UNIVERSITÀ DEGLI STUDI DI BERGAMO | Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione

Data Analysis Lab

ICT/Electronic Lab

Lecture 02: Introduction to data science

SPEAKER
Prof. Antonio Ferramosca

PLACE
University of Bergamo

Outline

1. Data science and the data-driven company
2. Data and its types
3. What we are going to do with data (supervised and unsupervised learning)
4. Static and dynamical models in supervised learning

Outline

1. **Data science and the data-driven company**
2. Data and its types
3. What we are going to do with data (supervised and unsupervised learning)
4. Static and dynamical models in supervised learning

«Data is the new oil»

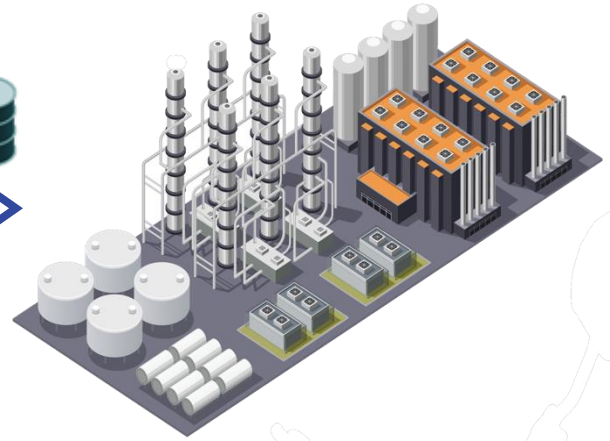
Introduction

Data are considered the **new oil**, due to their enormous value nowadays.

«Data is the new oil»

Crude oil extraction

Barrels of oil

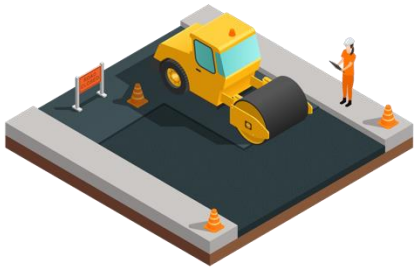


Refinement process

- Fuels
- Oils
- …

- Automobiles,
- Planes,
- Generators,
- Engines,
- ...

Asphalt

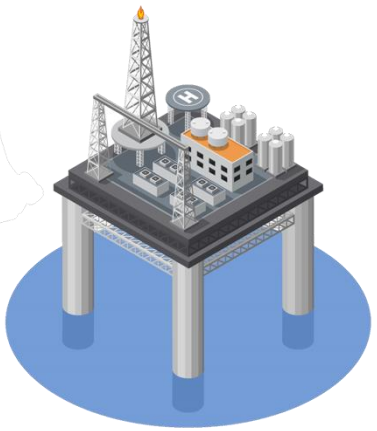


- Infrastructures,
- Streets,
- …

Plastic

- Bottles,
- Containers,
- Films,
- ...

«Data is the new oil»

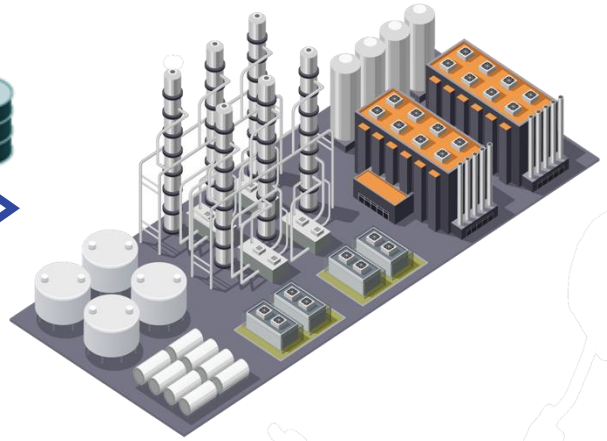


~~Crude oil~~ extraction

~~Barrels of~~ oil

DATA

DATUM

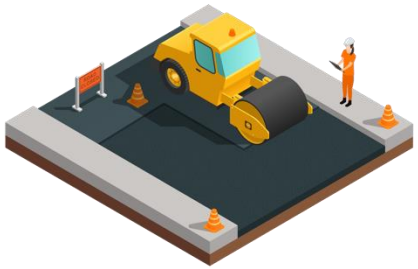


Refinement process

- Fuels
- Oils
- …

- Automobiles,
- Planes,
- Generators,
- Engines,
- ...

Asphalt



- Infrastructures,
- Streets,
- …

Plastic

- Bottles,
- Containers,
- Films,
- ...

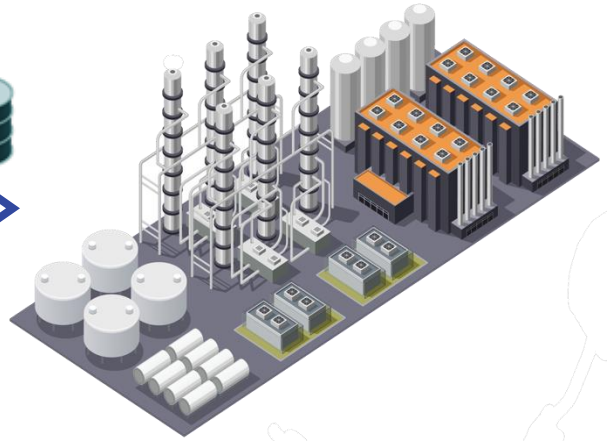
«Data is the new oil»

~~Crude oil~~ extraction

DATUM

~~Barrels of~~ oil

DATA

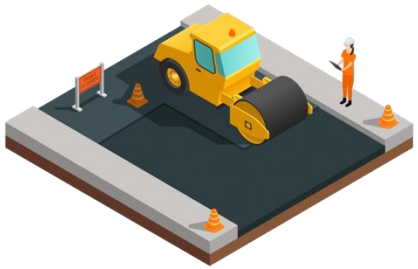


Refinement process

PRODUCT QUALITY

- Automobiles,
- Planes,
- Generators,
- Engines,
- ...

GOODS FORECAST

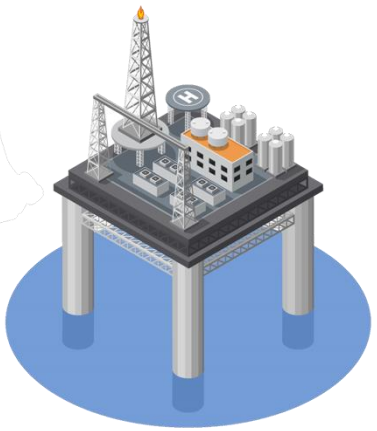


- Infrastructures,
- Streets,
- …

PROCESS OPTIMIZATION

- Bottles,
- Containers,
- Films,
- ...

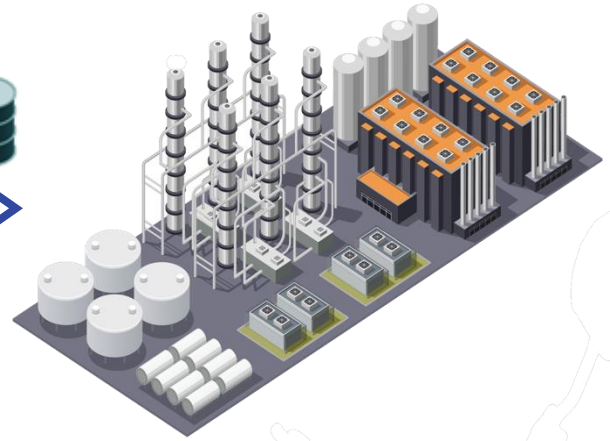
«Data is the new oil»



~~Crude oil~~ extraction

~~Barrels of~~ oil **DATA**

DATUM

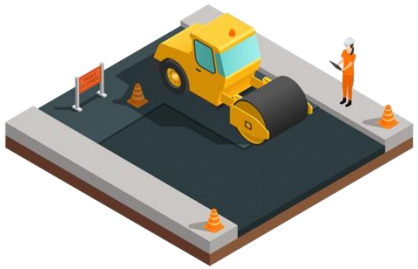


Refinement process

PRODUCT QUALITY

- **Machine parameters optimization**

GOODS FORECAST

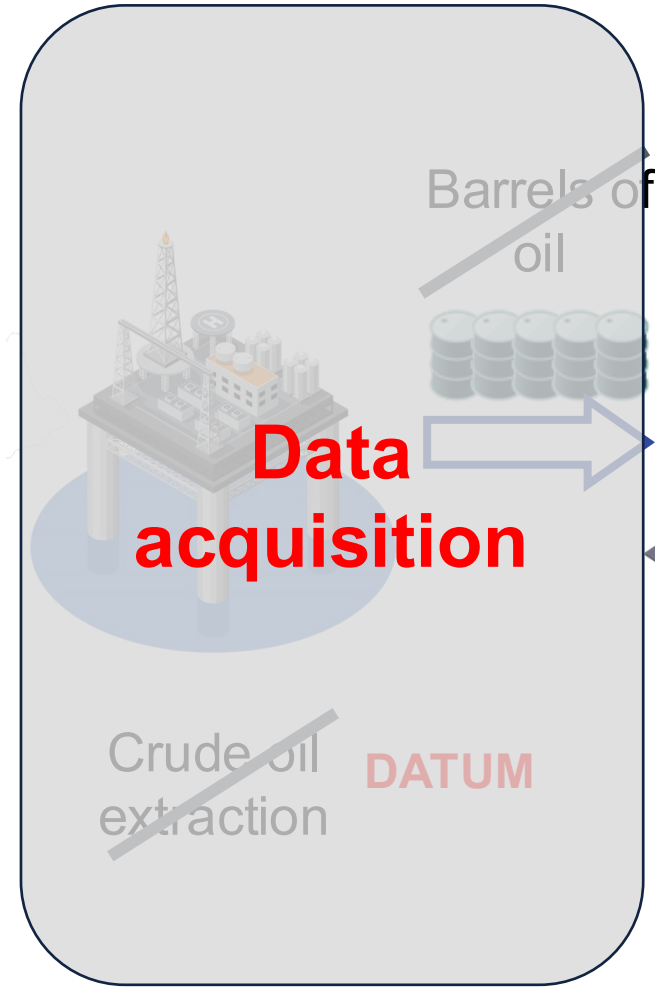


- **Production and purchasing management**

PROCESS OPTIMIZATION

- **Reduction of materials used**

«Data is the new oil»



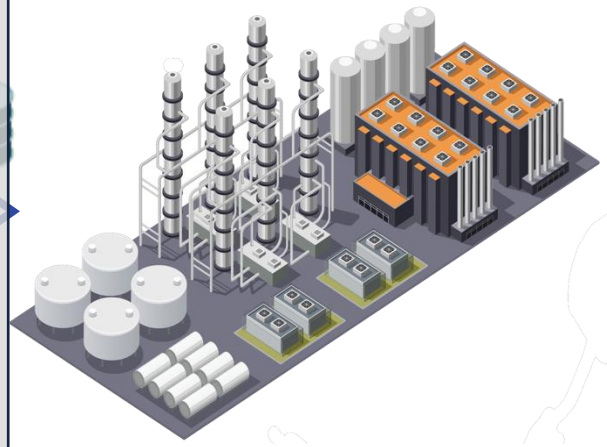
Barrels of oil

DATA

Data acquisition

Crude oil extraction

DATUM

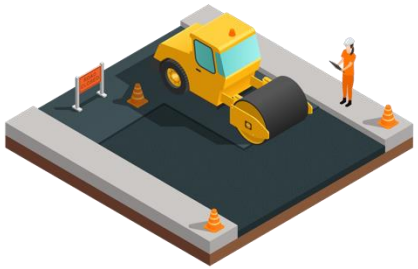


Refinement process

PRODUCT QUALITY

GOODS FORECAST

PROCESS OPTIMIZATION



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

«Data is the new oil»

Barrels of oil

DATA

Crude oil extraction

DATUM

Data acquisition

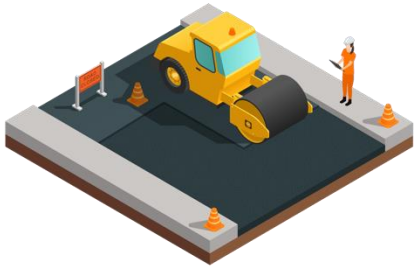
Descriptive analytics and reporting

Refinement process

PRODUCT QUALITY

- **Machine parameters optimization**

GOODS FORECAST

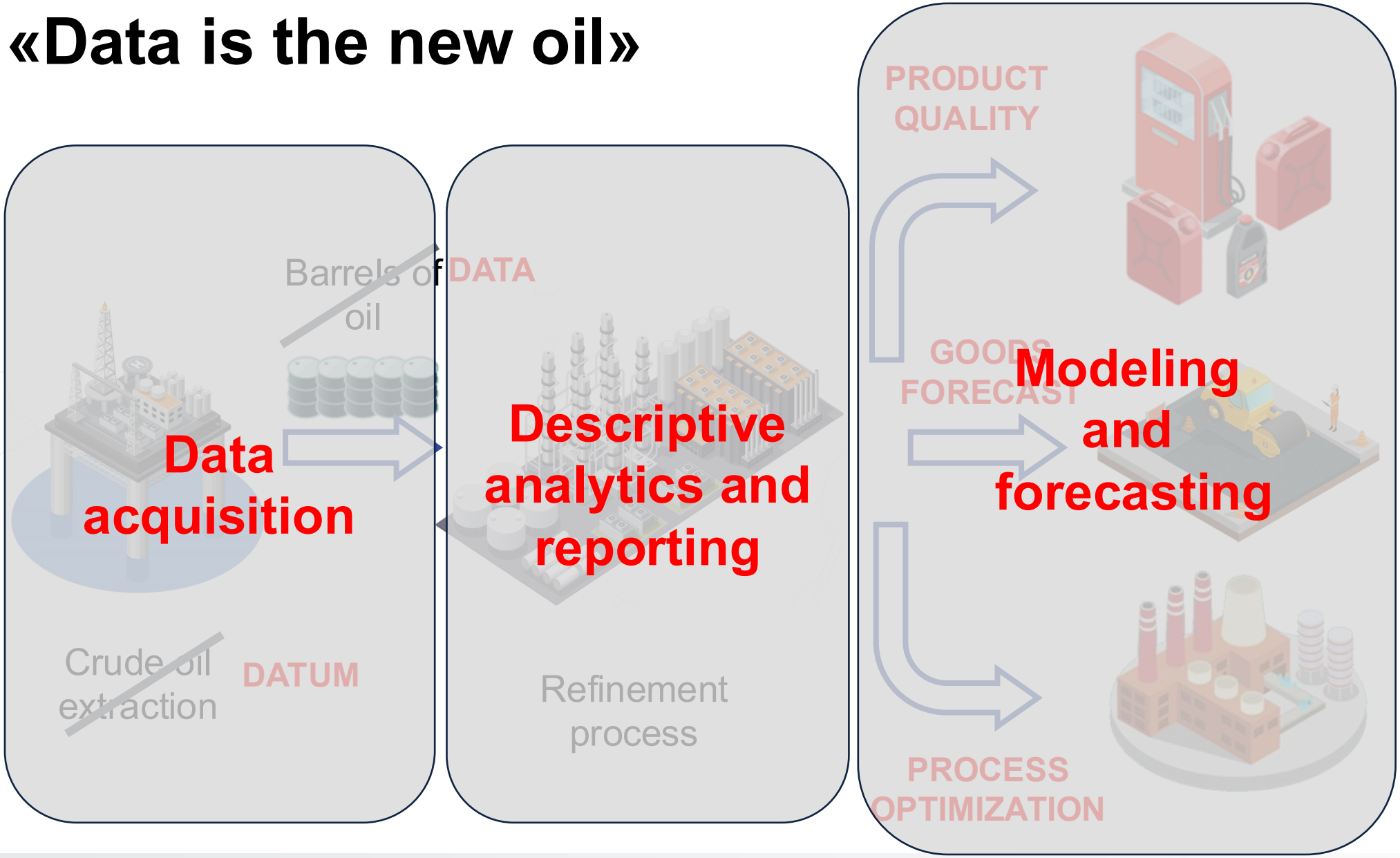


- **Production and purchasing management**

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

DATA

Data acquisition

Crude oil extraction

DATUM

Descriptive analytics and reporting

Refinement process

PRODUCT QUALITY

GOODS FORECAST

Modeling and forecasting

PROCESS OPTIMIZATION

Actions

- Machine parameters optimization
- Production and [illegible] management
- Reduction of materials used

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 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:

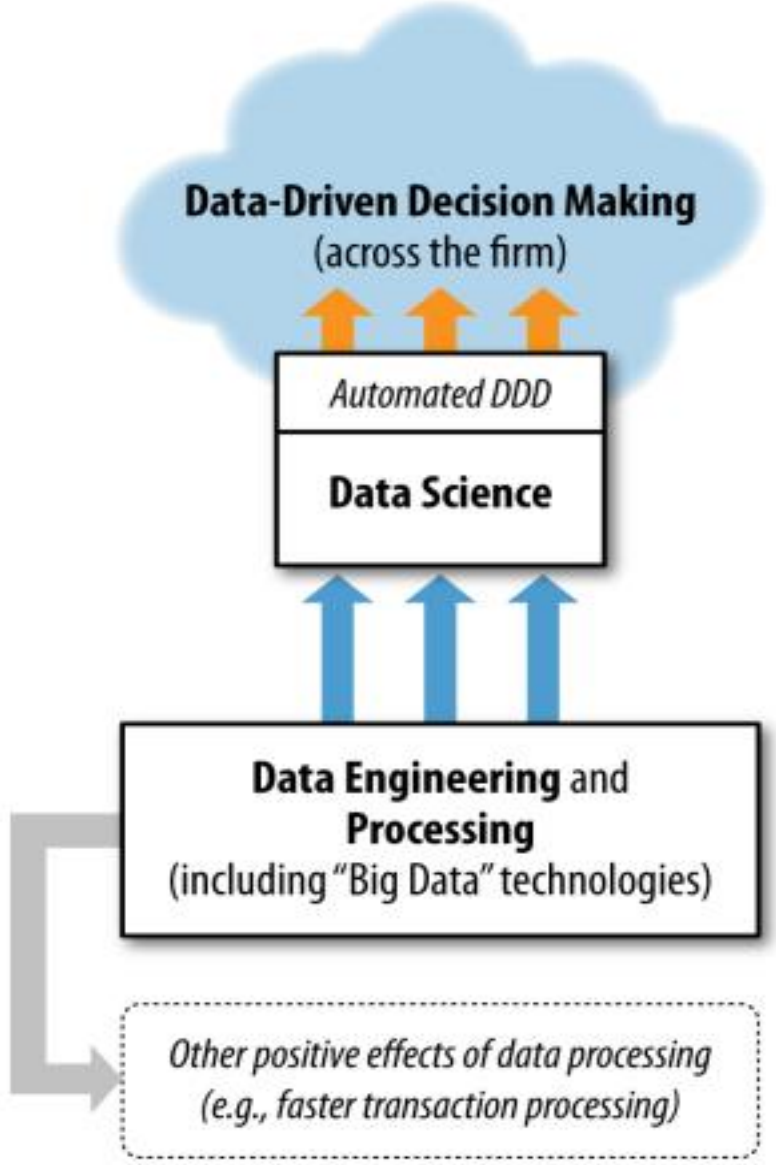
- Mathematics
- Statistics
- Machine learning and artificial intelligence
- Specialized programming

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

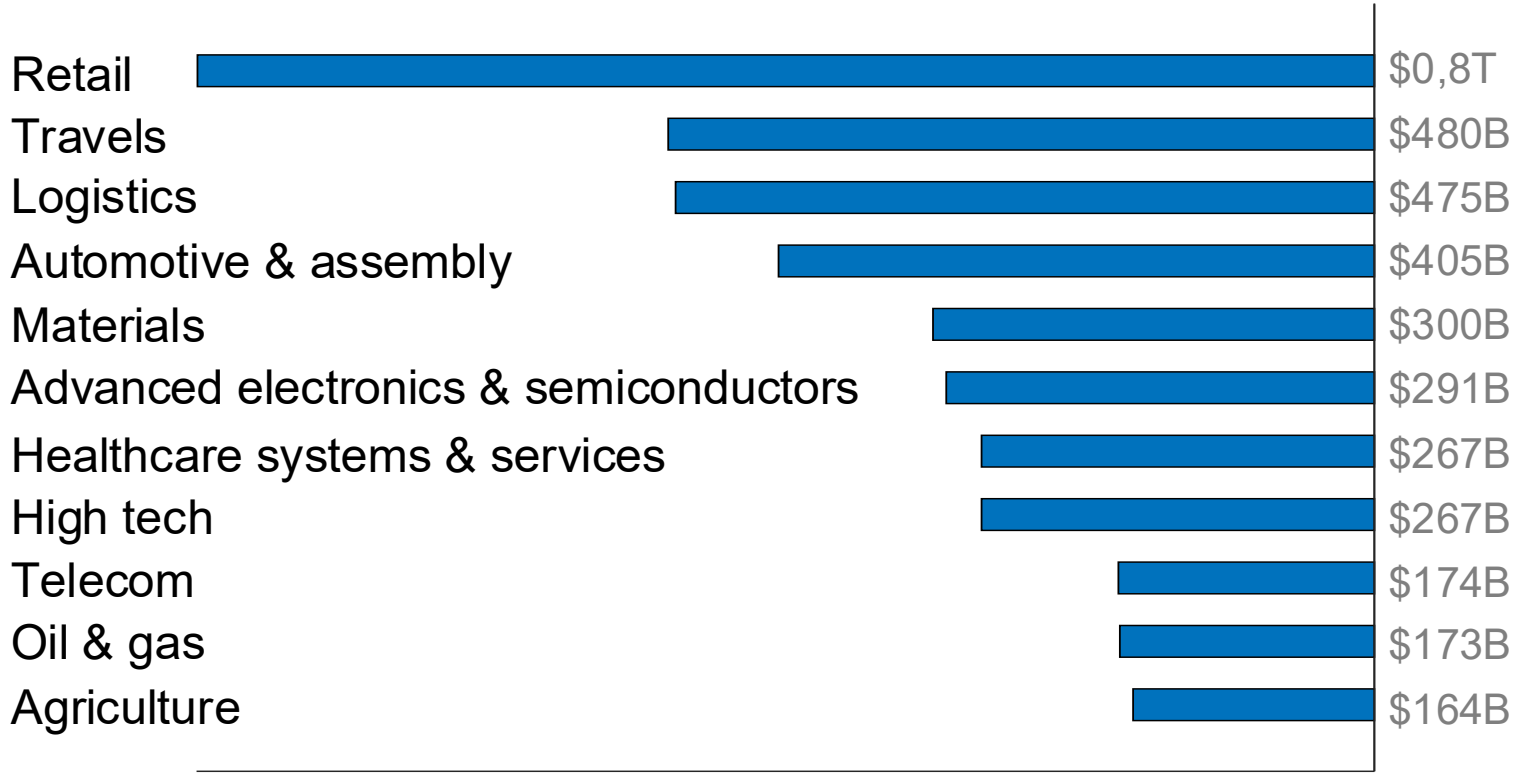
5% more productive [2]

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

Example in real life: football player

Scenario

29 years old football player with contract expiring soon

Aim

Maximize own performance and profit

Old style football player

«Data-driven» football player

analytics fc

Data science team

- What is the **best** football team **for me**?
- Since I do not score so many goals, **how can I prove my importance** in the game?
- How **would** the current team **have performed without me**, or with some «competitor» in my place?

«Data-driven» football player

analytics fc

Data science team

«Data-driven» football player

analytics fc

Data science team

«I am the **best midfielder for you** in this moment»

«Data-driven» football player

analytics fc

Data science team

+30%

MANCHESTER CITY 18 94

Outline

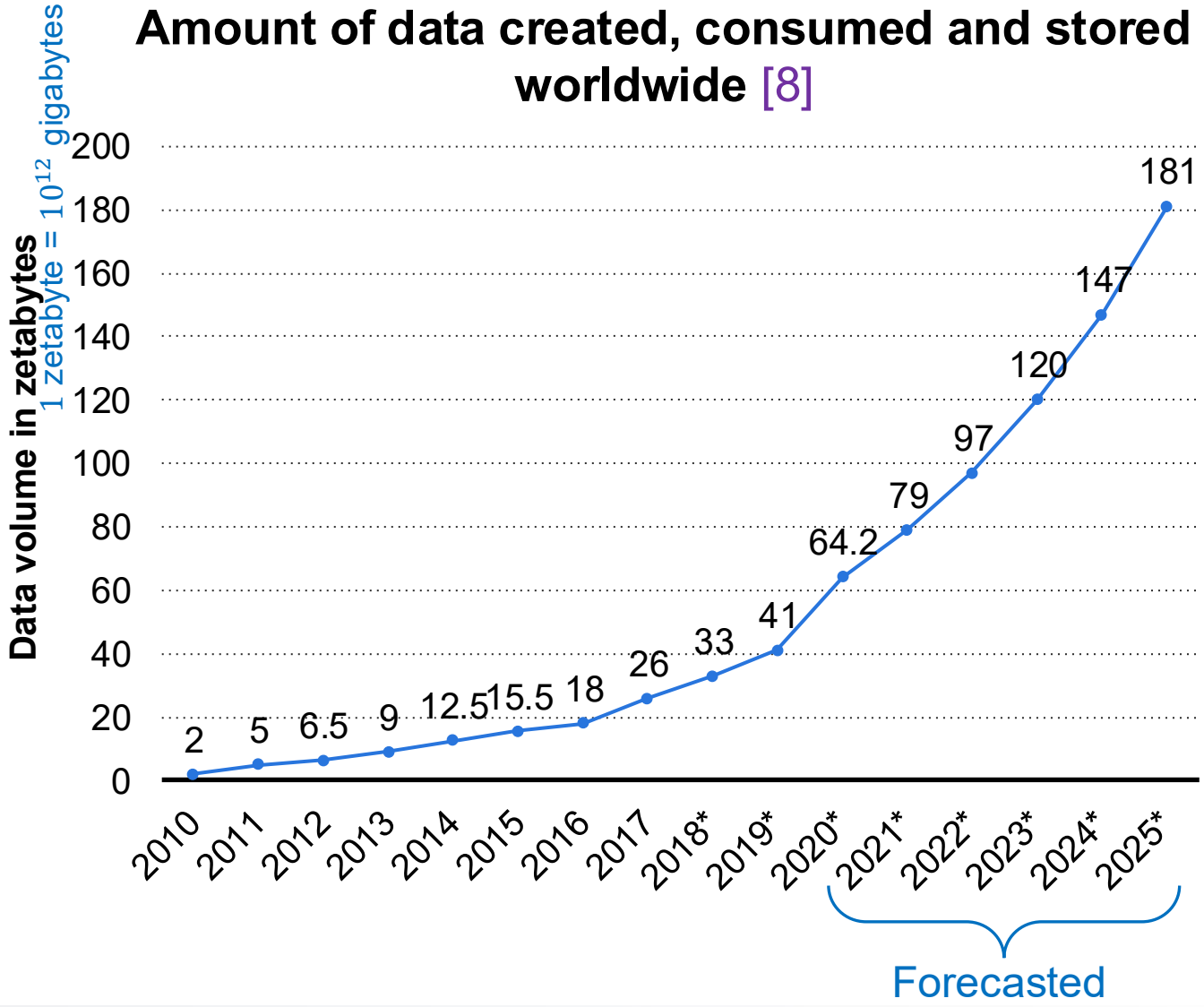
1. Data science and the data-driven company
2. **Data and its types**
3. What we are going to do with data (supervised and unsupervised learning)
4. Static and dynamical models in supervised learning

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

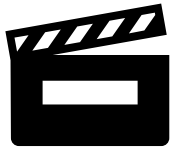


Image files

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. Data science and the data-driven company
2. Data and its types
3. **What we are going to do with data (supervised and unsupervised learning)**
4. Static and dynamical models in supervised learning

What are we going to do with data?

We can use data for:

- **Descriptive analysis** and **visualization**
- **Supervised learning** (in particular, regression and classification)
- **Unsupervised learning** (in particular, clustering and dimensionality reduction)

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
⋮	⋮	⋮

$\varphi \in \mathbb{R}$

$y \in \mathbb{R}$

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

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

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

Images are basically matrices of numbers that describe color intensity

Output: the class label

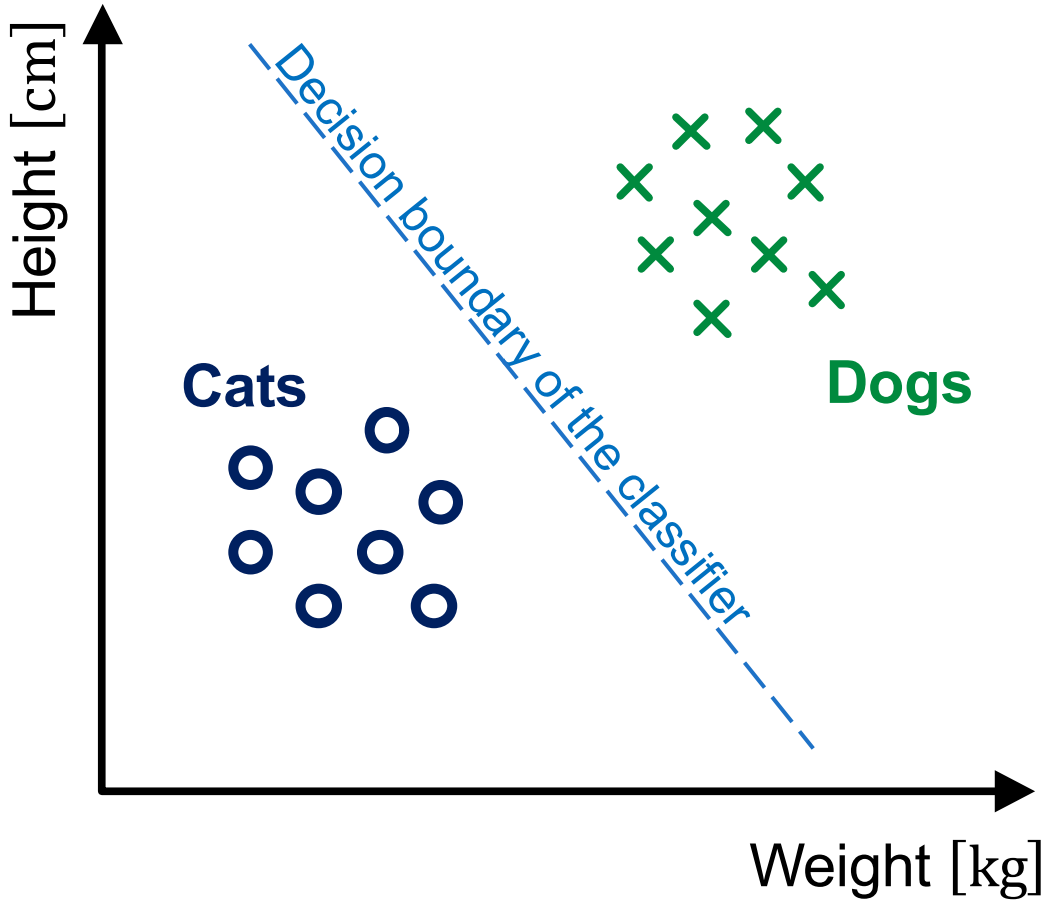
$$y \in \{\text{Cat}, \text{Not cat}\}$$

(single output)

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

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

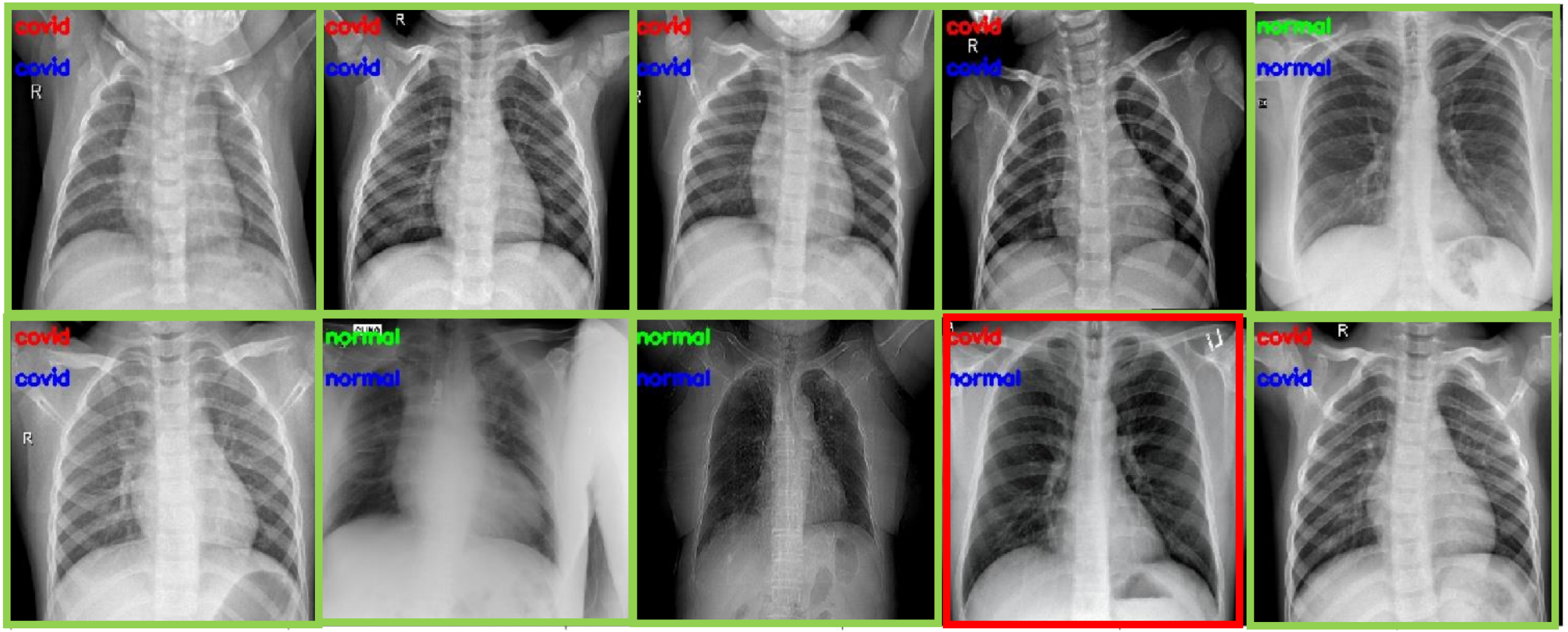
(single output)

Example: image classification

Predicted covid label

Predicted healthy label

True label



The classification problem

Estimating (or predicting) a categorical output from a set of features $\boldsymbol{\varphi}$ can be referred to as **classifying an observation with inputs $\boldsymbol{\varphi}$**, i.e. **assigning the observation to a category** (or class) among those in $\mathcal{C}$

Often, we are more interested in estimating the **probabilities** that $\boldsymbol{\varphi}$ belongs to each category in $\mathcal{C}$

The **most probable category** is then chosen as the class for $\boldsymbol{\varphi}$

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?

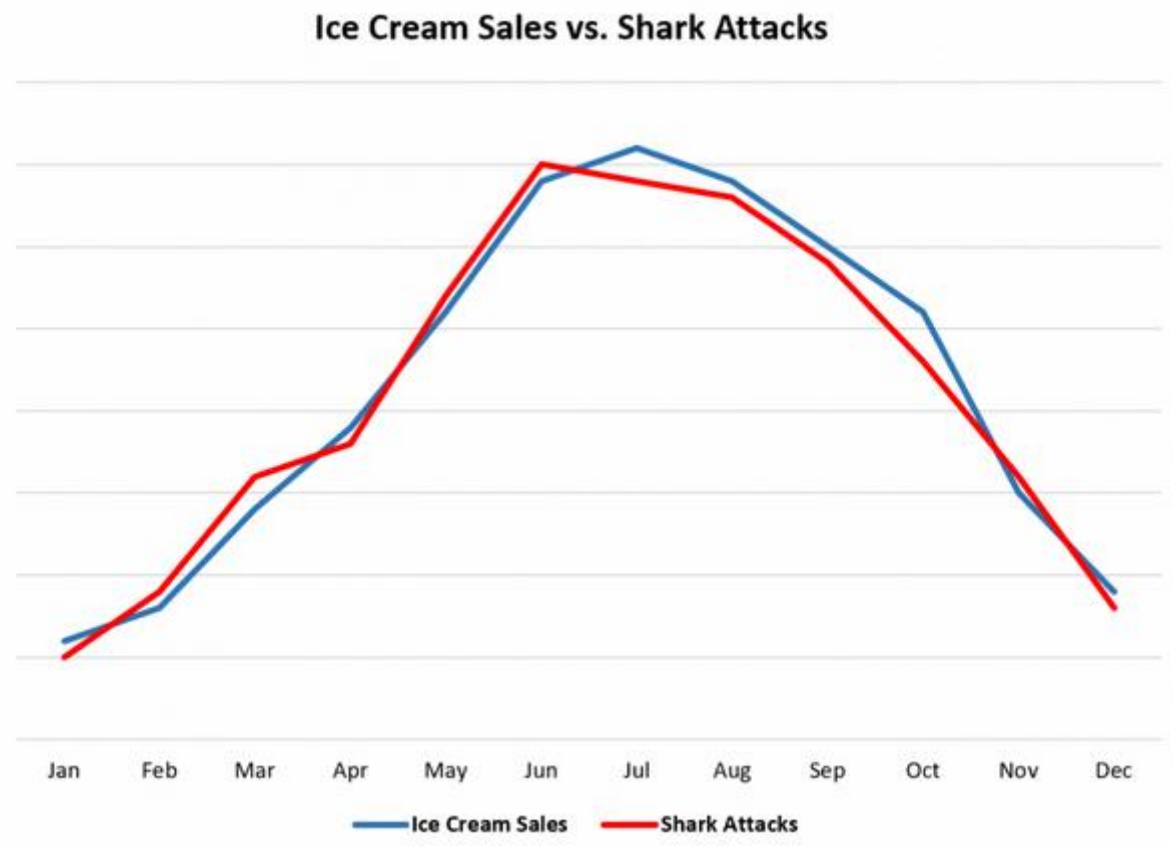
  Causal modeling typically involves substantial investments in data, such as randomized controlled experiments (**A/B tests**) and sophisticated methods for drawing causal observation data (**"counterfactual" analysis**)

  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

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!

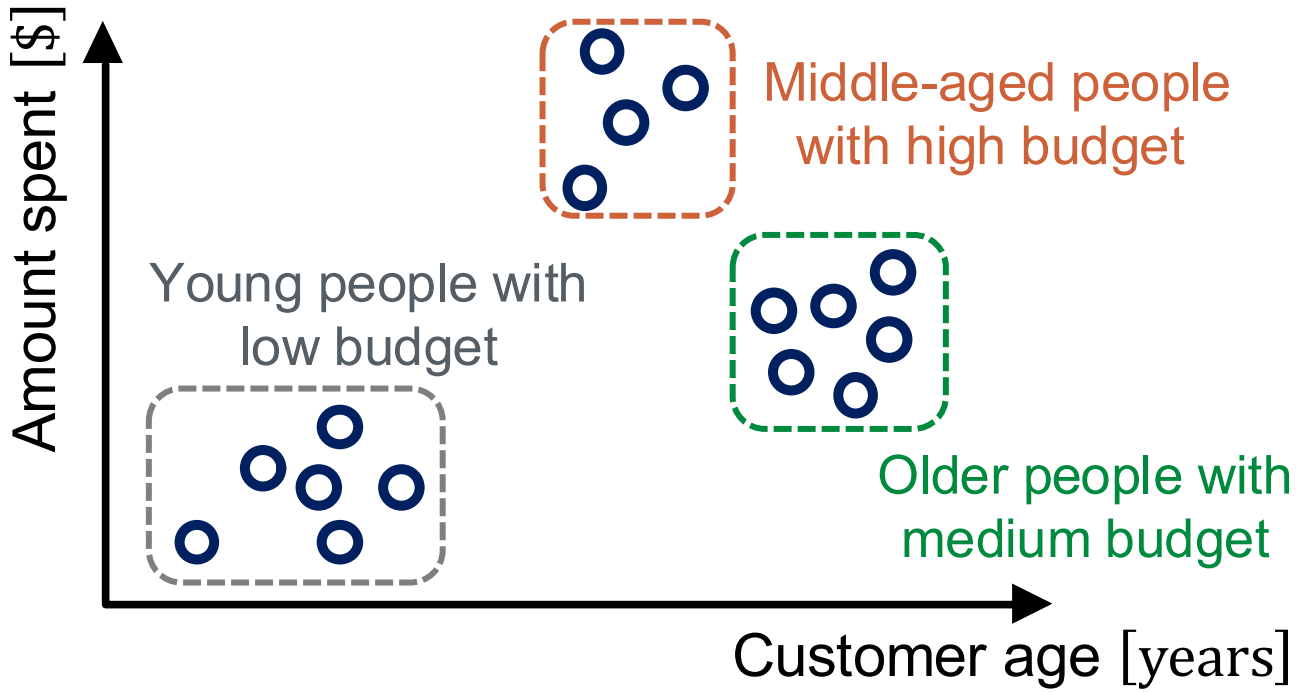
If we take a look at the data representing monthly ice cream sales and monthly shark attacks around the United States each year, we can see that the two variables are highly correlated

- 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

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



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

(customer age and amount spent)

Output: none

Data science tasks

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

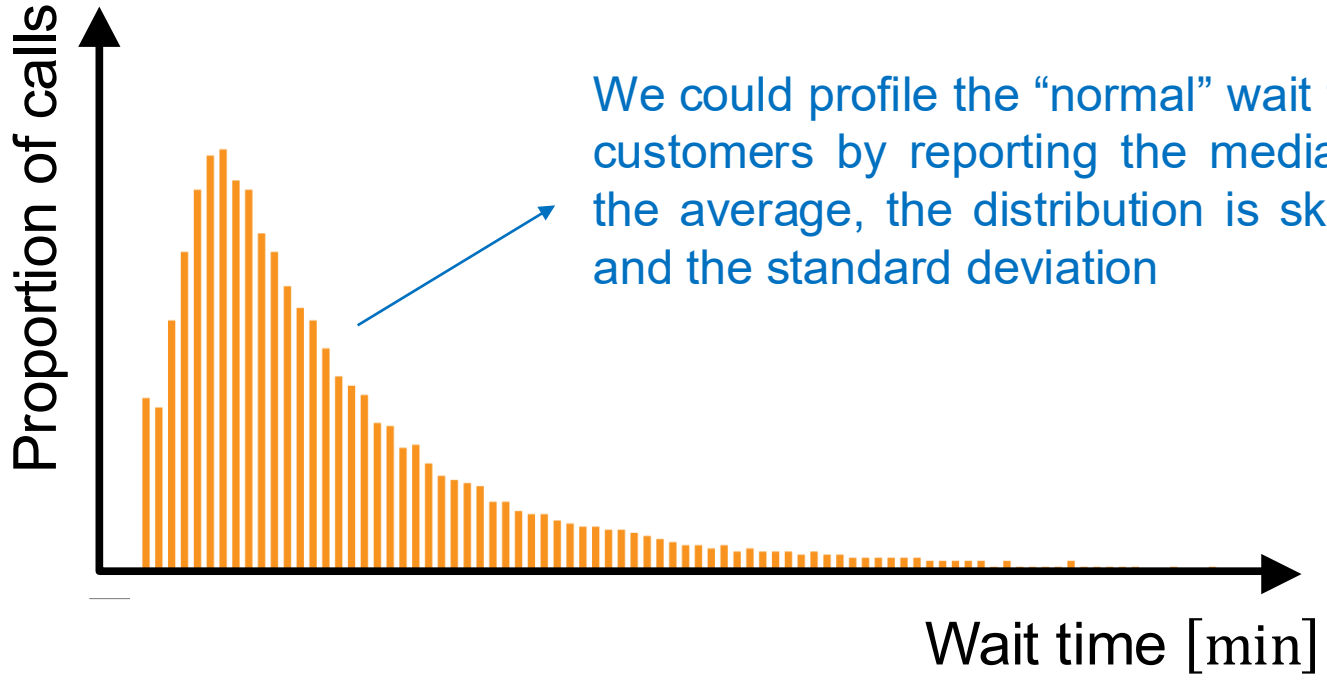
Example: ➢ What items are commonly purchased together? (**market basket analysis**)

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”)

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



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

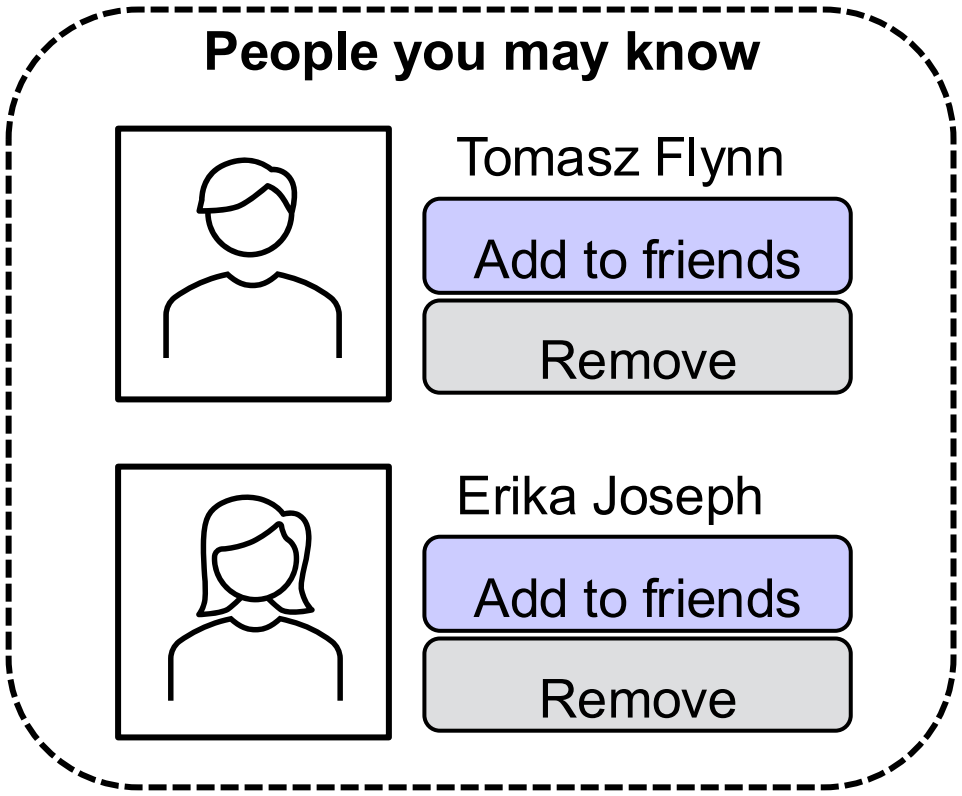
Output: none

Picture taken from [1]

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



: 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)

Inputs:

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

Output: none (in this example)

Picture taken from [1]

Data science tasks

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

Data science tasks vs algorithms

Data science task

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

Regression, classification, …

$\neq$

Algorithm (or method)

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

Neural networks, KNN, K-means clustering, …

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

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

- Different algorithms can solve the same data science task

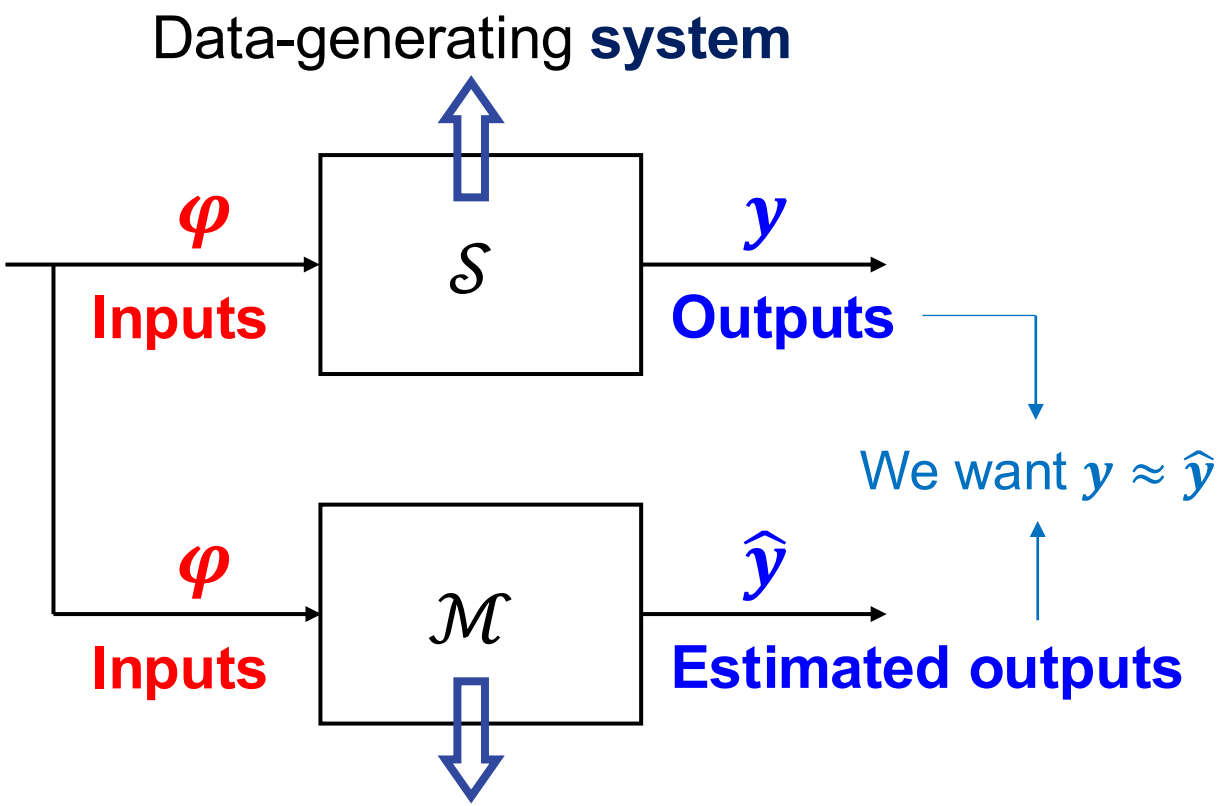
  A regression problem can be solved by the linear regression method, neural networks and KNN

Outline

1. Data science and the data-driven company
2. Data and its types
3. What we are going to do with data (supervised and unsupervised learning)
4. **Static and dynamical models in supervised learning**

Models in supervised learning

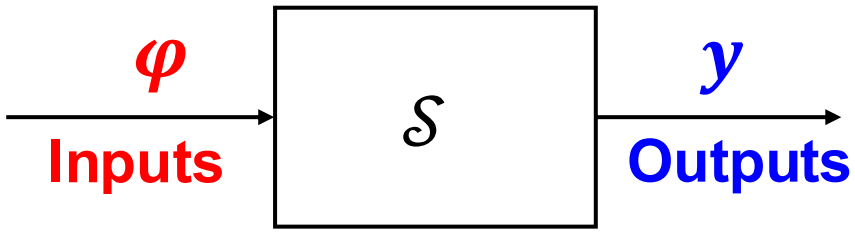
Most supervised learning methods rely on mathematical **models** that describe the relationship between the **inputs** and the **outputs**



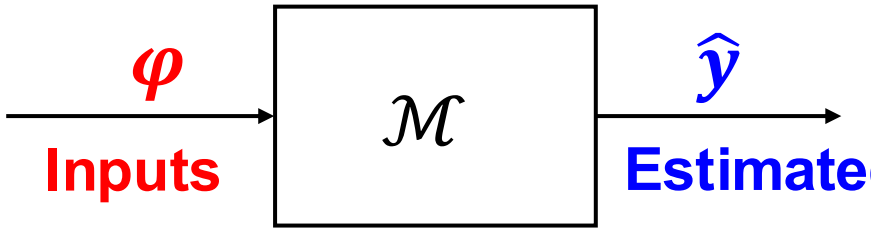
Supervised learning methods estimate $\mathcal{M}$ from data

Models in supervised learning

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



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

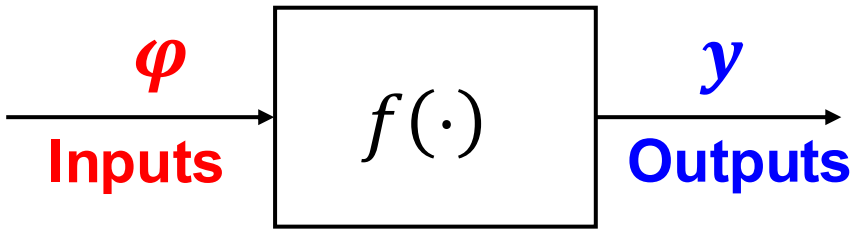


$\equiv$ $\hat{y} = \hat{f}(\varphi)$

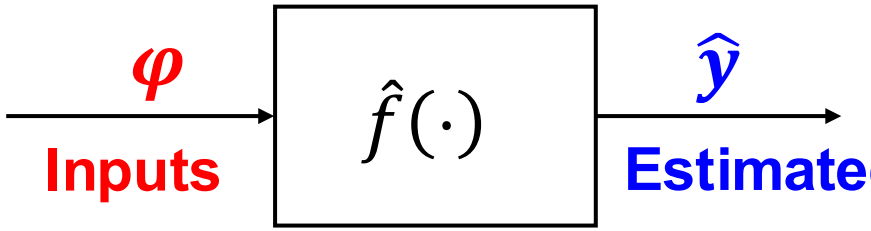
The goal of supervised learning methods is to learn a function $\hat{f}(\cdot)$ that approximates $f(\cdot)$ well **on the whole domain** of φ

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 $\mathcal{D} = \left\{\left(\boldsymbol{\varphi}(1), y(1)\right), \ldots, \left(\boldsymbol{\varphi}(N), y(N)\right)\right\}$

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

(N observations in total)

Static systems (and models)

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 (**but also dynamical systems!**) is the **linear model**

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

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

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

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**

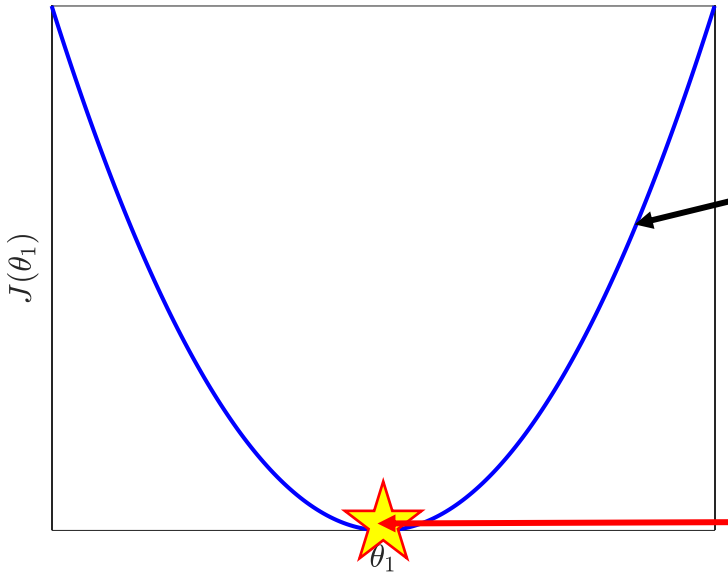
A typical cost in the regression setting is the following

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

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



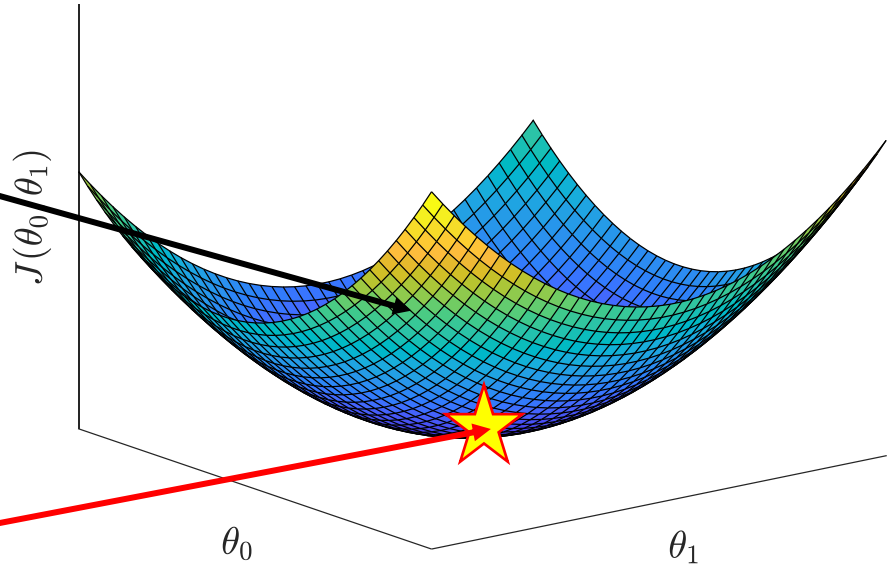
Cost function

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

Minimizer of the cost function:

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

Multiple parameters θ

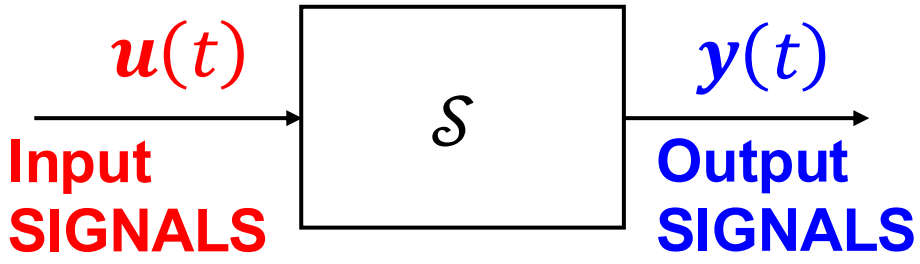


This rationale is followed by the **linear regression method**

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

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 trend**

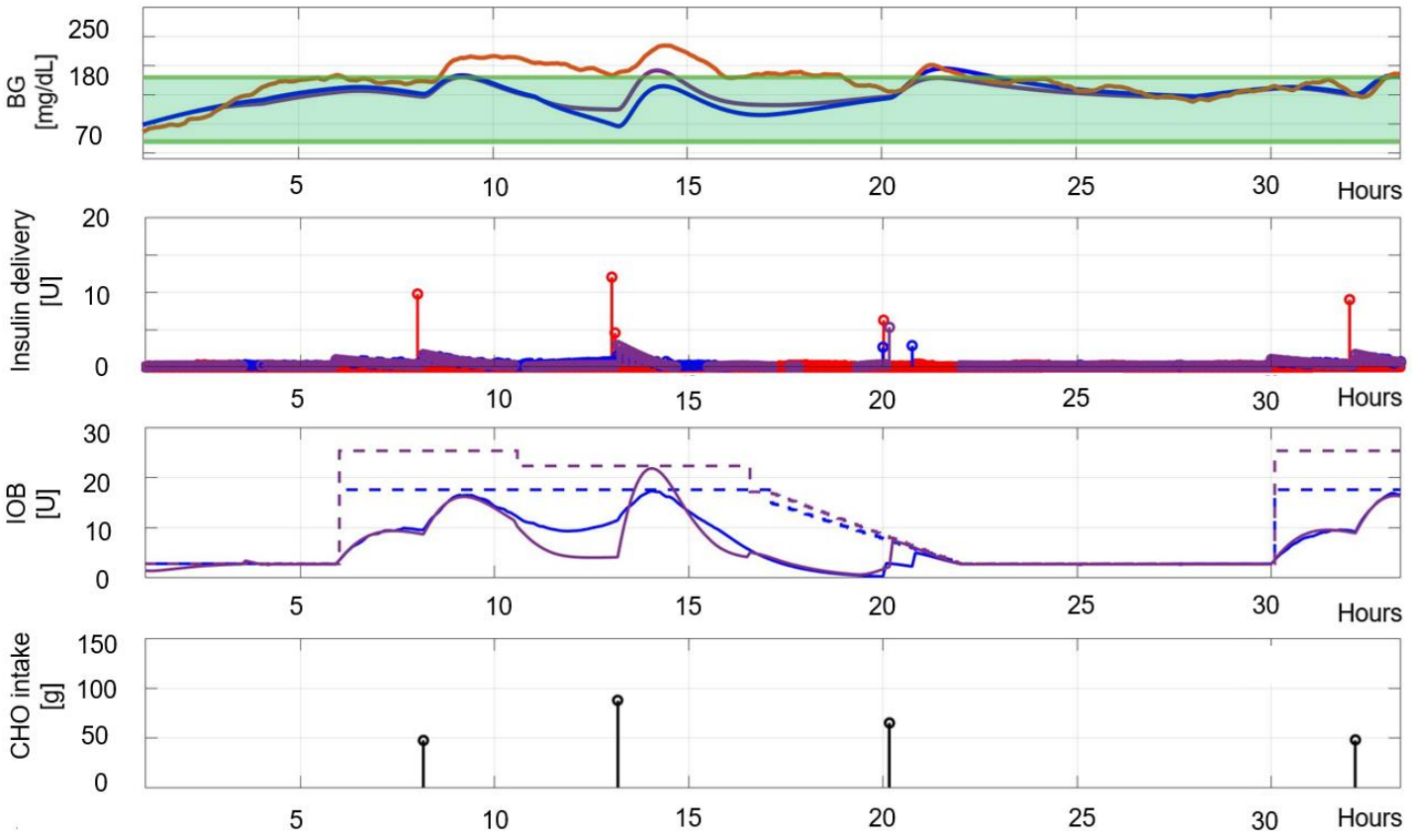
Dynamical systems usually involve the **time**: the **outputs** $y(t)$ at a certain time t **depend on the outputs at previous times**

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)

T1 Diabetes patient model:

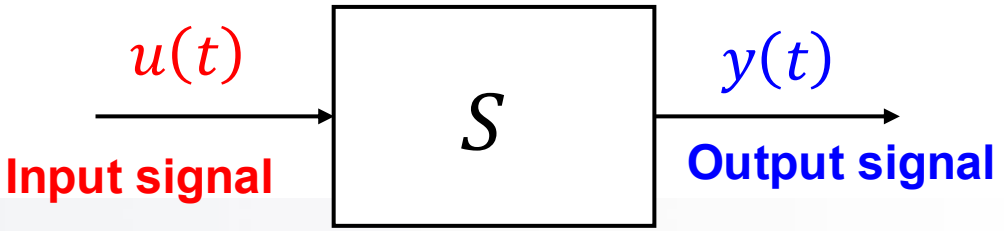


Output: Blood Glucose

Input: Insulin delivery

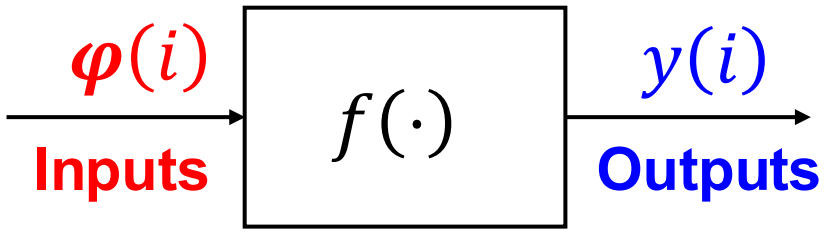
CHO Intake

The BG $y(t)$ at a certain time t **depends on its values at previous times**

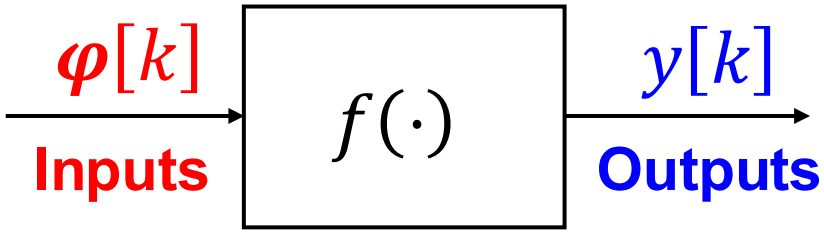


Static vs dynamical systems

Static systems



Dynamical systems



- For **static systems**, we will index the observations with the index i
- For **dynamical systems**, we will index the observations with the index k

k can be interpreted as the k-th sampling step

In either case, the aim will be **to learn $f(\cdot)$ from data**

- In the static case, we talk about (model) **"learning"**
- In the dynamical case, we talk about (system) **"identification"**

Both are supervised learning tasks!

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?

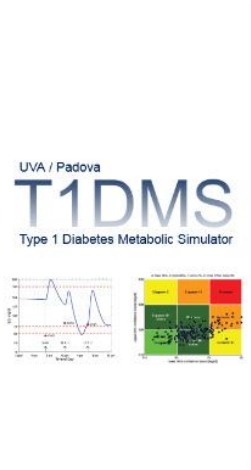
Prediction vs inference: prediction is not necessarily concerned with the structure of the model $\hat{f}(\cdot)$ and its complexity ($\hat{f}(\cdot)$ can be seen as a black-box) while inference uses the model to understand the relationship between each input and each output

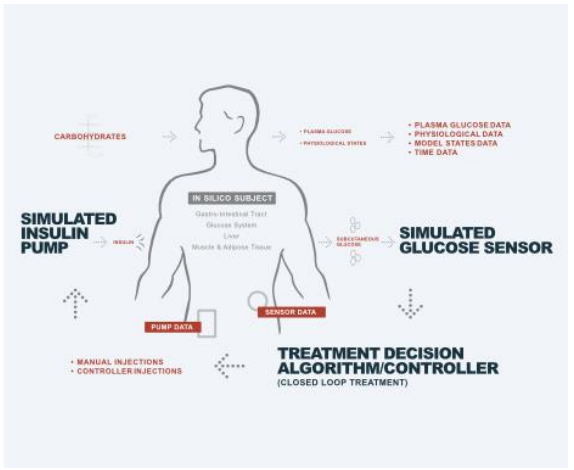
Why do we need models?

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



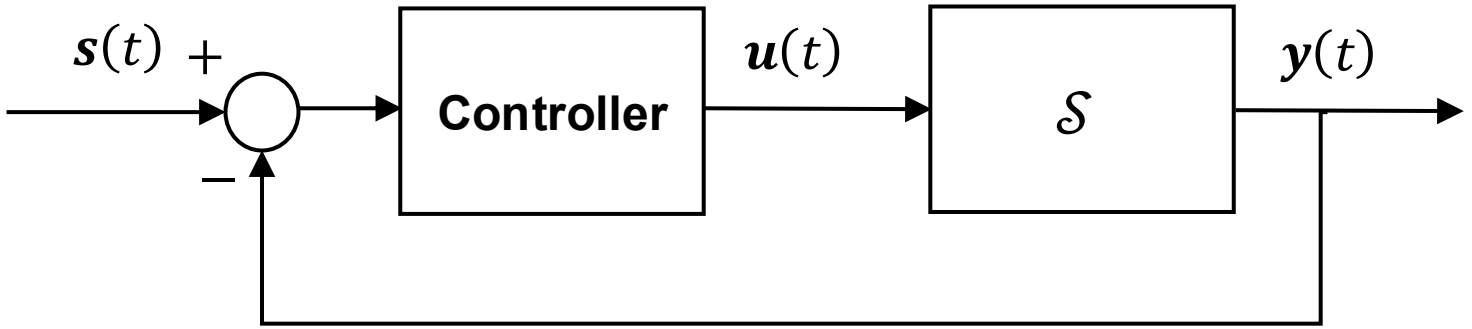


Why do we need models?

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 $\boldsymbol{s}(t)$ (setpoints)

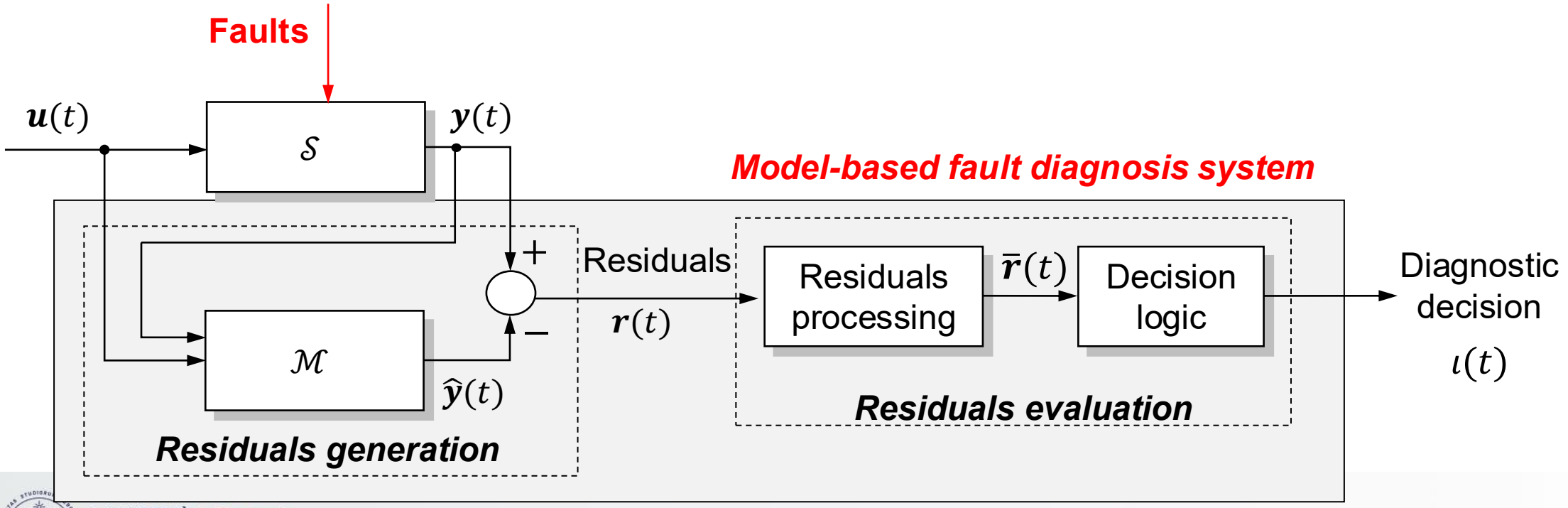


Why do we need models?

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



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.

UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione