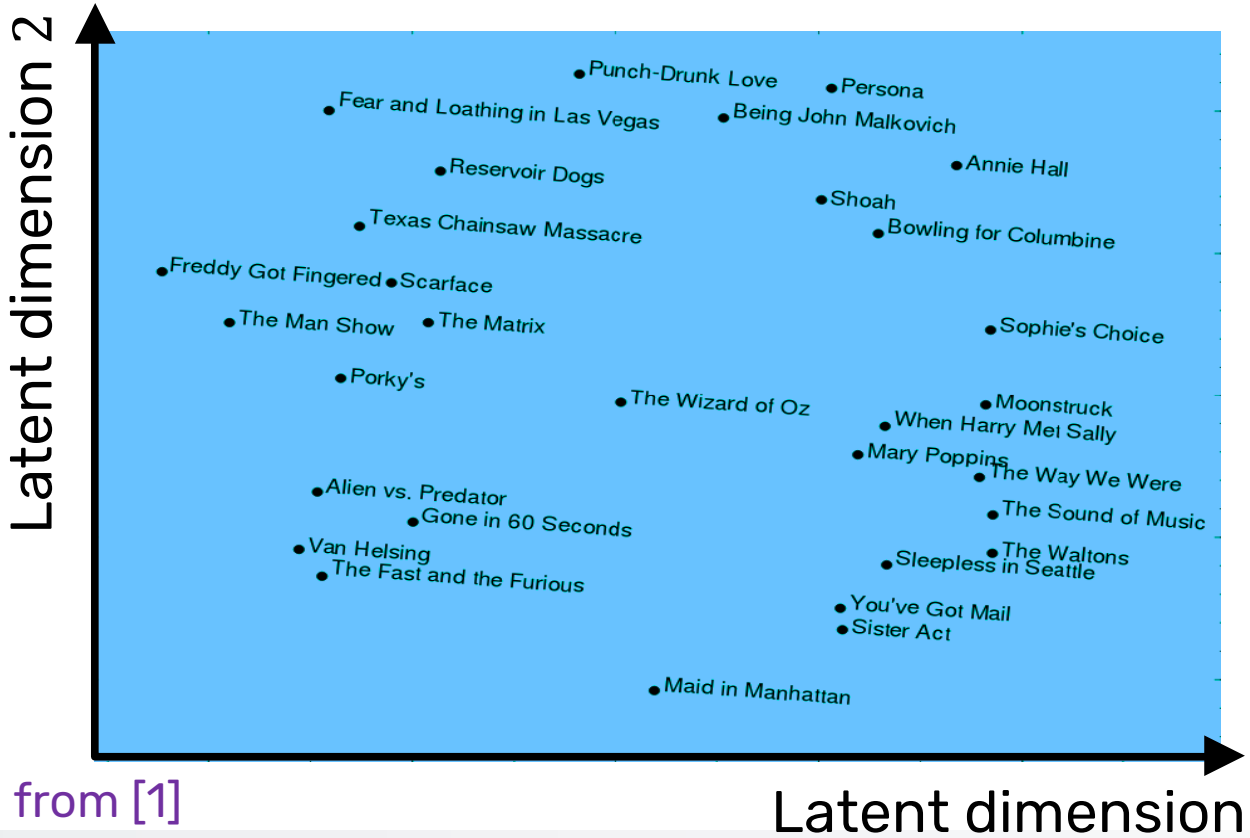


*: covered in this course : supervised : unsupervised

Data science tasks

- **Dimensionality reduction***: take a large dataset (many **inputs** and, possibly, many **outputs**) and replace it with a smaller dataset, retaining as much information as possible

Example: ➢ Represent a collection of movies in a two-dimensional space (Netflix Prize)



Picture taken from [1]

Inputs:

- Movie title
- Year of release
- User id
- User rating
- Rating date

Output: none (in this example)

*: covered in this course : supervised : unsupervised

Data science tasks

- **Similarity matching**: find similar entities based on data known about them

  Example: ➢ Recommendation systems

Inputs:

- Song titles
- Song genres
- Audio signals
- ⋮
- User ratings
- ⋮

Clustering is used for exploratory data analysis ("can we partition the data into different groups of similar entities?"), similarity matching has the specific goal of finding similar entities

Output: none (in this example)

*: covered in this course : supervised : unsupervised

Data science tasks vs methods

Data science task

(the problem that we are trying to solve, what we are trying to do)

Regression, classification, …

$\neq$

Method (or algorithm)

(how we solve it, a sequence of operations to follow)

Neural networks, K-nearest neighbors, K-means clustering, …

- Different data science tasks can be solved by the same methods

  K-means clustering can be used both for clustering and similarity matching

- Different methods can solve the same data science task

  A regression problem can be solved by the linear regression method, neural networks and K-nearest neighbors

In this course, **we will study methods for solving different data science tasks**

Syllabus

1. **Introduction to data science**
2. **Exploratory data analysis**
3. **Recap of statistics**
4. **Maximum likelihood estimation**
5. **Linear regression** (regression)
6. **Logistic regression** (classification)
7. **Bias-variance trade-off**
8. **Overfitting and regularization**
9. **Validation and cross-validation**
10. **Decision trees** (regression and classification)
11. **Neural networks** (regression, classification, dimensionality reduction…)
12. **Convolutional neural networks** (regression, classification, …)
13. **Clustering methods** (clustering)
14. **Principal component analysis** (dimensionality reduction)

Bonus

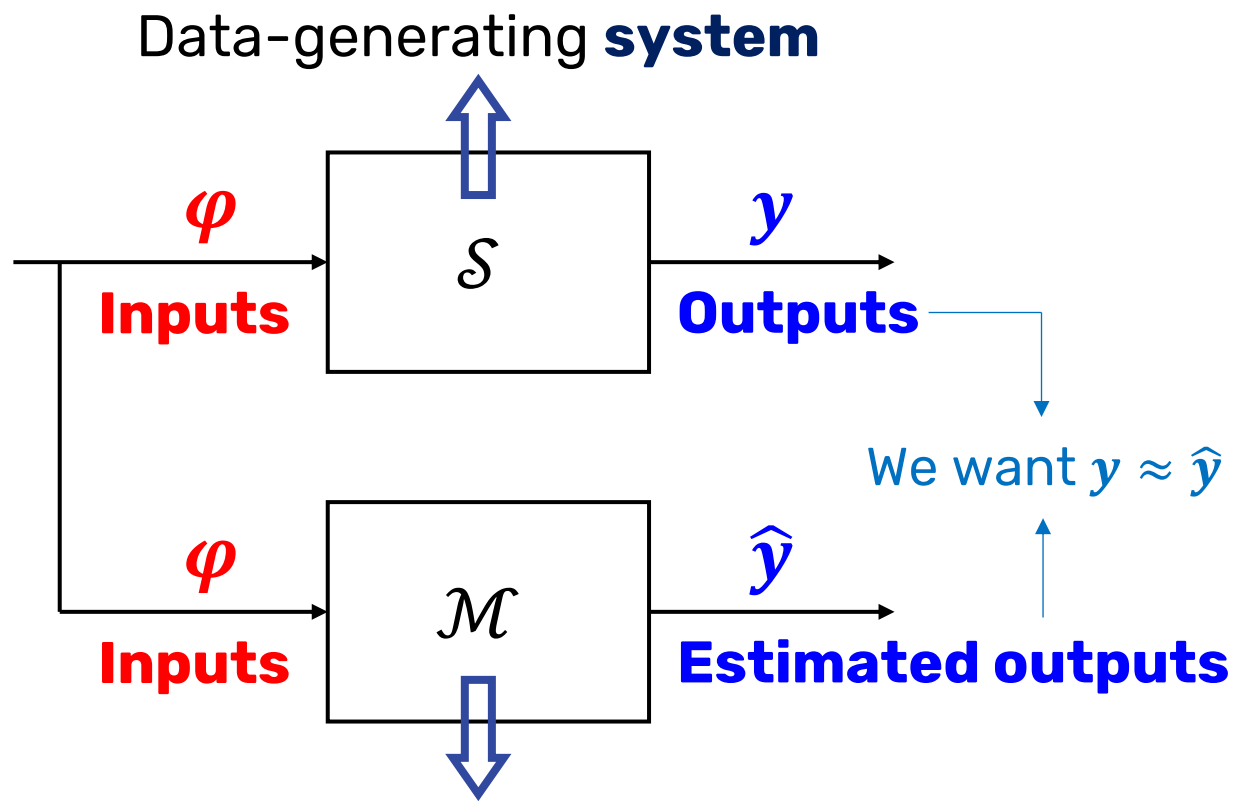
: supervised : unsupervised

Outline

1. Course introduction
2. Data science and the data-driven company
3. Data and its types
4. What we are going to do with data (supervised and unsupervised learning)
5. **Static and dynamical models in supervised learning**
6. From business problems to data science tasks
7. The data mining life cycle (CRISP-DM)

Models in supervised learning

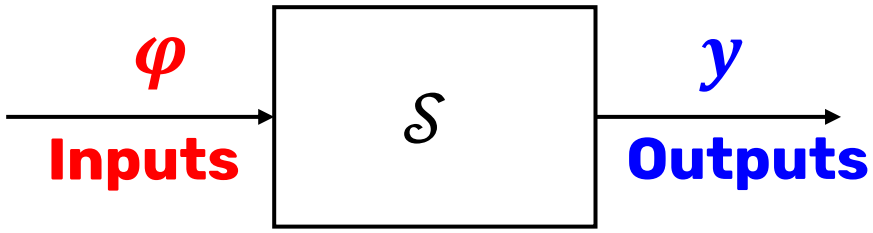
Most supervised learning methods rely on mathematical **models** that describe the relationship between the **inputs** and the **outputs**



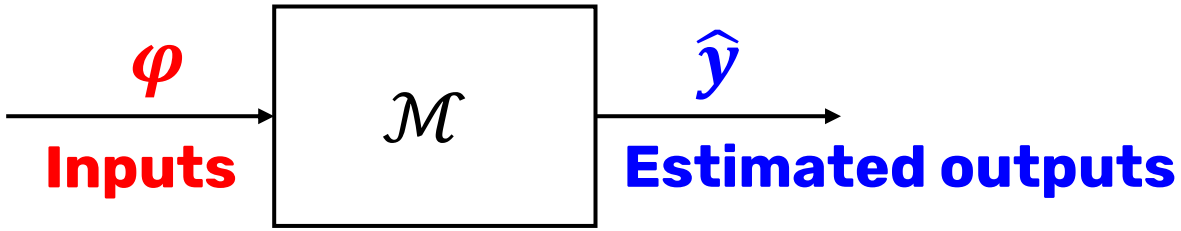
Supervised learning methods estimate $\mathcal{M}$ from data

Models in supervised learning

We view both $\mathcal{S}$ and $\mathcal{M}$ as mathematical functions that map **inputs** (**features**) to **outputs** (**targets**)



$\equiv \quad y = f(\varphi)$

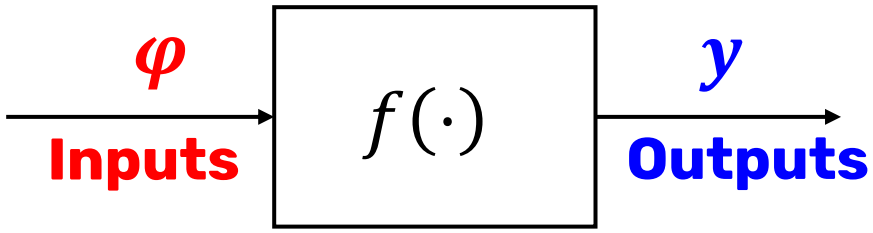


$\equiv \quad \hat{y} = \hat{f}(\varphi)$

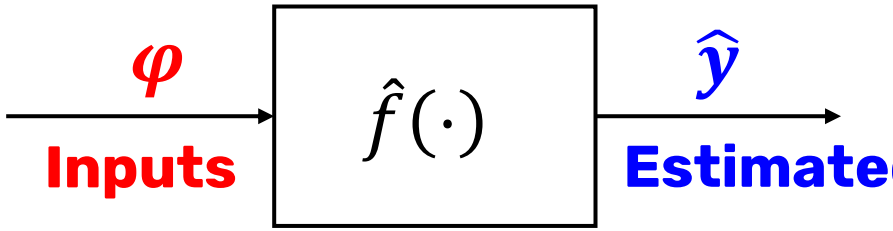
The goal of supervised learning methods is to learn a function $\hat{f}(\cdot)$ that approximates $f(\cdot)$ well **on the whole domain** of φ

Models in supervised learning

We view both $\mathcal{S}$ and $\mathcal{M}$ as mathematical functions that map **inputs** (**features**) to **outputs** (**targets**)



$\equiv \quad y = f(\varphi)$



$\equiv \quad \hat{y} = \hat{f}(\varphi)$

The goal of supervised learning methods is to learn a function $\hat{f}(\cdot)$ that approximates $f(\cdot)$ well **on the whole domain** of φ

Dataset notation

Before moving on, we introduce the following notation that we will use for any dataset

House area [feet^2]	# bedrooms	Price [k$]
⋮	⋮	⋮
523	1	115
645	1	150
708	2	210
⋮	⋮	⋮

$$\boldsymbol{\varphi}(i) = \begin{bmatrix} 523 \\ 1 \end{bmatrix} \qquad y(i) = 115$$

We refer to each row of the dataset as an **observation**

i-th observation (in this case it represents a house but, in general, it can be any entity)

$$\left(\boldsymbol{\varphi}(i), y(i)\right)$$

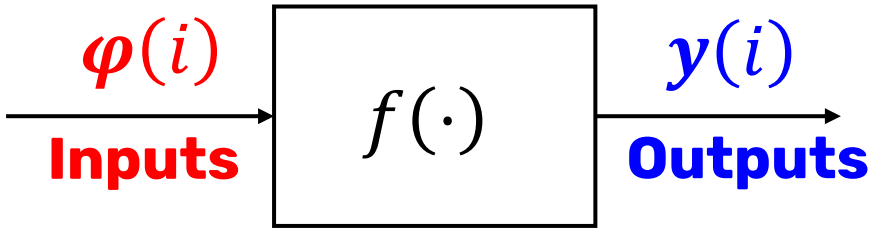
We denote the dataset as $\mathcal{D} = \left\{\left(\boldsymbol{\varphi}(1), y(1)\right), \dots, \left(\boldsymbol{\varphi}(N), y(N)\right)\right\}$

$$= \left\{\left(\boldsymbol{\varphi}(i), y(i)\right)\right\}_{i=1}^{N}$$

(N observations in total)

Static systems (and models)

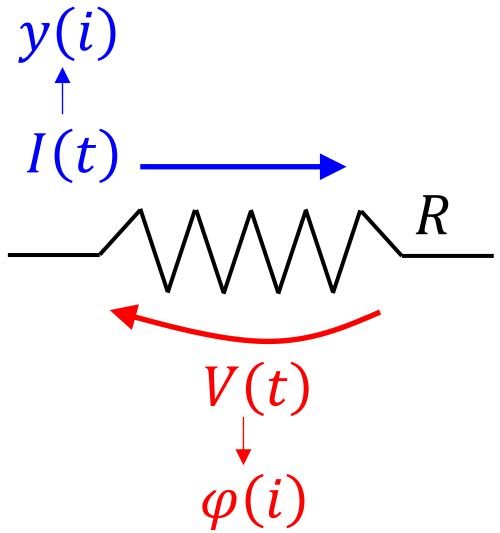
A system whose **outputs** can be determined directly from the **inputs** is said to be a **static system** ("memoryless" system)



Example: Ohm's law

$$I(t) = \frac{V(t)}{R}$$

$$f(\varphi(i))$$



The output $I(t)$ at time t only depends on the input $V(t)$ at the same time instant

We can view each voltage/current measurement by itself (i.e. as an observation $(\varphi(i), y(i))$ in its own right), we do not need to consider $V(t)$ and $I(t)$ as signals

"The time t can be omitted"

Static systems (and models)

Static systems need **not** describe **only** physics phenomena

House area [feet^2]	# bedrooms	Price [k$]
523	1	115
645	1	150
708	2	210
⋮	⋮	⋮

$f(\cdot)$: mapping from house area and # bedrooms to price

Image	Label
	Cat
	Not cat
	Cat
	Not cat

$f(\cdot)$: mapping from image to label

Learning static systems

In the **regression** setting, the simplest model that can be used to describe static systems is the **linear model**

$$\underset{1\times 1}{y(i)} = \theta_0 + \theta_1\varphi_1(i) + \cdots + \theta_{d-1}\varphi_{d-1}(i) + \epsilon(i) = \sum_{j=0}^{d-1}\theta_j\,\varphi_j(i) + \epsilon(i)$$

$$= \underset{1\times d}{\boldsymbol{\varphi}(i)^\top}\underset{d\times 1}{\boldsymbol{\theta}} + \underset{1\times 1}{\epsilon(i)}$$

$[\ \cdots\ ]$ $\begin{bmatrix}\vdots\end{bmatrix}$

i –th observation

- $\varphi_0 = 1$
- $\boldsymbol{\varphi}(i) = [\varphi_0 \quad \varphi_1(i) \quad \cdots \quad \varphi_{d-1}(i)]^\top \in \mathbb{R}^{d\times 1}$
- $\boldsymbol{\theta} = [\theta_0 \quad \theta_1 \quad \cdots \quad \theta_{d-1}]^\top \in \mathbb{R}^{d\times 1}$
- $y(i) \in \mathbb{R}$

- The vector $\boldsymbol{\theta}$ is called **parameters vector** → to be found by minimizing a cost function
- The vector $\boldsymbol{\varphi}(i)$ is called **features vector** for the i-th observation → attributes of entities
- The quantity $\epsilon(i)$ is the **error** due to not perfect explanation of $y(i)$ using $\boldsymbol{\varphi}(i)$

Learning static systems

To **"learn"** means to **estimate the values** of the parameters in $\boldsymbol{\theta} = [\theta_0 \quad \theta_1 \quad \cdots \quad \theta_{d-1}]^\top$

Key idea: find the values of $\boldsymbol{\theta}$ that **minimize** a "cost" (or "loss"), i.e., an "error" or "something bad" → it is good to minimize something bad

- This is achieved through **optimization**

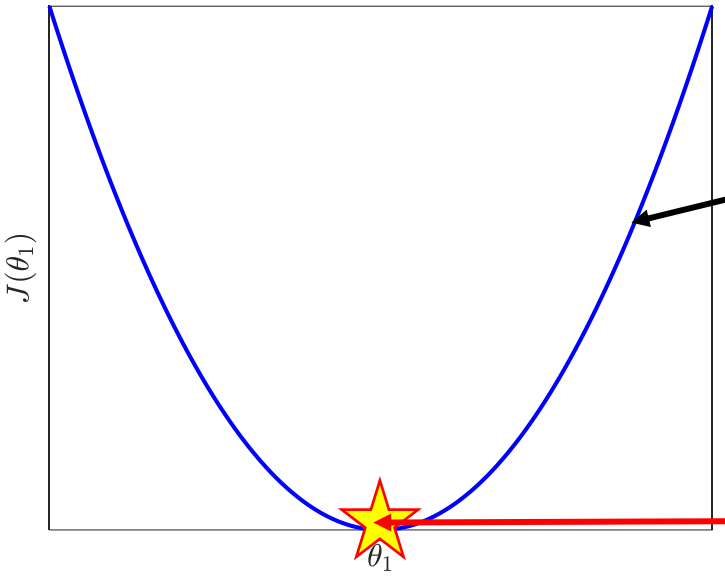
A typical cost in the regression setting is the following

$$J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(y(i) - \boldsymbol{\varphi}(i)^\top\boldsymbol{\theta})^2 = \frac{1}{N}\sum_{i=1}^{N}\epsilon(i)^2$$

With this cost, we are **minimizing the sum of the squared errors** between the observed outputs (i.e. those reported in our dataset) and the outputs estimated by the linear model

Learning static systems

Scalar (single) parameter θ



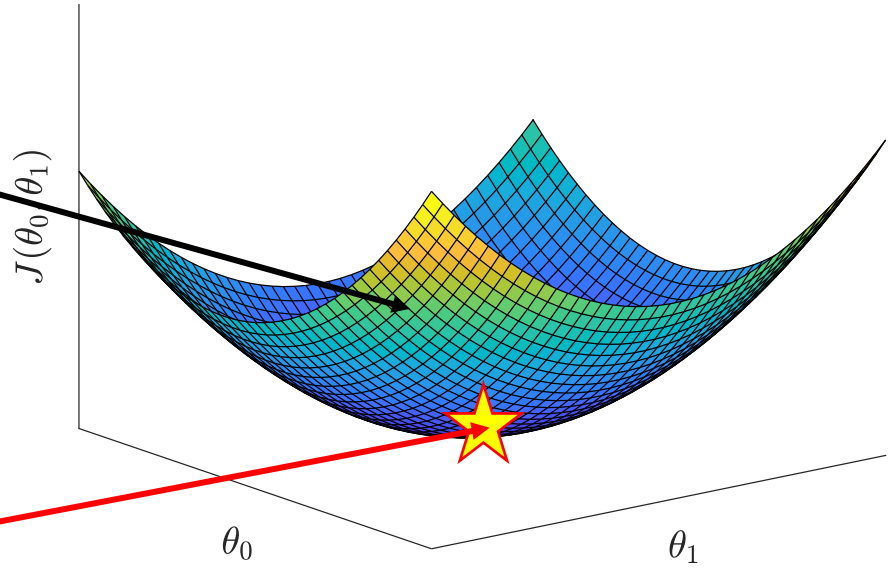
Cost function

$$J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \epsilon(i)^2$$

Minimizer of the cost function:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

Multiple parameters θ

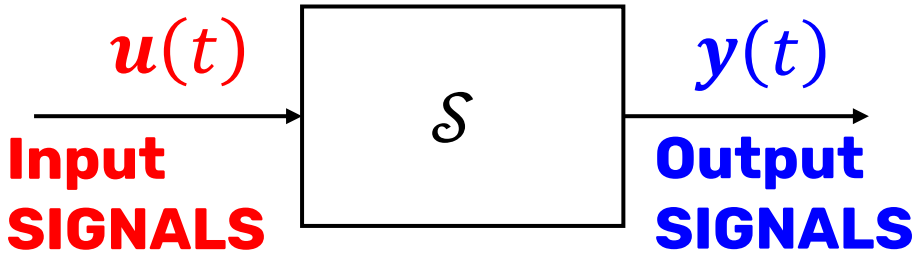


This rationale is followed by the **linear regression method**

$$\hat{y}(i) = \hat{f}\left(\boldsymbol{\varphi}(i)\right) = \boldsymbol{\varphi}(i)^{\mathsf{T}}\widehat{\boldsymbol{\theta}}$$

Dynamical systems (and models)

A system whose **outputs** (at a certain time instant) cannot be determined directly from the **inputs** (at the same time instant) is said to be a **dynamical system**



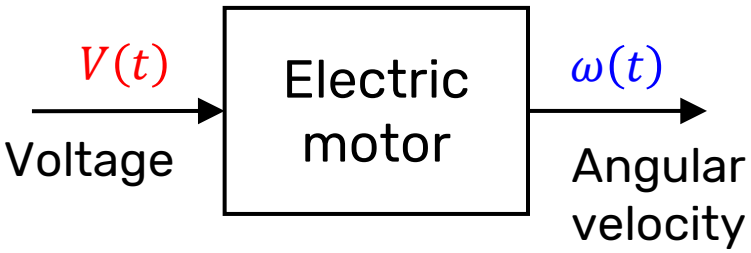
Dynamical models are mathematical models that describe the future evolution of the variables involved as a **function of their past trends**

Dynamical systems usually involve the **time**: the **outputs** $y(t)$ at a certain time t **depend on the outputs at previous times**

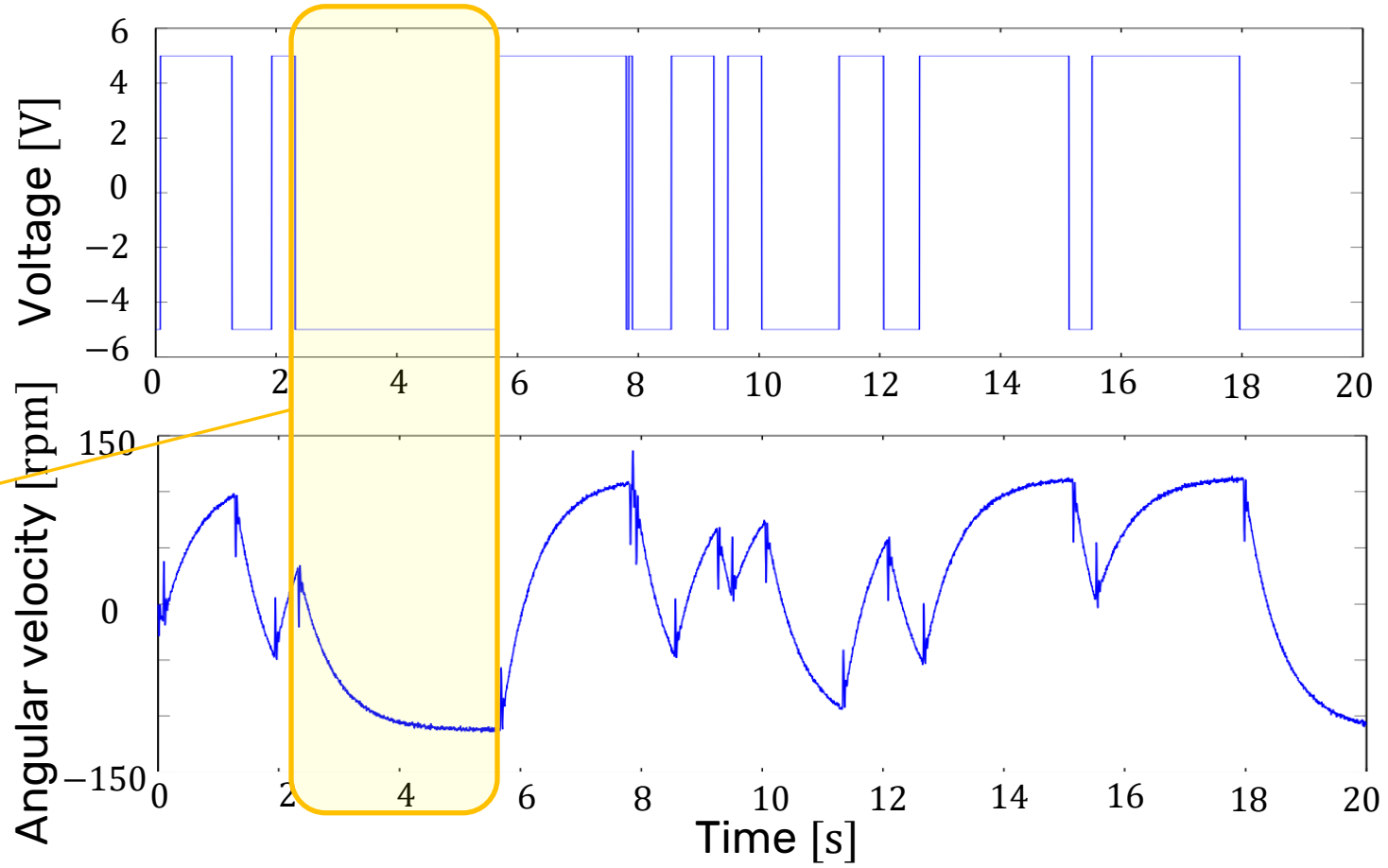
This dependency on the past endows the model with a **"memory"** (i.e. the dynamics)

Dynamical systems (and models)

This dependency on the past endows the model with a **"memory"** (i.e. the dynamics)



We are dealing with a dynamical system because, although **the input is constant, the output keeps evolving**

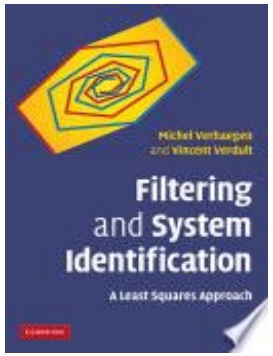


Learning dynamical systems

- In the control systems community, the concept of learning dynamical systems is typically referred to as **system identification**
- In any case, **system identification belongs to the supervised learning framework**:
  - Instead of dealing with datasets of observations, each observation representing an entity (e.g., a house), we have at our disposal datasets of signals (e.g., the electric motor's voltage and current signals). In this context, an observation is the collection of measurements of signals of interest at a certain time
  - Models used to describe dynamical systems must embed the time dependency in some way (e.g., we can employ transfer function models, state-space models, …)

Learning dynamical systems

- System identification methods are **not covered** in this course but are based on the concepts that we are going to learn in the following lectures
- If you are interested in system identification, these UniBg courses are what you are looking for:
  - [38003-1] - Identificazione dei modelli e analisi dei dati
  - [38095] - Adaptive learning, estimation and supervision of dynamical systems
- A good book on the topic is:

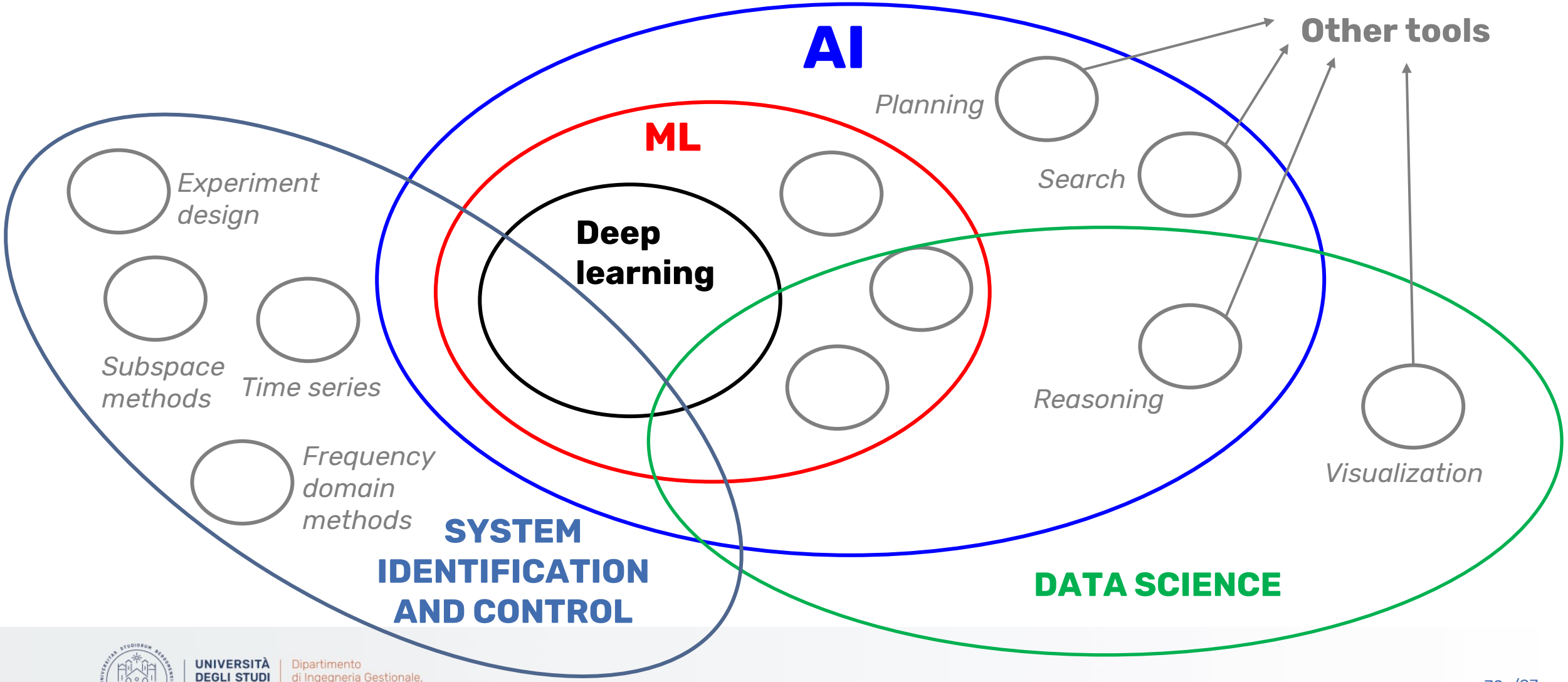


Michel Verhaegen, Vincent Verdult

Filtering and system identification: a least squares approach

Cambridge University Press (2007)

Machine Learning (ML), Artificial Intelligence (AI), Data Science and System Identification



Why do we need models?

All in all, we need a model to **better understand the phenomena** that are of our interest.

Models are useful for:

- **Decision-making:** suppose that we are testing a new vaccine. We have two groups of people. We give the vaccine to the first group (test group) and a placebo to the second one (control group). Then, we measure some variables from the patients. How can we determine if the vaccine was effective or not?

- **Communication:** a model allows to communicate to third parties the main insights and results of your analysis

Why do we need models?

All in all, we need a model to **better understand the phenomena** that are of our interest.

Models are useful for:

- **Prediction:** forecast the values that the output variables will assume based on the values assumed by the inputs variables and on which we have no data about

House area [feet^2]	# bedrooms	Price [k$]
523	1	115
645	1	150
708	2	210
⋮	⋮	⋮

How much does a 600 feet^2 house with 2 bedrooms cost?

Why do we need models?

All in all, we need a model to **better understand the phenomena** that are of our interest.

Models are useful for:

- **Inference**: understand how changes in the inputs affect the outputs

House area [feet^2]	# bedrooms	Price [k$]
523	1	115
645	1	150
708	2	210
⋮	⋮	⋮

- Does increasing house area increase the house price (and by how much)?
- Is # bedrooms actually associated with the price of a house?

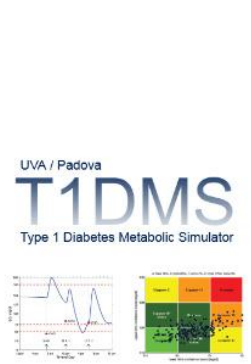
Prediction vs inference: prediction is not necessarily concerned with the structure of the model $\hat{f}(\cdot)$ and its complexity ($\hat{f}(\cdot)$ can be seen as a "black-box") while inference uses the model to understand the relationship between each input and each output

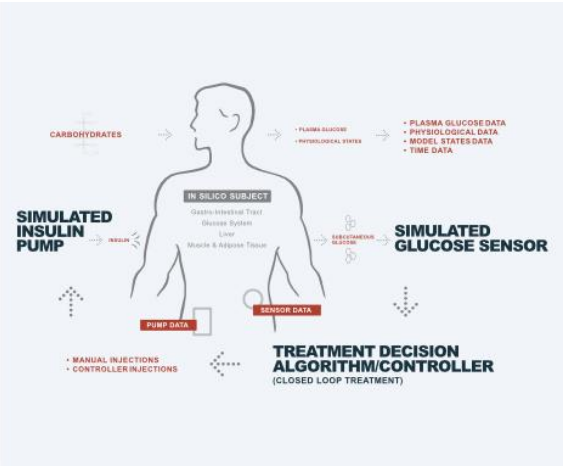
Why do we need models?

All in all, we need a model to **better understand the phenomena** that are of our interest.

Models are useful for:

- **Simulation:** we can simulate, with a computer, the response (outputs) of a model due to certain inputs. By looking at the model's response, we can get a better grasp of the modeled system



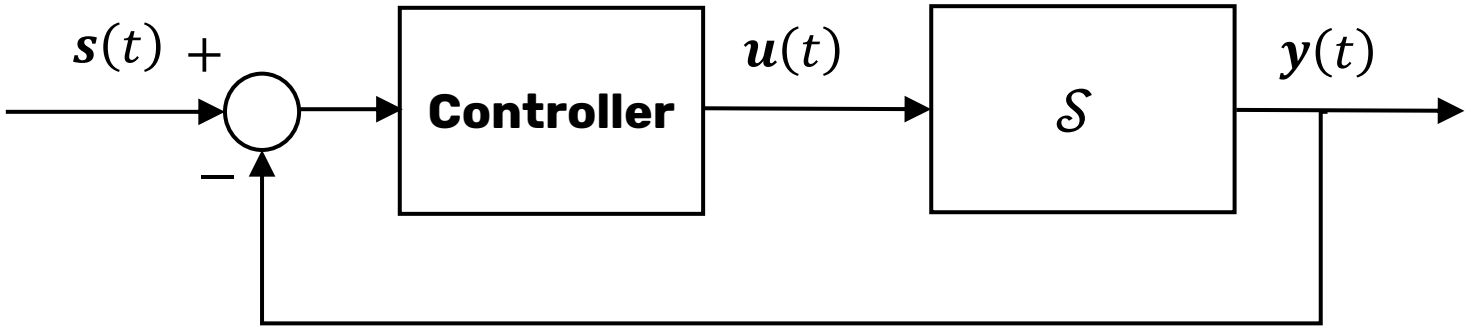


Why do we need models?

All in all, we need a model to **better understand the phenomena** that are of our interest.

Models are useful for:

- **Control:** often, in control engineering, we need a model of a system to design a controller that limits the deviation of the controlled variables $\boldsymbol{y}(t)$ from the reference variables $\boldsymbol{s}(t)$ (t represents the time)

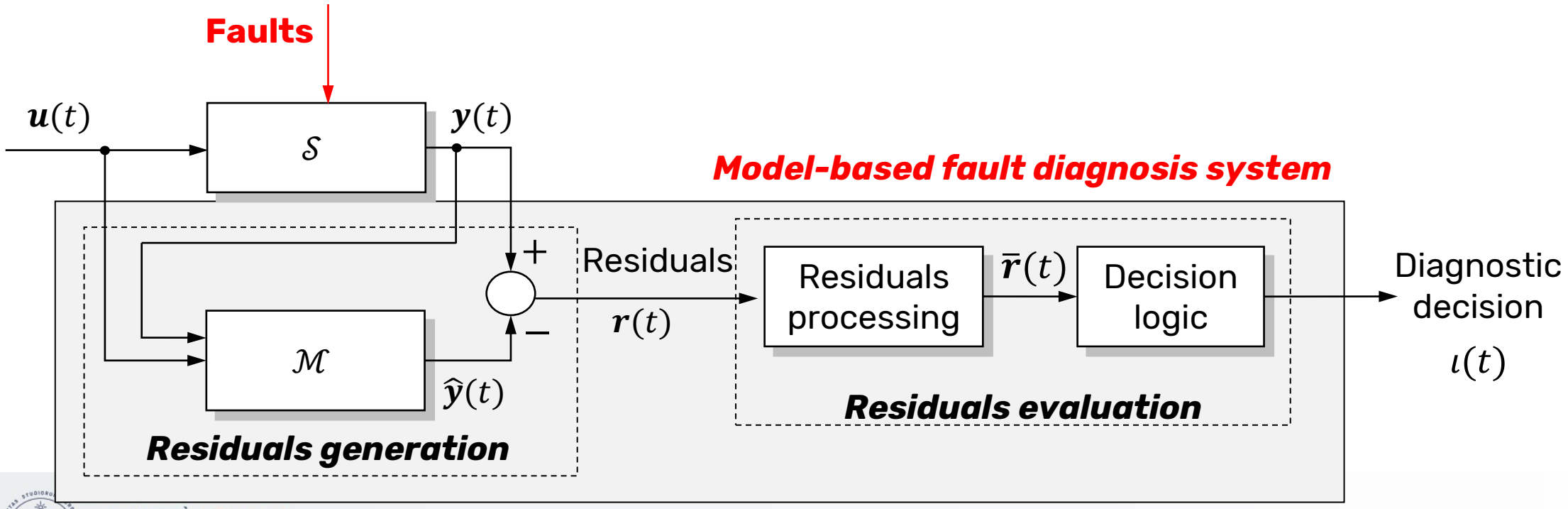


Why do we need models?

All in all, we need a model to **better understand the phenomena** that are of our interest.

Models are useful for:

- **Fault diagnosis:** we can check the presence of faults by comparing signals that come from the real system with those simulated by the estimated model

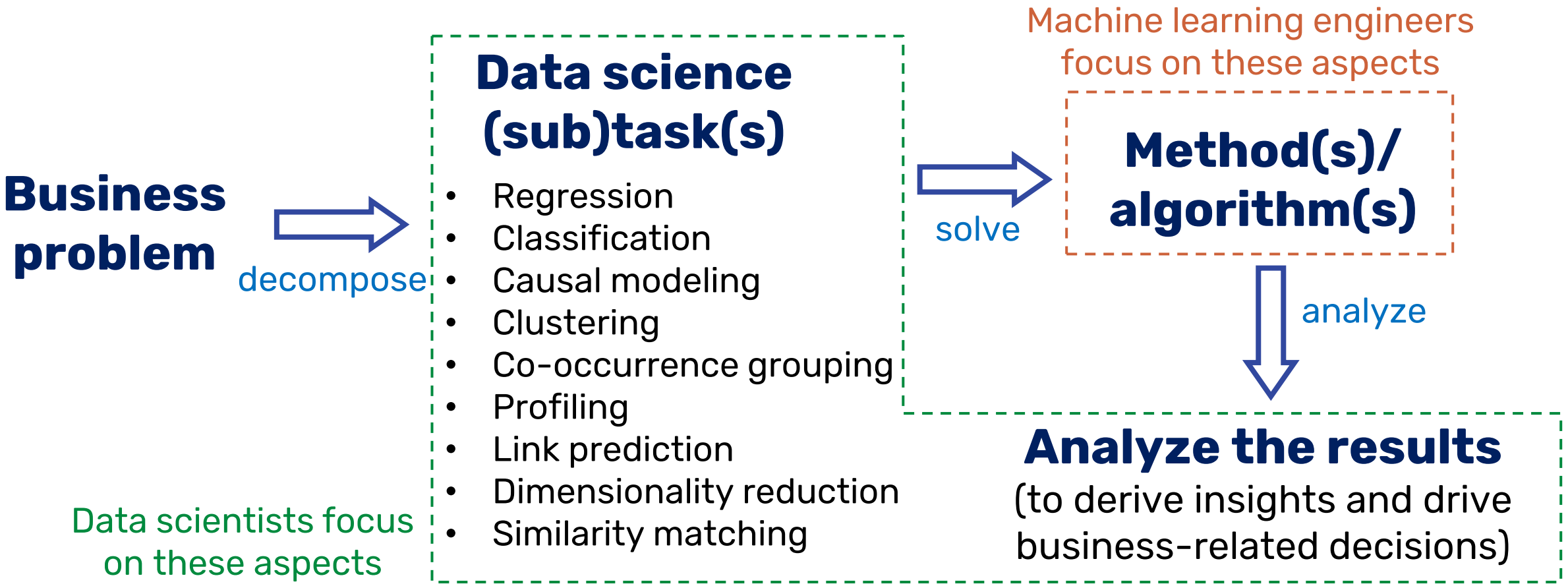


Outline

1. Course introduction
2. Data science and the data-driven company
3. Data and its types
4. What we are going to do with data (supervised and unsupervised learning)
5. Static and dynamical models in supervised learning
6. **From business problems to data science tasks**
7. The data mining life cycle (CRISP-DM)

Business problems as data science tasks

Each data-driven project is **unique**. First and foremost, **decompose** the business problem into data science subtasks that can be solved by **existing methods**



Business problems as data science tasks

- Spam e-mail detection system Classification
- Credit approval Classification
- Fraud detection Profiling
- Recognize objects in images Classification
- Find the relationship between house prices and house sizes Regression
- Predict the stock market Regression
- Market segmentation Clustering
- Market basket analysis Co-occurrence grouping
- Language models (word2vec) Similarity matching
- Social network analysis Link prediction
- Low-order data representations Dimensionality reduction
- Movies recommendation Similarity matching
- A/B testing Causal modeling

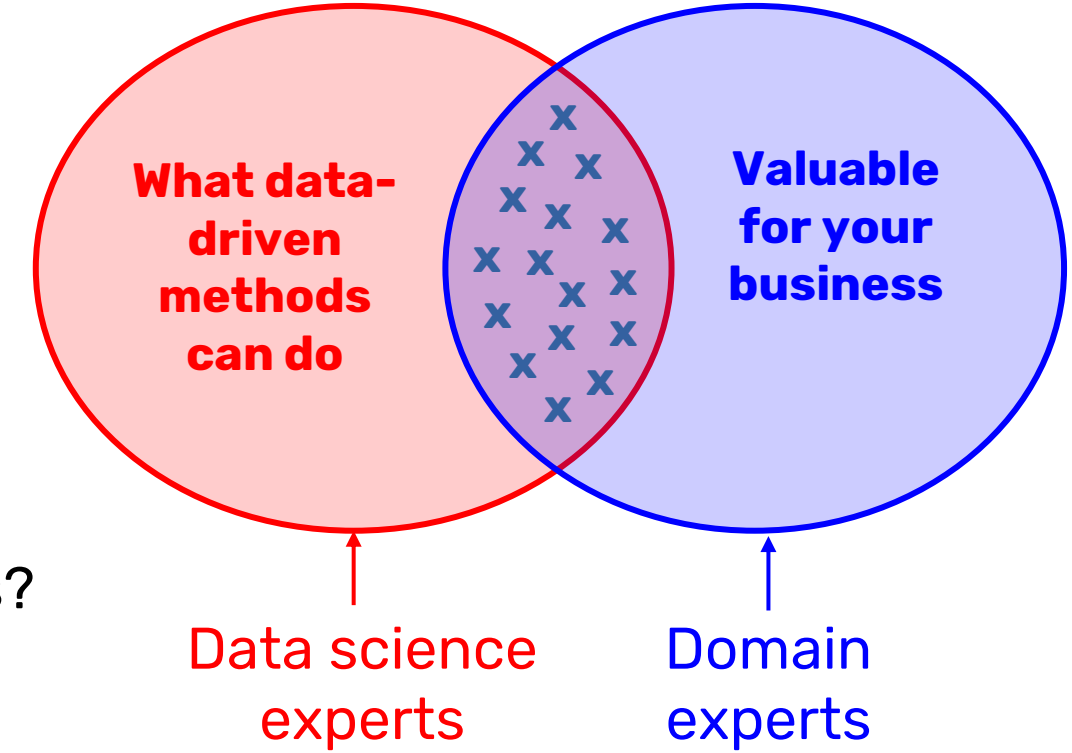
Selecting data-driven projects

Focus on data science and machine learning projects that are **valuable** and **feasible**

Think about automating **tasks** rather than automating **jobs**

What are the main **drivers** of the business values?

What are the main **pain points** in your business?



Selecting data-driven projects

MANUFACTURING LINE MANAGER

Data science

- Optimize production yield

Machine learning

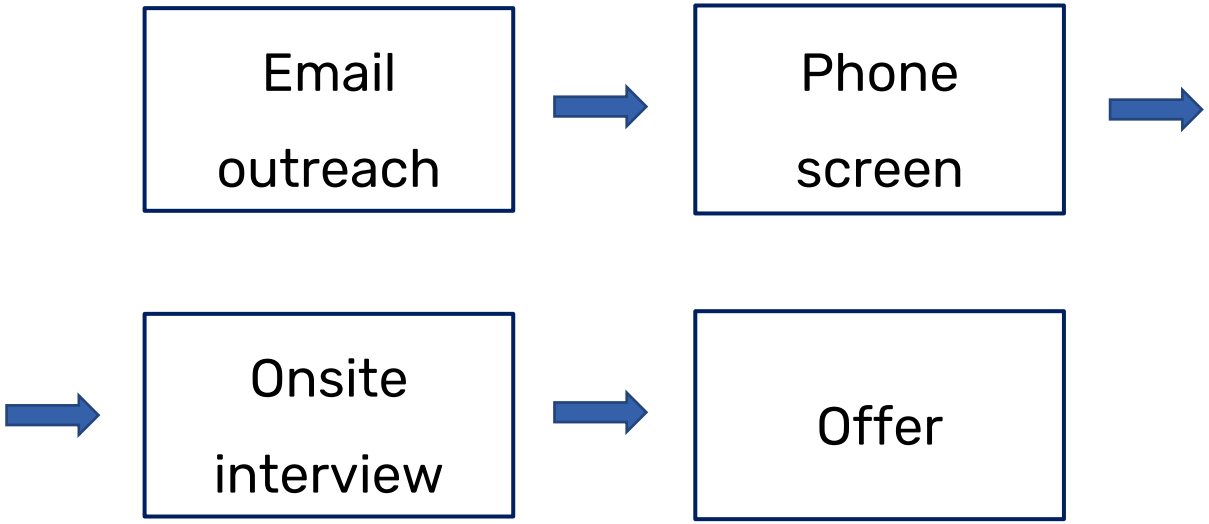
NO DEFECT NO DEFECT DEFECT

- Automatic visual inspection

Selecting data-driven projects

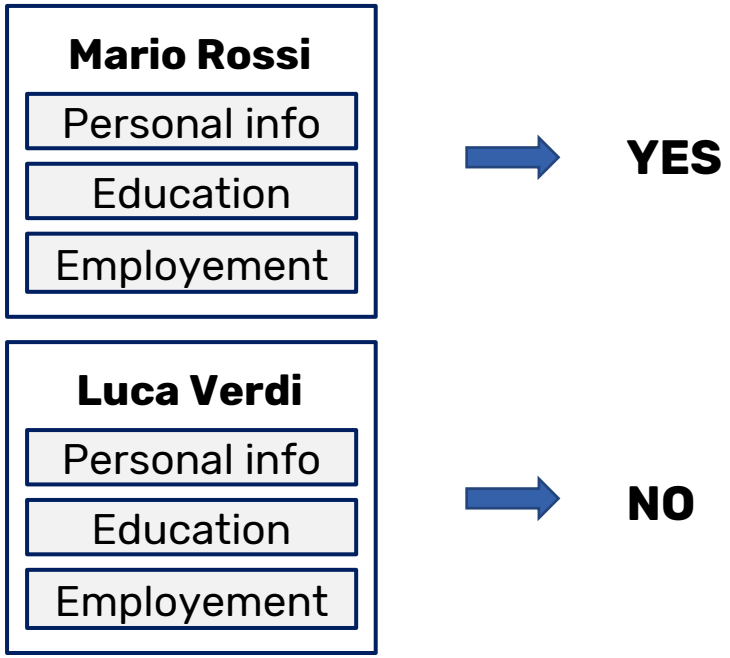
RECRUITING

Data science



- Optimize recruiting process

Machine learning



- Automatic resume screening

Selecting data-driven projects

MARKETING

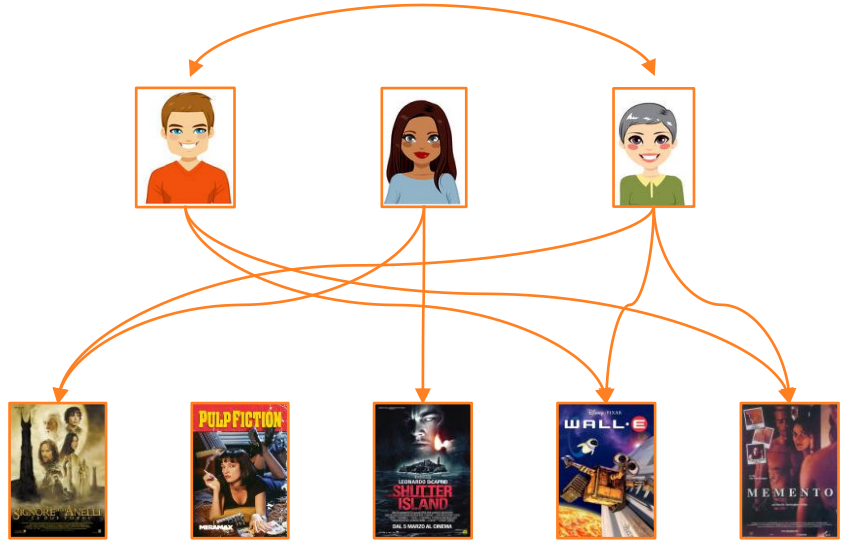
Data science

Version A

Version B

- A/B testing websites

Machine learning



- Recommendation system

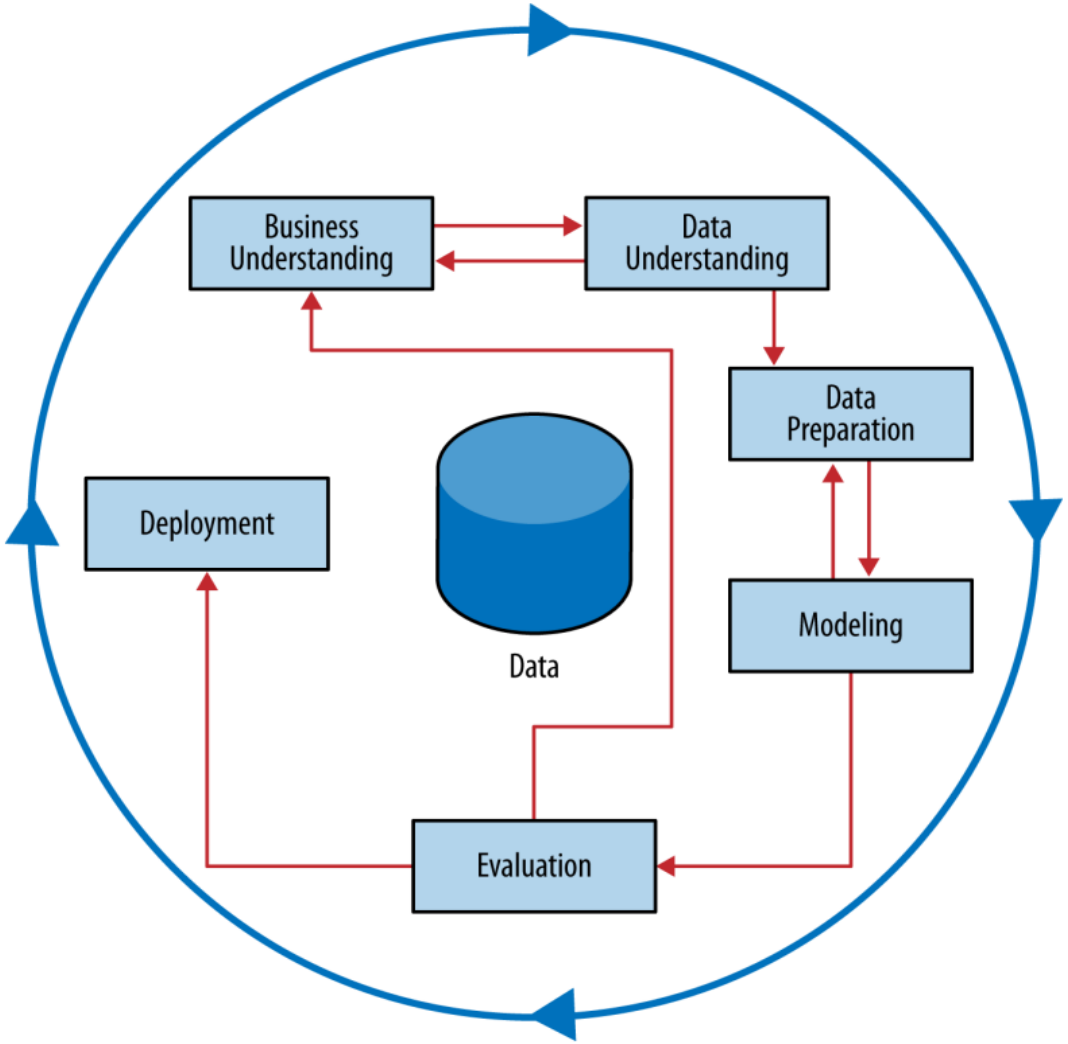
Outline

1. Course introduction
2. Data science and the data-driven company
3. Data and its types
4. What we are going to do with data (supervised and unsupervised learning)
5. Static and dynamical models in supervised learning
6. From business problems to data science tasks
7. **The data mining life cycle (CRISP-DM)**

CRISP-DM process

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

Picture taken from [1]



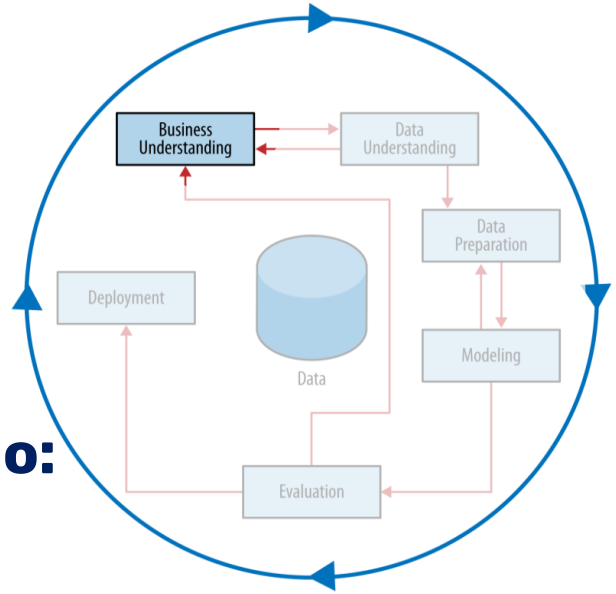
Iteration is the rule rather than the exception:

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

CRISP-DM: Business understanding

Cast the business problem into one or more data science tasks

- Regression
- Classification
- Causal modeling
- Clustering
- Co-occurrence grouping
- Profiling
- Link prediction
- Dimensionality reduction
- Similarity matching

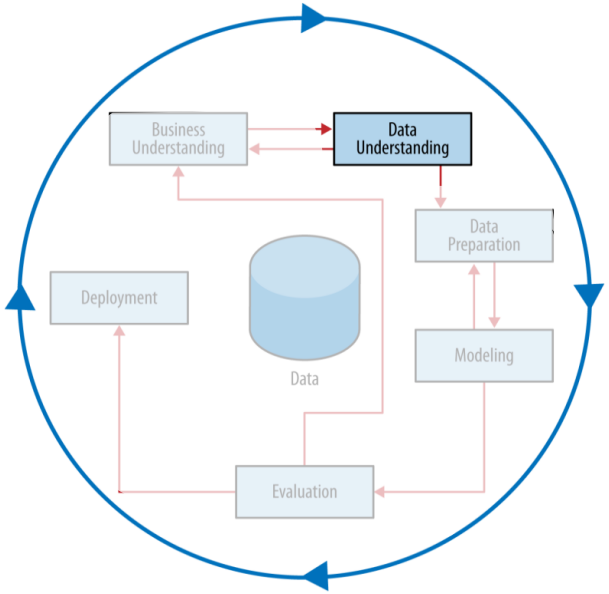


Think carefully about the **use scenario:**

- **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

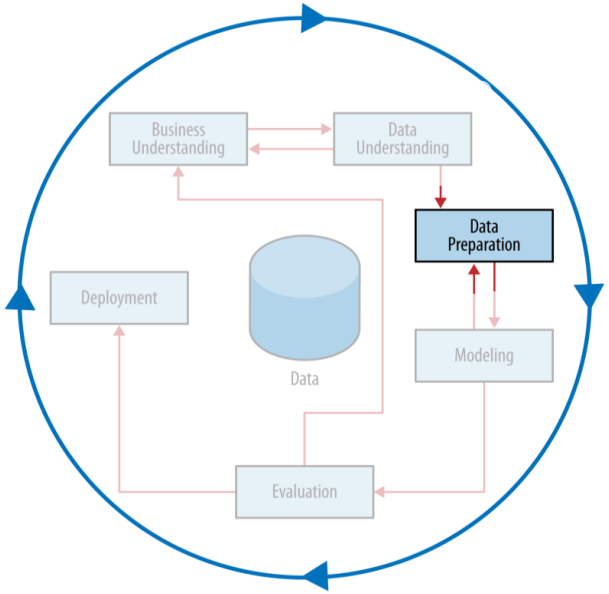
Are the data at our disposal **related to the business problem**?

Can we use a **proxy** for the data that we do not have?

As data understanding progresses, the **solution paths** may differ

CRISP-DM: Data preparation

Clean and prepare the data for usage



Usually, data mining algorithms require **data in a specific format** which is different from the one that is readily available

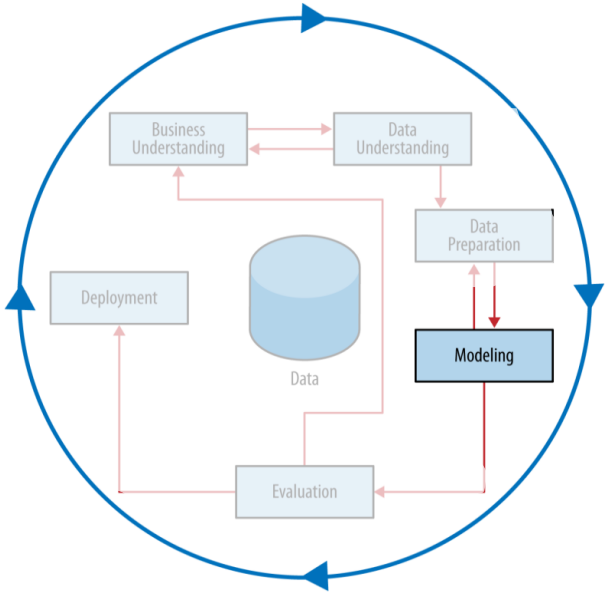
- Convert string to numbers, infer missing data, import data from excel files, …

Data preprocessing/cleaning/labeling **(most of data science project time is spent here)** [5]

Pay attention to not use historical data that **will not be available** when decisions need to be made

CRISP-DM: Modeling

Estimate a mathematical model to extract patterns from data



In most cases, **standard methods** can be directly applied on data

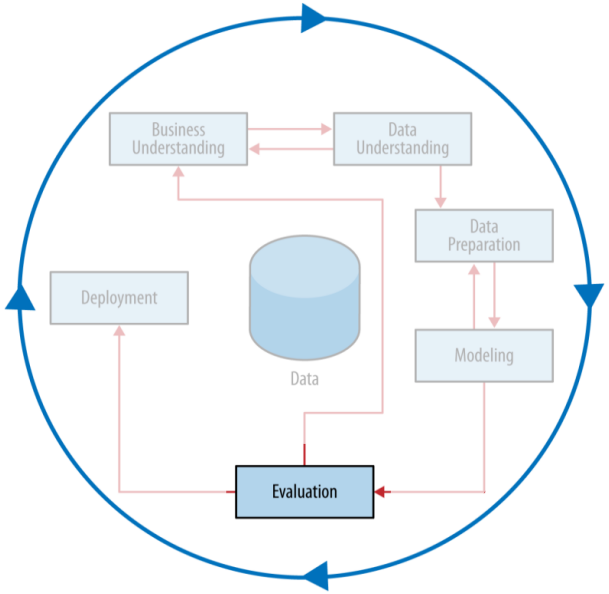
The aim is to find a model that performs well **on unseen data**

The type of the model is chosen based on:

- What data science **task** we want to solve
- Performance **measures**
- Availability of **libraries** for deployment

CRISP-DM: Evaluation

Assess the validity of the results



We could find patterns that exist only in the particular dataset that we have at our disposal **(overfitting)**

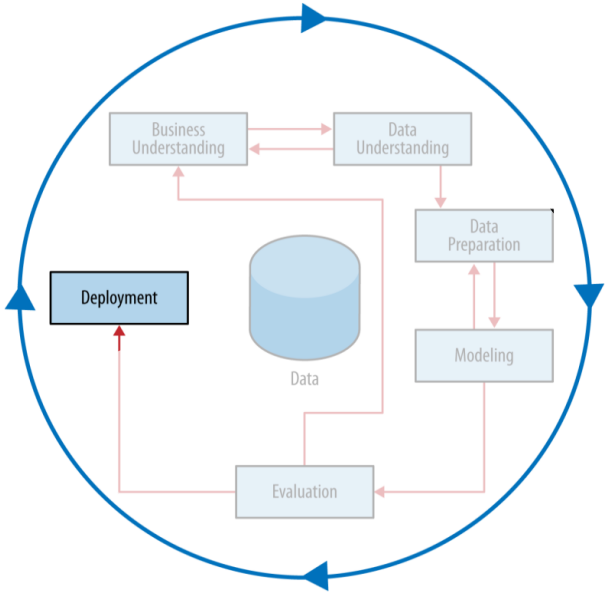
Does the model **satisfy** the original business goals?

The devised solution and the model's decisions should be **comprehensible** by the stakeholders

Usually, evaluation is performed **before deploying**. In this case, build environments that **closely mimic** the real use scenario

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 existing technologies

This step can require a notable **investment in time**. Usually, the data science team builds a prototype that is then passed on to the development team

For this reason, it is suggested to **include 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)

Example: CRISP-DM

Suppose that a bank company asks us to **recognize fraudulent transactions** (i.e., without the owner's consent) that may happen when a credit card is stolen or account credentials are taken over. How should we organize this data science project following the **CRISP-DM framework**?

Business understanding

Business problem: credit card fraud detection

⇩

Data science task:

- Classification? (fraud/not fraud)
- Profiling? (understand the typical spending behaviour of customers ⇔ fraud = difference from typical behaviour)

Data understanding

- What data do we have on transactions? E.g., amount paid, location, …
- What data do we have on customers? E.g., age, assets, …
- Do we have **labelled** data? I.e., do we actually know if transactions are fraudulent or not?

Supervised or unsupervised learning?

Example: CRISP-DM

Customers are likely to call the bank whenever they see a transaction that was not made by them
⇒ **we have labelled data**

Business understanding

Business problem: credit card fraud detection

⇩

Data science task: Classification (fraud/not fraud)

Data understanding

- Transaction data (**with** label fraud/not fraud)
- Customer data

Data preparation

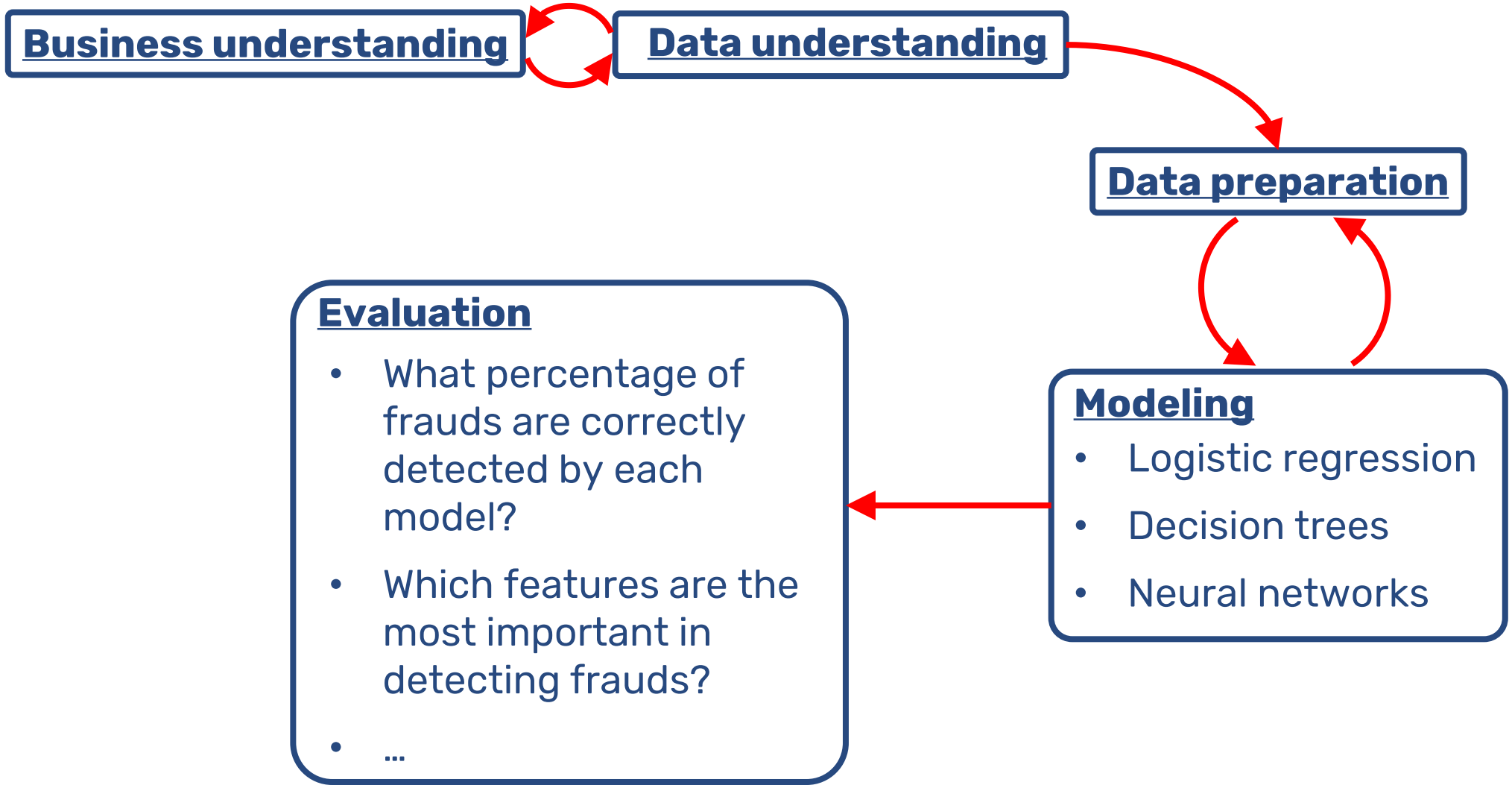
- Gather and clean relevant data from the bank databases
- Merge transaction and customer data
- …

Modeling

- Logistic regression
- Decision trees
- Neural networks

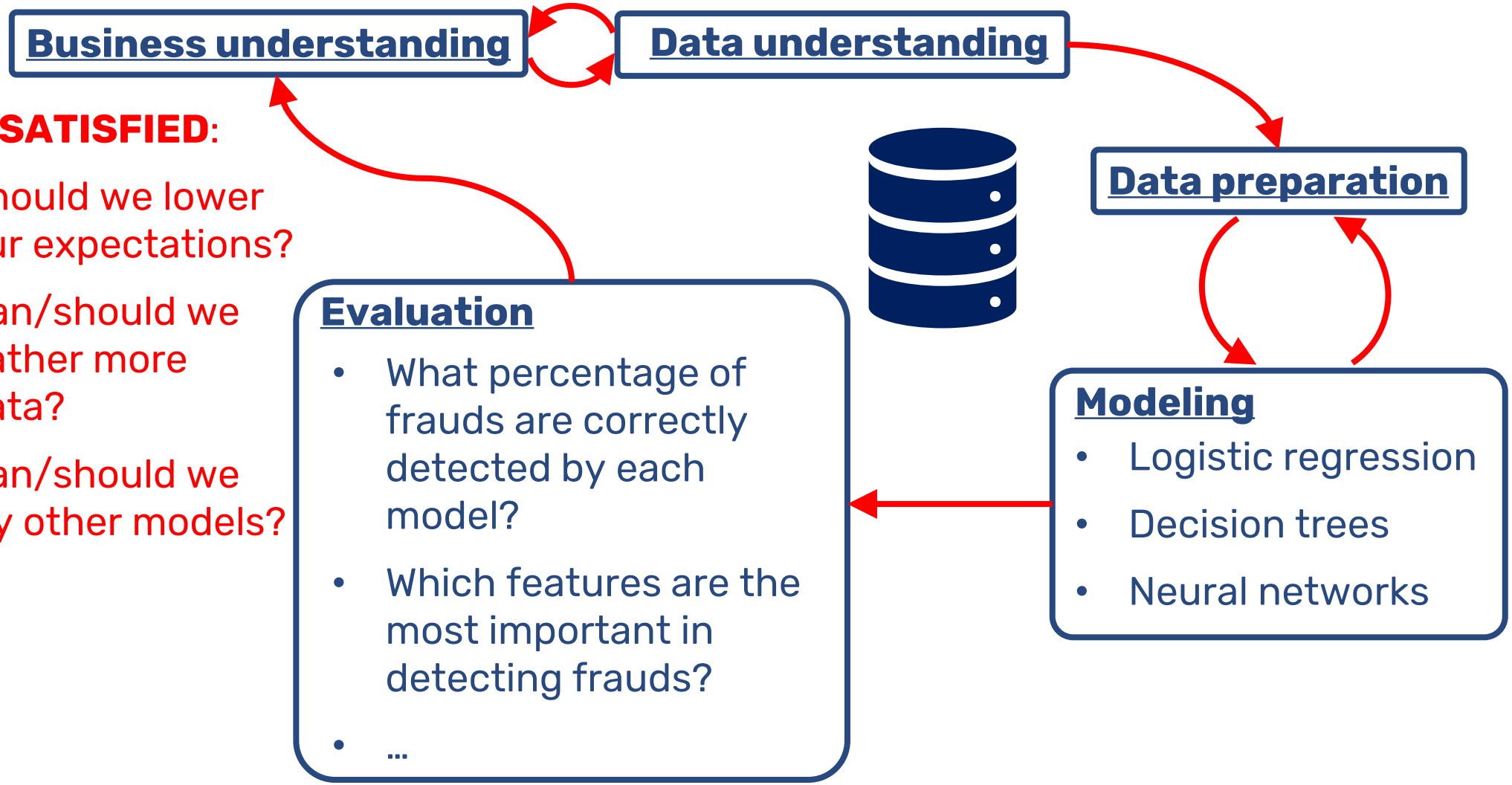
- Do we need more transaction data?
- Should we exploit additional information on the customers?
- …

Example: CRISP-DM

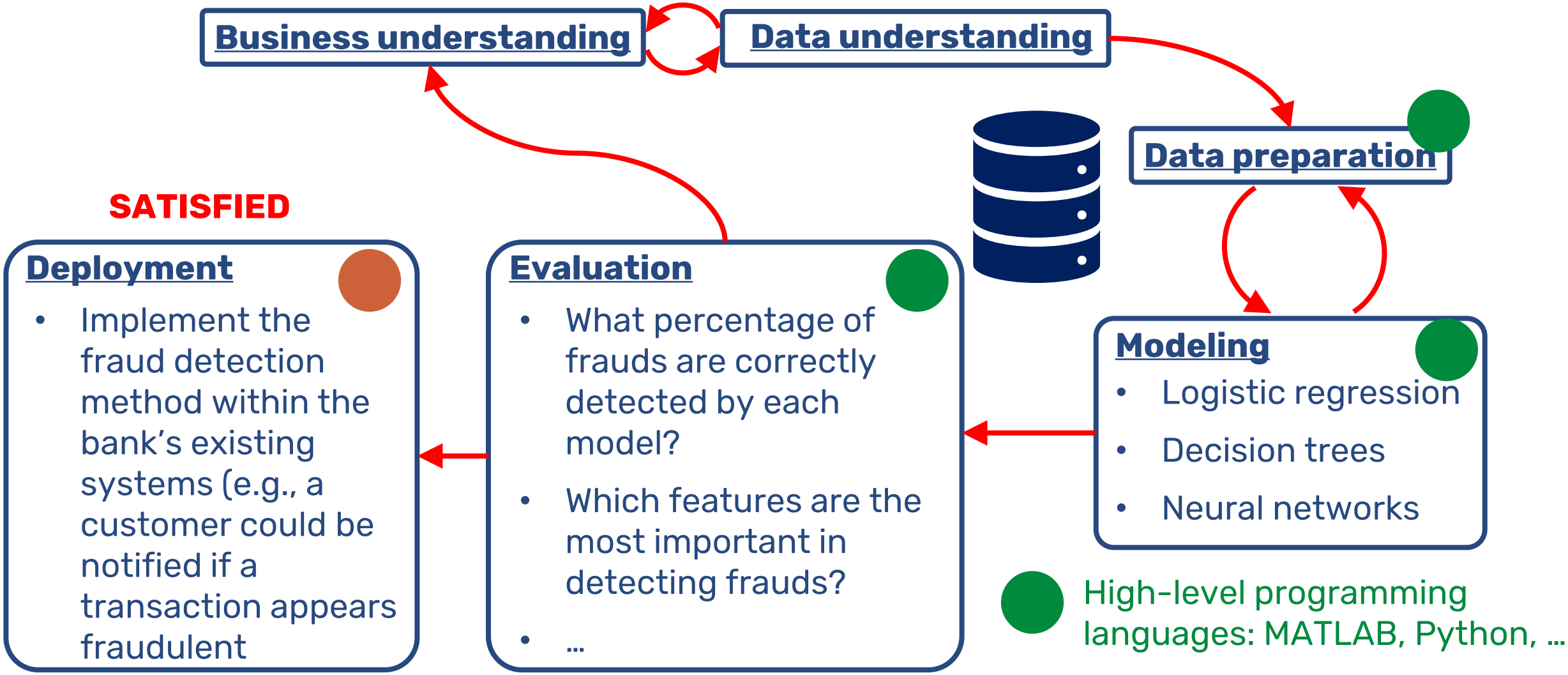


Say that the best-performing model is able to detect 90% of frauds. Are the stakeholders **satisfied**?

Example: CRISP-DM



Example: CRISP-DM



Quiz

1. Which of these data science tasks can be tackled by unsupervised learning methods? *Select all that apply.*
   - ❑ Clustering
   - ❑ Regression
   - ❑ Link prediction
   - ❑ Causal modeling

2. There is a one-to-one correspondence between a business problem and a data science task.
   - ❑ True
   - ❑ False

3. Which of these business problems can be tackled by clustering methods? *Select all that apply.*
   - ❑ Reduce the dimensionality of data for visualization purposes
   - ❑ Market segmentation (e.g. identify customer segments with common needs and similar purchasing habits)
   - ❑ Find groups of patients that exhibit comparable symptoms
   - ❑ Predict the prices of products based on their characteristics

UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

References

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. **Chapters 1-2**.
2. 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
3. Nucleus Research, 2014. http://bit.ly/XQFDbv.
4. Notes from the AI frontier: Modeling the impact of AI on the world economy, 2018.
5. Pyle, D. "*Data Preparation for Data Mining*". Morgan Kaufmann, 1999. **Chapter 1**.
6. G. James, D. Witten, T. Hastie, R. Tibshirani. *"An Introduction to Statistical Learning*". 2° Edition, Springer, 2021. **Chapters 1-2**.
7. Data scientist: The Sexiest Job the 21st Century, 2012.
8. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025, 2022.
9. Correlation does not imply causation: 5 real-world examples, 2021.