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 1: Introduction to data science

SPEAKER
Prof. Mirko Mazzoleni

PLACE
University of Bergamo

Syllabus

1. **Introduction to data science**
   - 1.1 The business perspective
   - 1.2 Data analysis processes
2. **Data visualization**
3. **Maximum Likelihood Estimation**
4. **Linear regression**
5. **Logistic regression**
6. **Bias-Variance tradeoff**
7. **Overfitting and regularization**
8. **Validation and performance metrics**
9. **Decision trees**
10. **Neural networks**
11. **Machine vision**
    - 11.1 Classic approaches
    - 11.2 CNN and deep learning
12. **Unsupervised learning**
    - 12.1 k-means and hierarchical clustering
    - 12.2 Principal Component Analysis
13. **Fault diagnosis**
    - 13.1 Model-based fault diagnosis
    - 13.2 Signal-based fault diagnosis
    - 13.3 Data-driven fault diagnosis

Outline

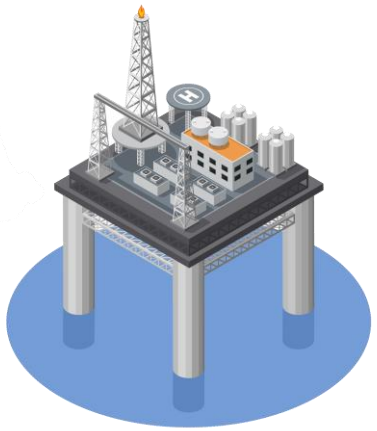
1. The data-driven company
2. Data types and usage
3. What we are going to do with data
4. From business problems to data-driven tasks
5. Supervised and unsupervised problems
6. Data science and machine learning processes

Outline

1. **The data-driven company**
2. Data types and usage
3. What we are going to do with data
4. From business problems to data-driven tasks
5. Supervised and unsupervised problems
6. Data science and machine learning processes

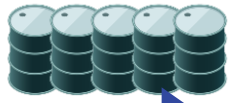
«Data is the new oil»

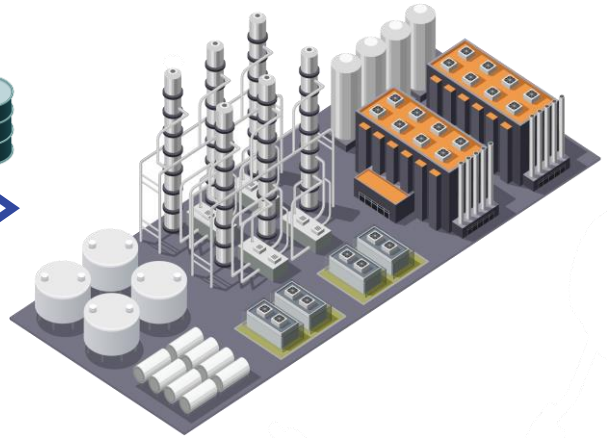
«Data is the new oil»



Crude oil extraction

Barrels of oil



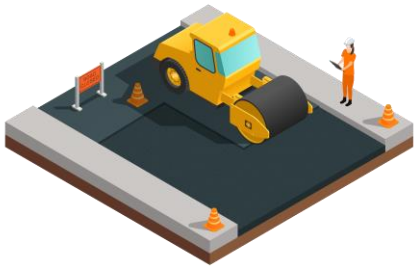


Refinement process

- Fuels
- Oils
- …

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

Asphalt

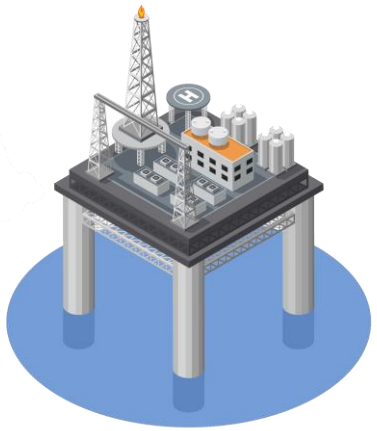


- Infrastructures
- Streets
- …

Plastic

- Bottles,
- Containers,
- Films,
- …

«Data is the new oil»

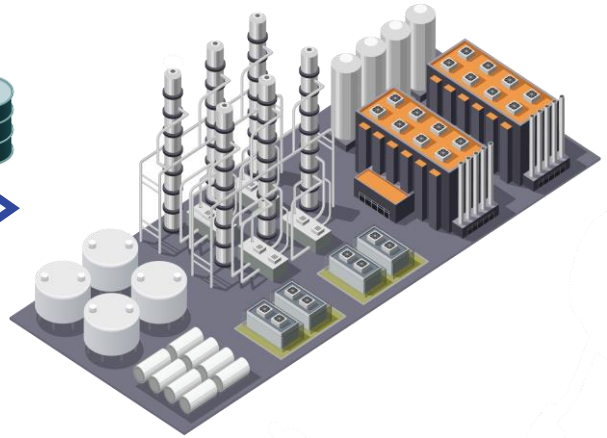


~~Crude oil~~ extraction

DATUM

~~Barrels of oil~~

DATA

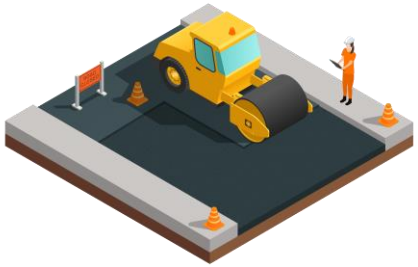


Refinement process

- Fuels
- Oils
- …

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

Asphalt

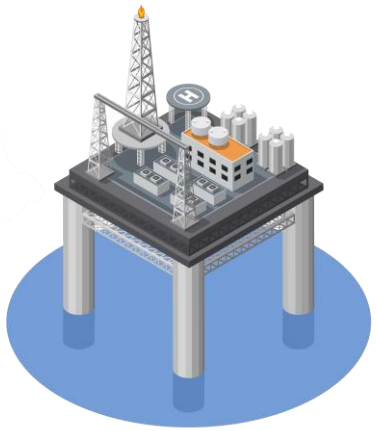


- Infrastructures
- Streets
- …

Plastic

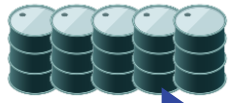
- Bottles,
- Containers,
- Films,
- …

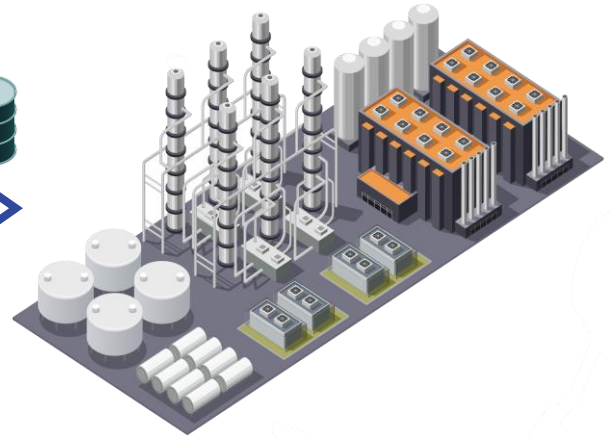
«Data is the new oil»



~~Crude oil~~ extraction **DATUM**

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



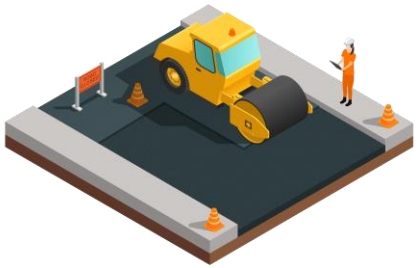


Refinement process

PRODUCT QUALITY

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

GOODS FORECAST

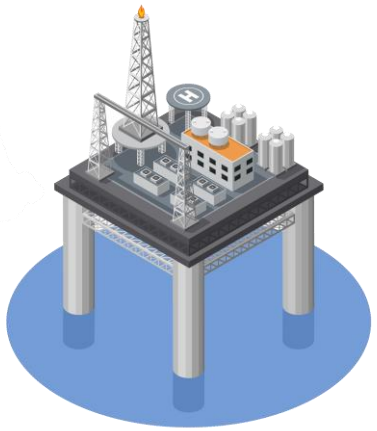


- Infrastructures
- Streets
- …

PROCESS OPTIMIZATION

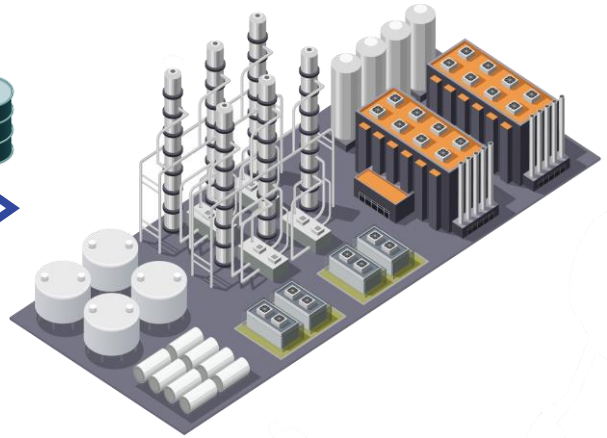
- Bottles,
- Containers,
- Films,
- …

«Data is the new oil»



Crude oil extraction **DATUM**

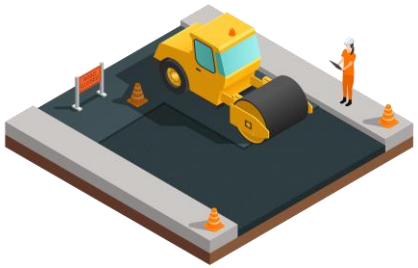
Barrels of oil **DATA**



Refinement process

PRODUCT QUALITY

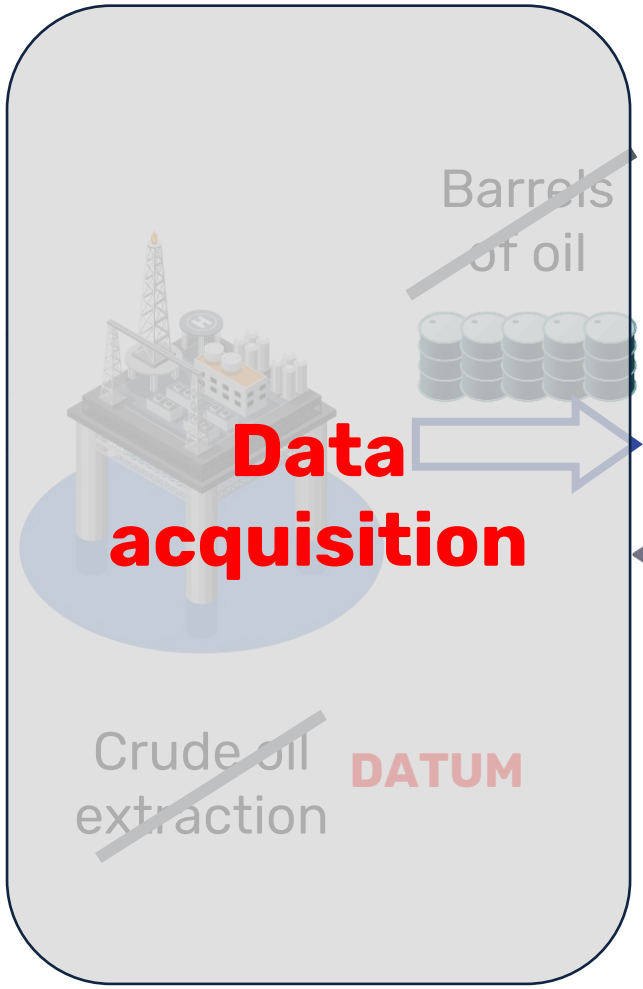
GOODS FORECAST



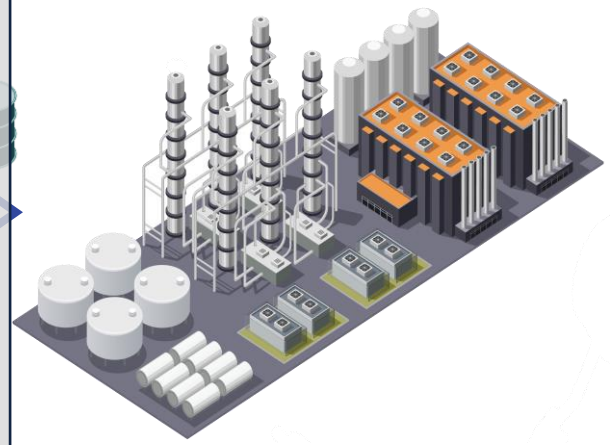
PROCESS OPTIMIZATION

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

«Data is the new oil»



DATA

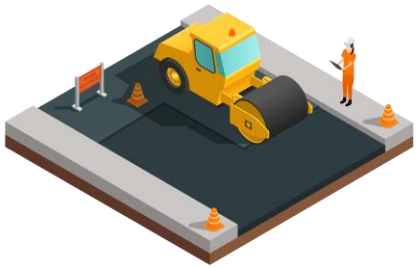


Refinement process

PRODUCT QUALITY

- Machine parameters optimization

GOODS FORECAST

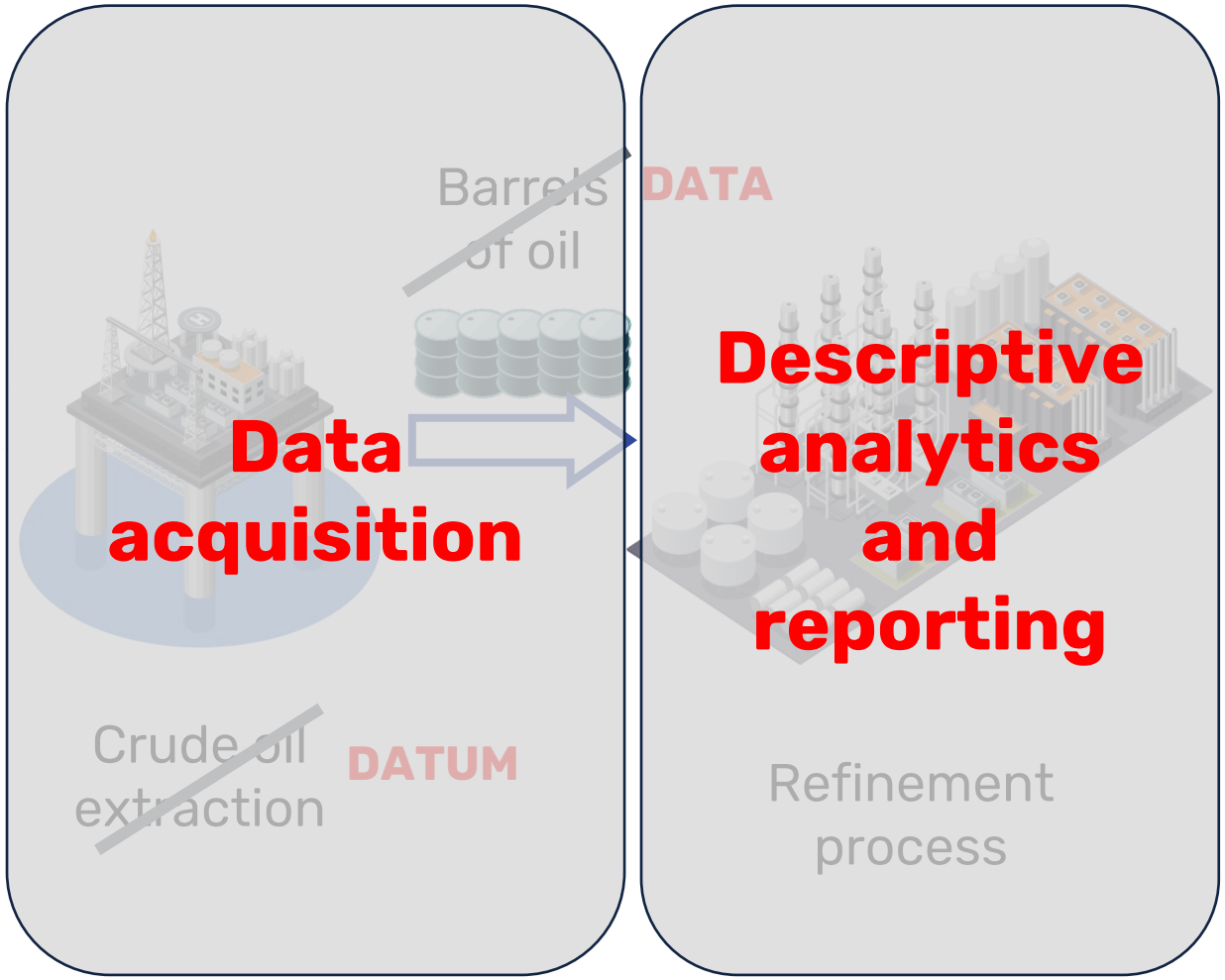


- Production and purchasing management

PROCESS OPTIMIZATION

- Reduction of materials used

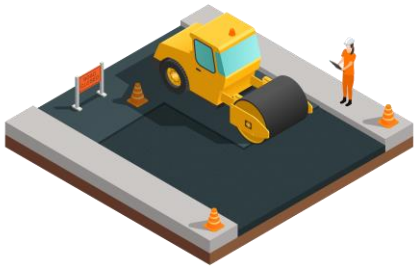
«Data is the new oil»



PRODUCT QUALITY

- **Machine parameters optimization**

GOODS FORECAST

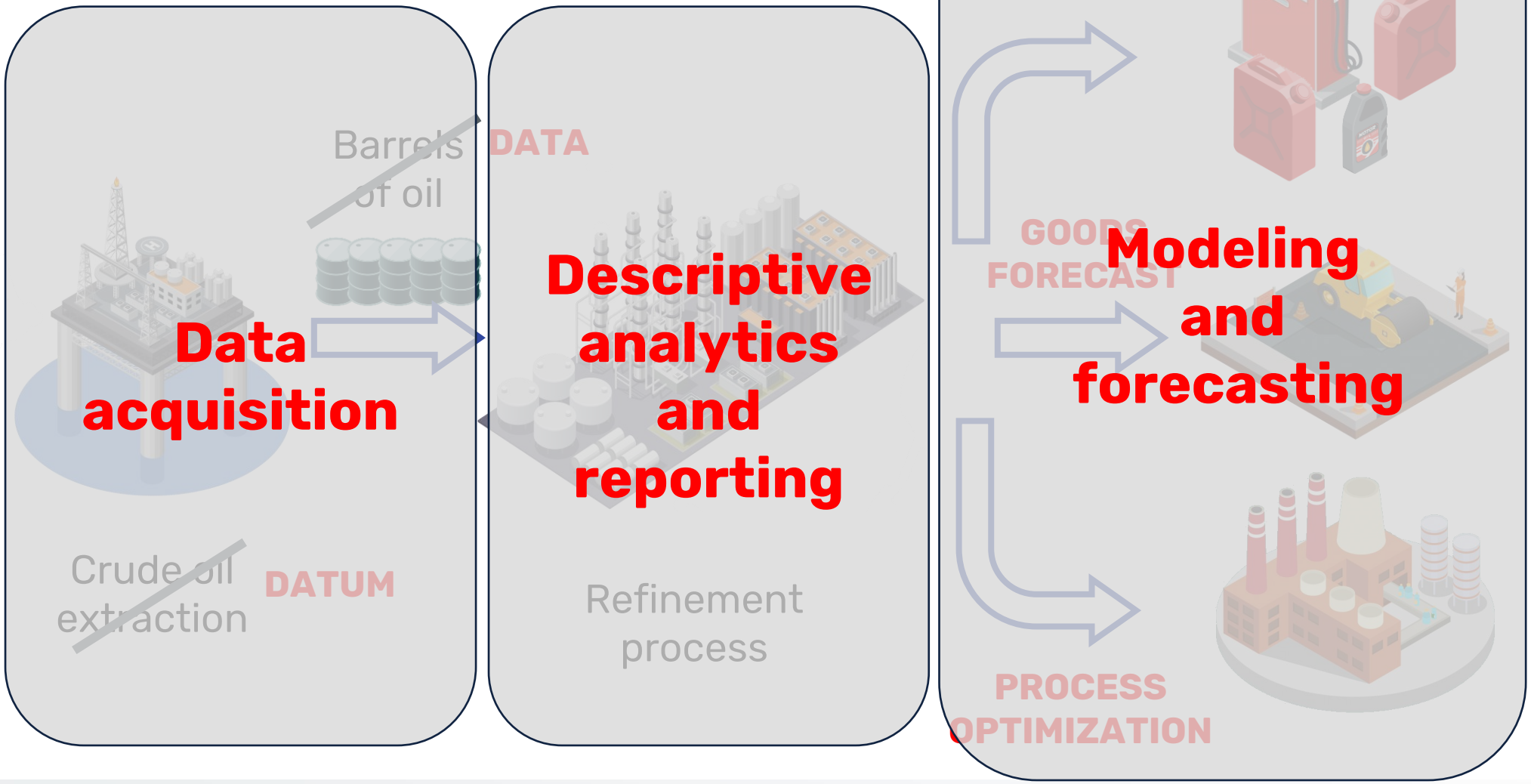


- **Production and purchasing management**

PROCESS OPTIMIZATION

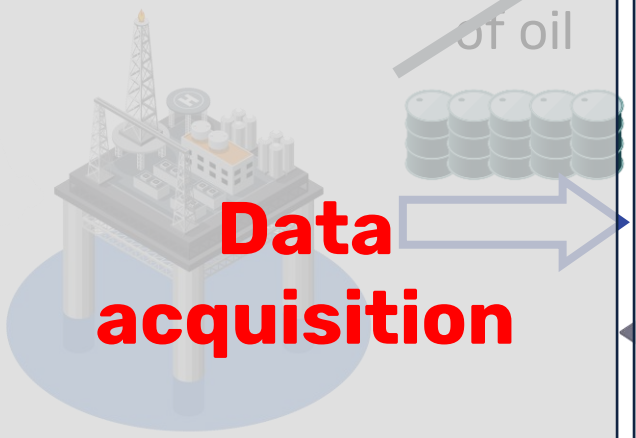
- **Reduction of materials used**

«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

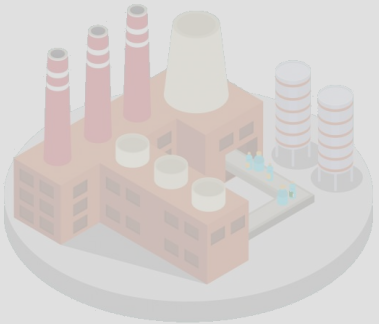
Refinement process

PRODUCT QUALITY

GOODS FORECAST

Modeling and forecasting

PROCESS OPTIMIZATION



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

Actions

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

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

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

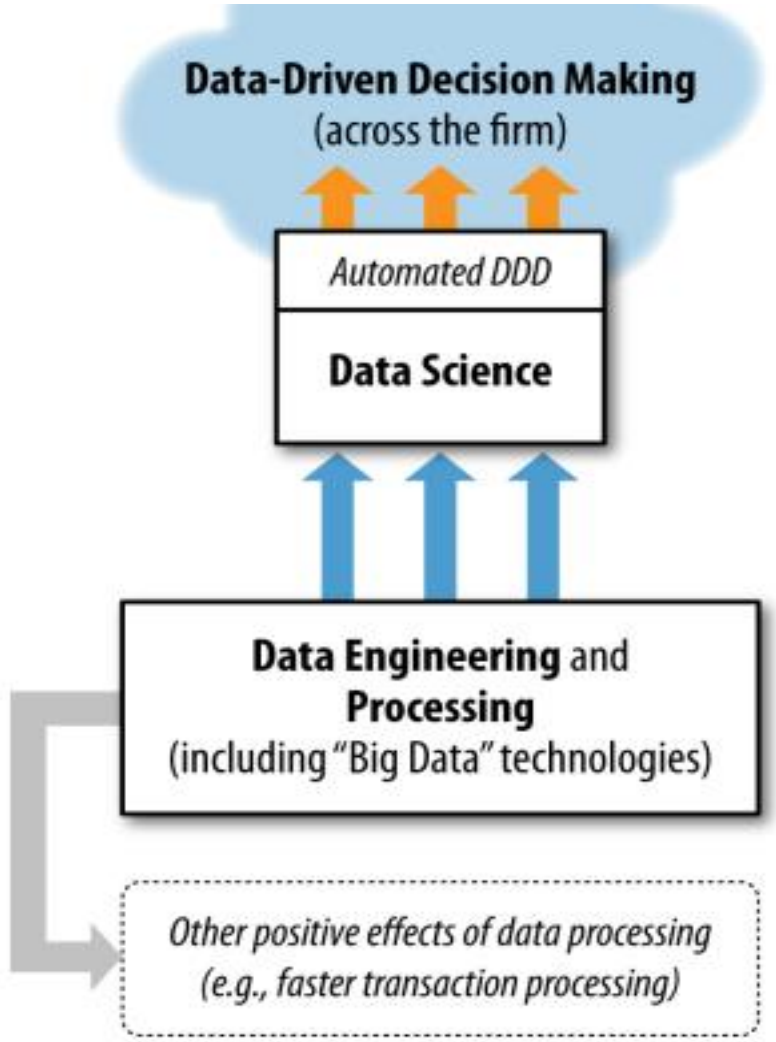
- Virtually every aspect of business is now open to **data collection** (operations, manufacturing, supply-chain management, customer behaviour, marketing campaigns)
- Collected information need to be **analyzed properly** in order to get **actionable results**
- A huge amount of data requires **specific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

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

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



Anti-hippo culture

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

—Jonathan Rosenberg

The road to become 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 on them

4

Data Informed

The 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

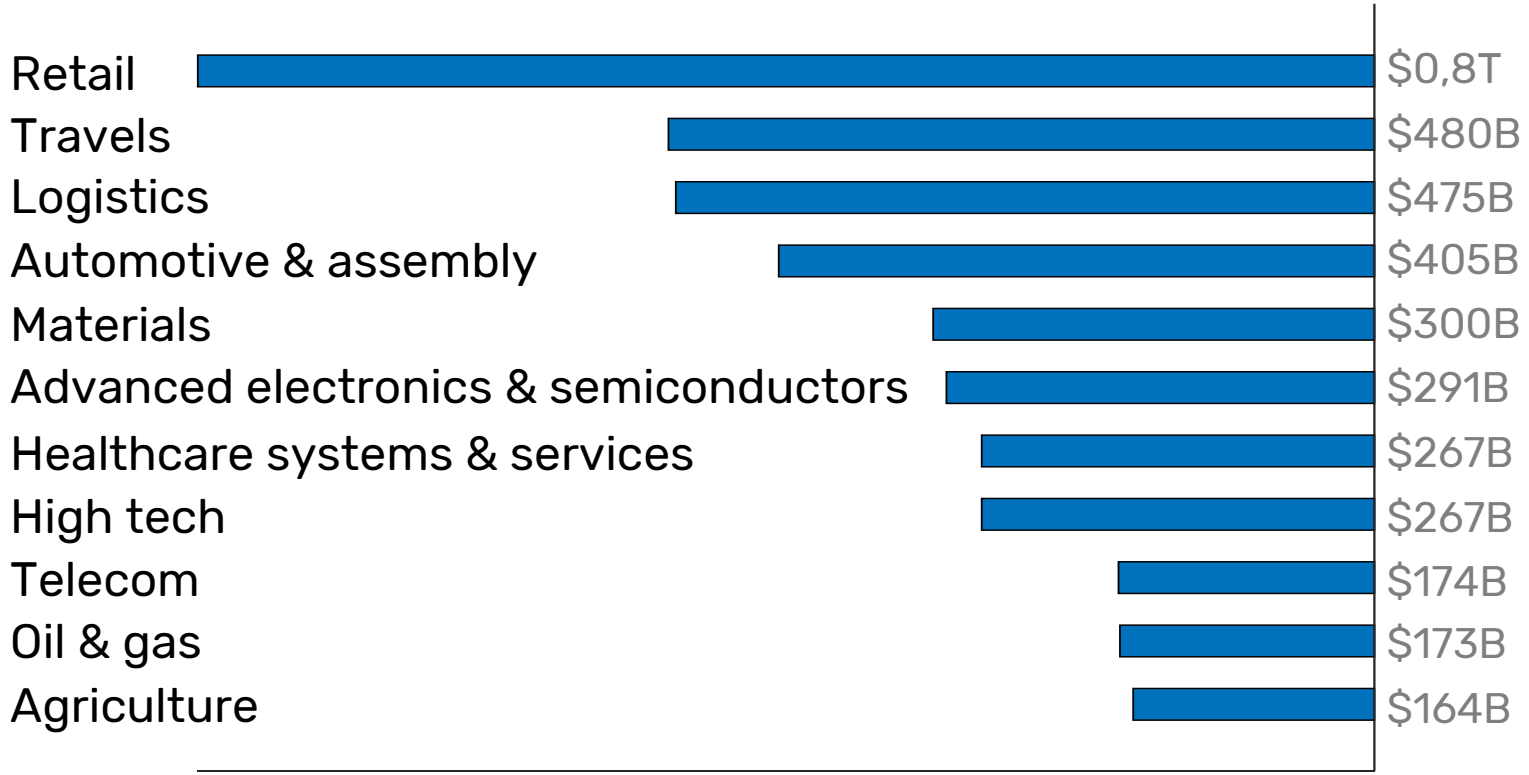
5% more productive [2]

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

Why become data-driven?

Business value created by Artificial Intelligence up to 2030 [4]

$13 Trillions



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

Outline

1. The data-driven company
2. **Data types and usage**
3. What we are going to do with data
4. From business problems to data-driven tasks
5. Supervised and unsupervised problems
6. Data science and machine learning processes

Data types

Data can have different formats. The most typical is that of a **table** (tabular data)

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

Data are dirty

Garbage IN, garbage OUT

Data problems:

- Missing values
- Not correct values

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

Structured data

Different data types

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

Organizing data

It is important to specify your **acceptance criteria**: when is our ML system good?

NO DEFECT

NO DEFECT

DEFECT

AIM: detect defects with 95% accuracy

Provide the data science team a dataset (often more than one) on which to **validate** their performance

Validation set

Organizing data

Training data

Input

Output

NO DEFECT

DEFECT

NO DEFECT

Estimate the model

Validation data

Real Output

NO DEFECT

NO DEFECT

DEFECT

Evaluate the model

Output of the algorithm

NO DEFECT

NO DEFECT

NO DEFECT

66,7% accuracy

Data problems and how to solve them

Issues that your source should solve

- Values are missing
- Zeros replace missing values
- Data are missing where you know should be there
- Rows or values are duplicated
- Spelling is incosistent
- Date fromats are incosistent
- Units are not specified
- Categories are badly chosen
- Field names are ambiguous
- Provenance is not documented
- Suspicious values are present
- Data are too coarse
- Totals differs from published aggregates
- Spreadsheet has 65536 rows
- Spreadsheet has dates in 1900, 1904, 1969 or 1970
- Text has been converted to numbers
- Numbers have been stored as text

Data problems and how to solve them

Issues that you should solve

- Text is garbled
- Line endings are garbled
- Data are in a PDF
- Data are too granular
- Data were entered by humans
- Data are intermingled with formatting and annotations
- Aggregations were computed on missing values
- Sample is not random
- Margin-of-error is too large
- Margin-of-error is unknown
- Sample is biased
- Data have been manually edited
- Inflation skews the data
- Natural\seasonal variation skews the data
- Timeframe has been manipulated

Data problems and how to solve them

Issues that a third-party expert should help you solve

- Author is untrustworthy
- Collection process is opaque
- Data assert unrealistic precision
- There are inexplainable outliers
- An index masks underlying variation
- Results have been p-hacked
- Benford's law fails (an observation that in many real-life sets of numerical data, the leading digit is likely to be small)
- Too good to be true

Data problems and how to solve them

Issues that a programmer should help you solve

- Data are aggregated to the wrong categories or geographies
- Data are in scanned documents

Outline

1. The data-driven company
2. Data types and usage
3. **What we are going to do with data**
4. From business problems to data-driven tasks
5. Supervised and unsupervised problems
6. Data science and machine learning processes

Data usages

With data we can perform several analyses, such as:

- **Descriptive** analysis
- **Model learning**
- **Visualization**

In this course we will see all these aspects

Descriptive analysis

Descriptive analysis consists mainly in:

- **Summarize the data** (usually tabular data) and compute quantities such as averages, standard deviations
- **Check for missing values**, and try to understand **why they are missing**
- **Visualize the data** using different plot types, colors and formats
- **Counting** how many data satisfy a certain condition
- **Answer,** and provide a way to **generate,** questions about the data

Statistical Modeling: The Two Cultures

Modeling entails estimating models (functions) of the natural data-generating systems [6]

There are two goals in analyzing the data with a model:

1. **Information:** to extract some information about how nature is associating the response variables to the input variables **«Statistical approach»**

2. **Prediction:** to be able to predict what the responses are going to be given future input variables **«Machine learning approach»**

Statistical Modeling: The Two Cultures

There are two different approaches toward these goals:

1. The **«statistical approach»:** the focus is to get an interpretable model that relies on data and model assumptions

2. The **«machine learning approach»:** the focus is on prediction performance rather than interpretation. Assumptions on data and their distribution are less important

In this course we will cover both approaches

Model learning

Model learning consists in **estimating a model** from data. Consider the following data:

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

- **AIM:** predict house prices

Regression

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

X

X

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

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

Model learning

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

Picture	Label
	Cat
	Not cat
	Cat
	Not cat

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

Classification

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

Machine learning (ML) vs. data science

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

X

y

Machine learning

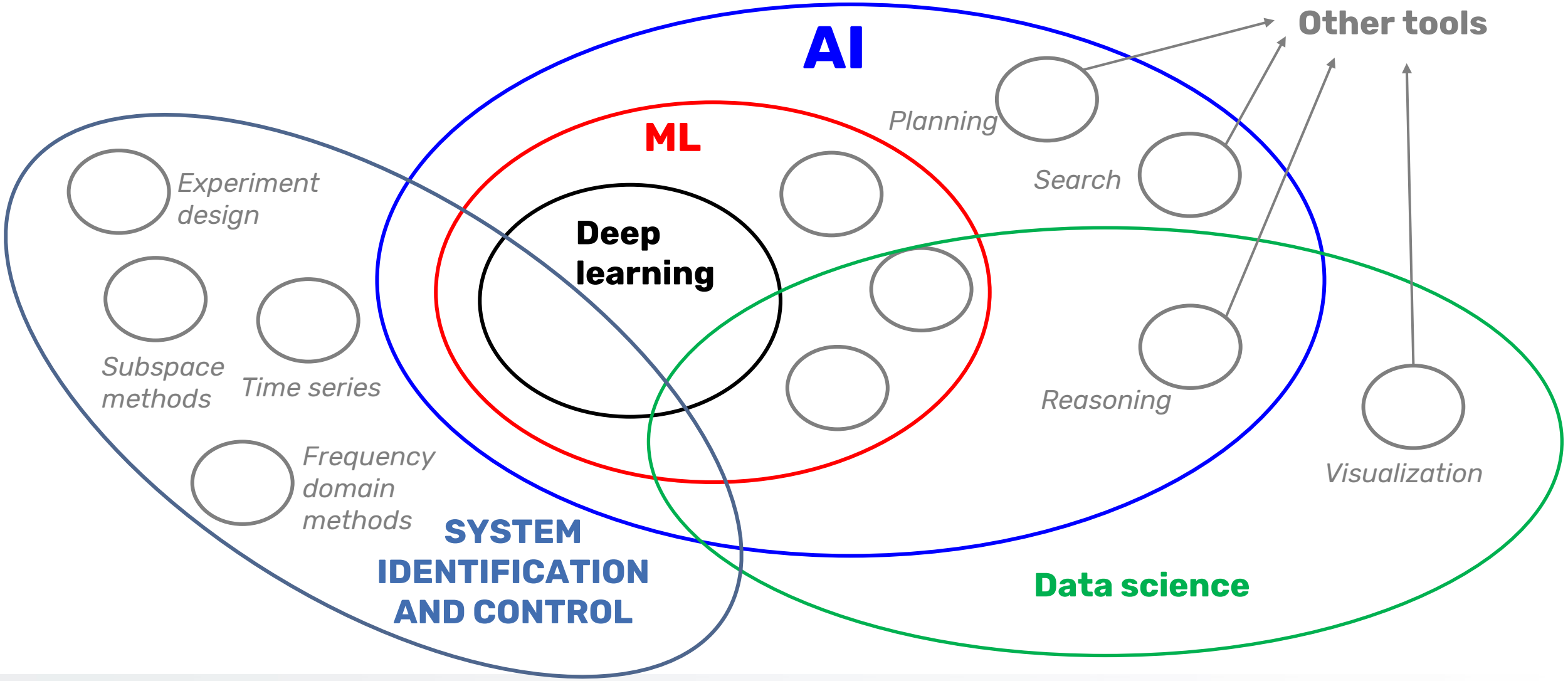
- Predict y given X
- Running software (web site\ mobile app)

→ **Output:** Code and program

Data science

- Houses with 3 bathrooms are more expensive than those with 2 bathrooms of the same size
- Recently renovated houses cost 15% more → **Output:** Slide deck

ML, data science, AI and dynamical systems



Outline

1. The data-driven company
2. Data types and usage
3. What we are going to do with data
4. **From business problems to data-driven tasks**
5. Supervised and unsupervised problems
6. Data science and machine learning processes

Business problems as data-driven tasks

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

There are many data science algorithms. However, they address a **handful** of tasks:

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

Business problems as data-driven tasks

- Spam e-mail detection system Classification
- Credit approval Classification
- Recognize objects in images Classification
- Find the relation between house prices and house sizes Regression
- Predict the stock market Regression
- Market segmentation Clustering
- Market basket analysis Co-occurence grouping
- Language models (word2vec) Similarity matching
- Social network analysis Link prediction
- Low-order data representations Data reduction
- Movies recommendation Similaritiy matching

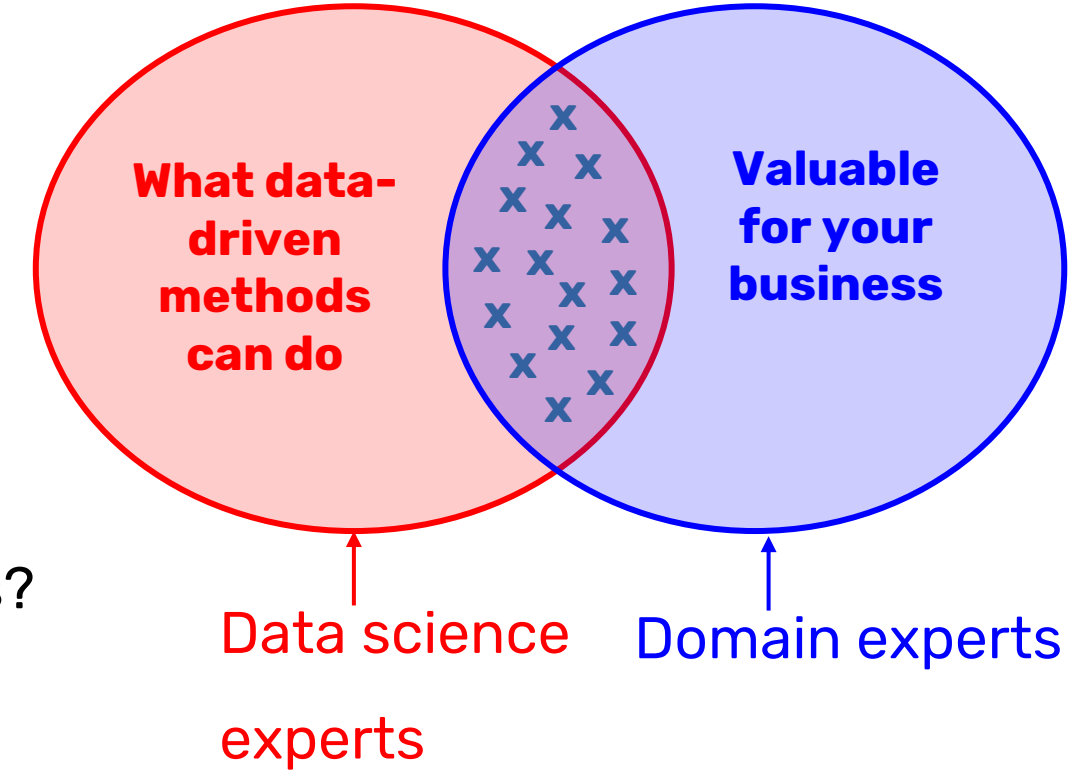
Selecting data-driven projects

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

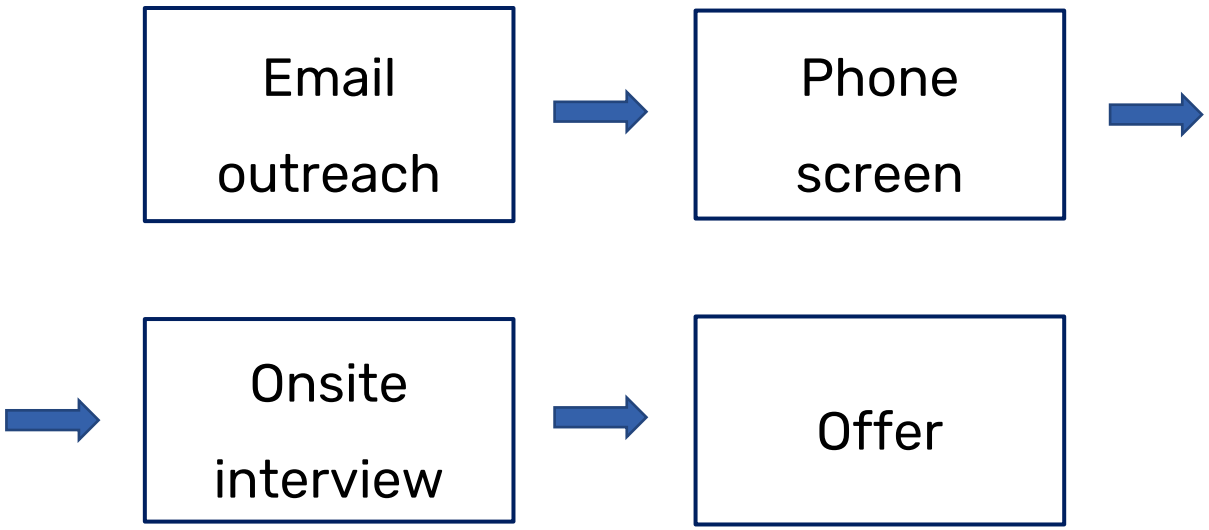
Machine learning

- Automatic visual inspection

Selecting data-driven projects

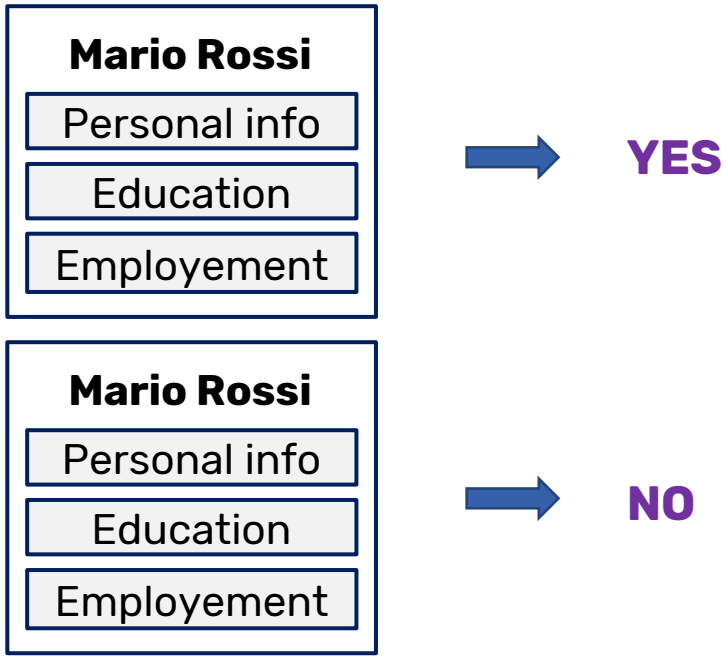
RECRUITING

Data science



- Optimize recruiting process

Machine learning

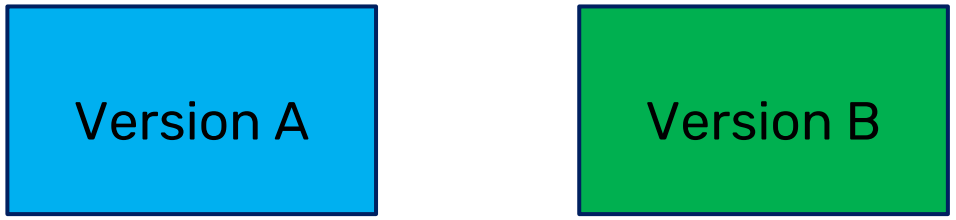


- Automatic resume screening

Selecting data-driven projects

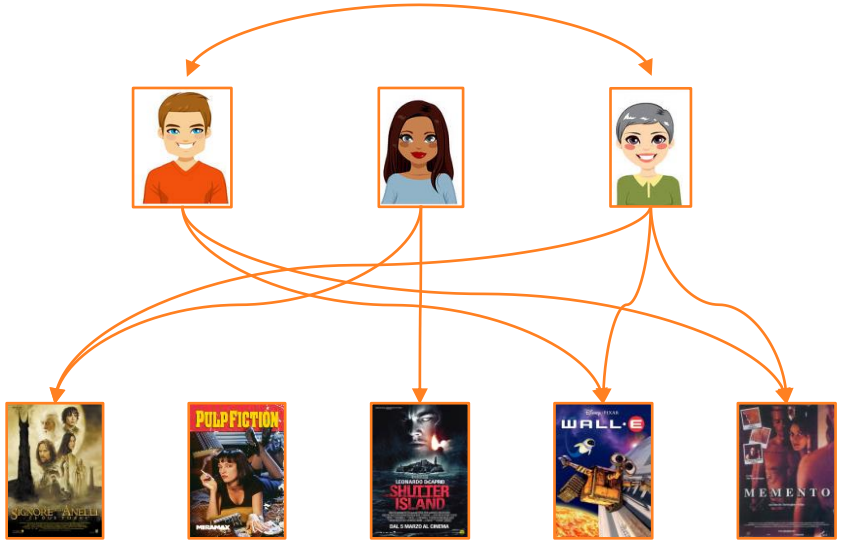
MARKETING

Data science



- A \ B testing websites

Machine learning



- Recommendations

Outline

1. The data-driven company
2. Data types and usage
3. What we are going to do with data
4. From business problems to data-driven tasks
5. **Supervised and unsupervised problems**
6. Data science and machine learning processes

Supervised vs unsupervised methods

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

Unsupervised

«Do our customers naturally fall into different groups?»

*There is not a specific purpose for the grouping. The aim is only to find **similarities** between individuals*

Supervised

«Can we find groups of customers who have particularly high likelihoods of canceling their service soon after their contracts expire? »

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

Supervised vs unsupervised methods

- Classification and class probability estimation
- Regression
- Causal modeling

Supervised

- Symilarity matching
- Link prediction
- Data reduction

Supervised or Unsupervised

- Clustering
- Co-occurrence grouping
- Profiling

Unsupervised

Supervised vs unsupervised methods

Supervised

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

Unsupervised

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

- Movies recommendation

Supervised or unsupervised

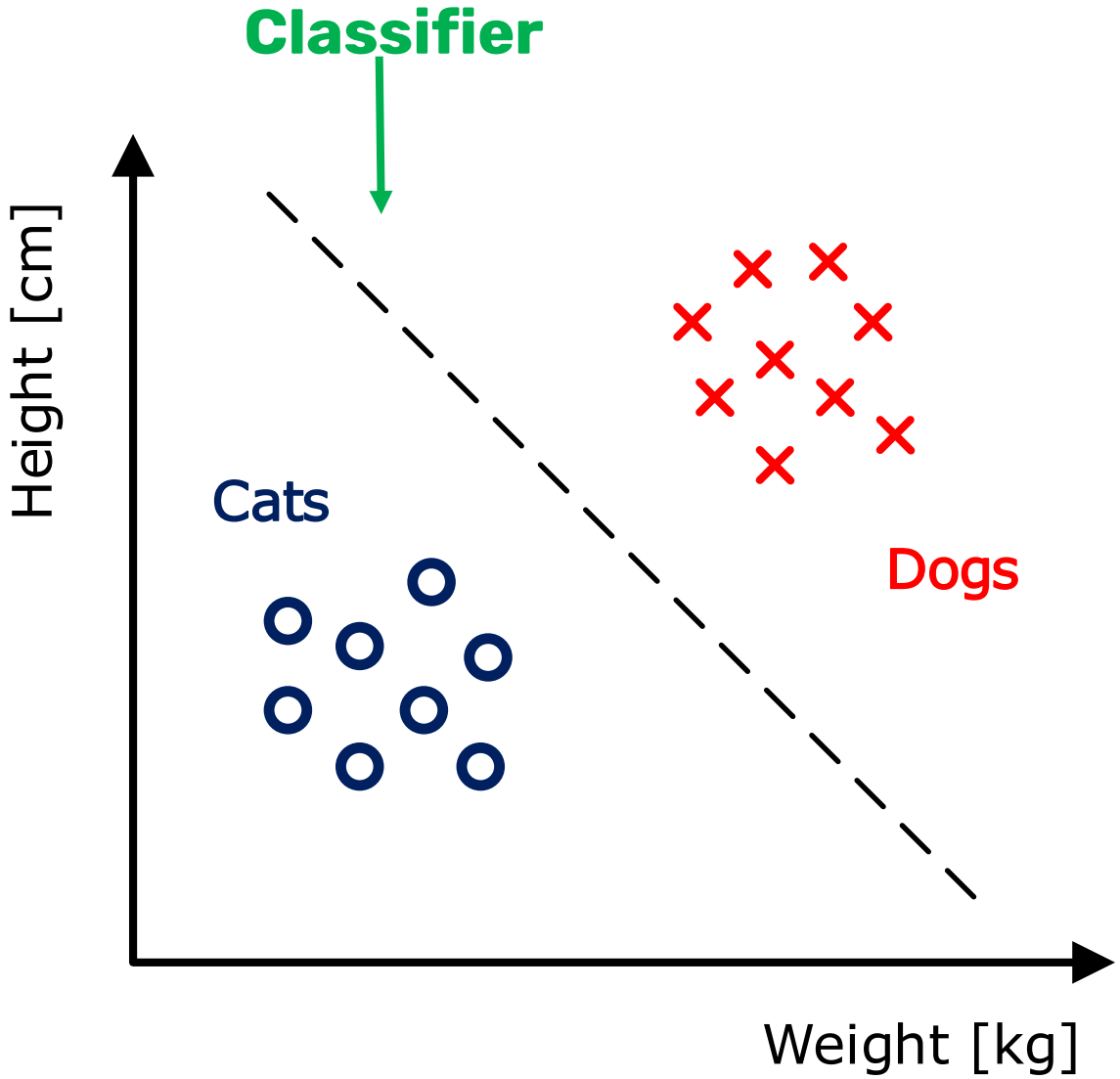
Components of learning (in general)

- Input: $\boldsymbol{x}$ (e-mail textual content) → each dimension is some e-mail attribute
- Output: y (spam / not spam?) → the decision that we have to take in the end
- Target function: $f: \mathcal{X} \to \mathcal{Y}$ (Ideal spam filter formula) → unknown, we have to learn it
- Data: $\mathcal{D} = \{(\boldsymbol{x}(1), y(1)), \ldots, (\boldsymbol{x}(N), y(N))\}$ (historical records of e-mail examples)
  - ✓ Each **feature vector** $\boldsymbol{x}$ consists of different **regressors** or **features**, i.e. information used to predict the output variable
- Hypothesis: $g: \mathcal{X} \to \mathcal{Y}, g \in \mathcal{H}$ (formula to be used) → g is an **approximation** of f

$\mathcal{H}$ is called the **Hypothesis space**. This, together with the **Learning algorithm**, forms the **learning model** [5]

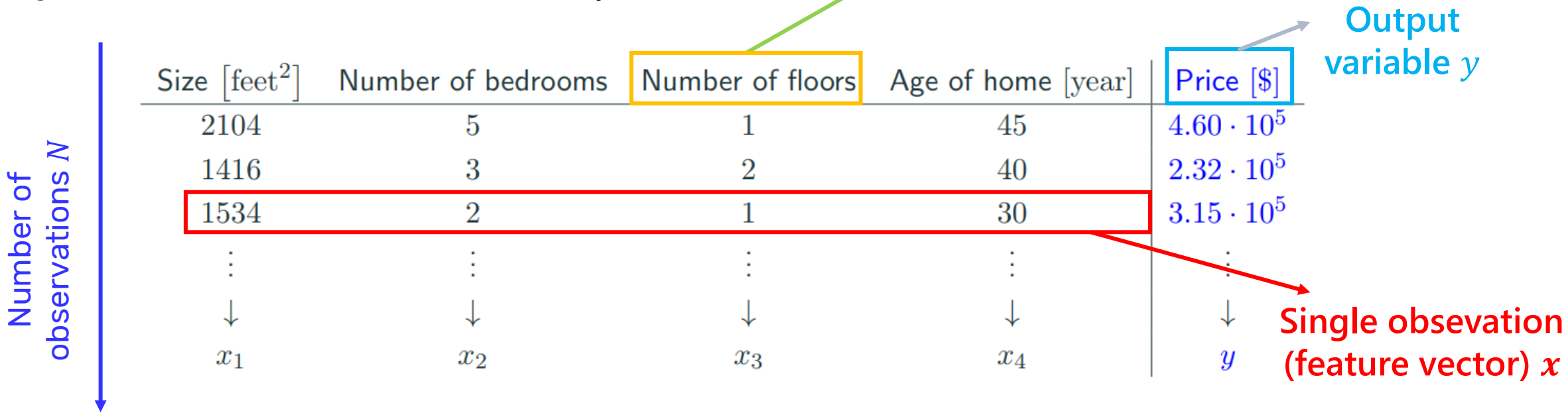
Supervised learning

- The «correct answer» (output label) y is given
- Predict y from a set of inputs $\boldsymbol{x} \in \mathbb{R}^{d \times 1}$
- **Regression:** predict a continuous output $y \in \mathbb{R}$ **(real value)**
- **Classification:** predict a discrete categorical output $y \in \{1,2,\dots,C\}$ **(class)**



Example: house prices regression

Suppose we want to find a linear function which relates the **measured** regressors x_1, x_2, x_3, x_4 with the **observed** output y



Size [feet^2]	Number of bedrooms	Number of floors	Age of home [year]	Price [$]
2104	5	1	45	$4.60 \cdot 10^5$
1416	3	2	40	$2.32 \cdot 10^5$
1534	2	1	30	$3.15 \cdot 10^5$
⋮	⋮	⋮	⋮	⋮
↓	↓	↓	↓	↓
x_1	x_2	x_3	x_4	y

- The number of rows is the number of data points (also known as number of observations) N
- The i-th observation is the vector $\boldsymbol{x}(i) = [x_1(i)\ \ x_2(i)\ \ x_3(i)\ \ x_4(i)]^\top \in \mathbb{R}^{4x1}$
- Each feature vector $\boldsymbol{x}$ has associated a response $y \in \mathbb{R}$ that we want to predict for new observations $\boldsymbol{x}^*$

Example: house prices classification

The components of the features vector are the same. The difference lies in the response variable, which now is a **class** (categorical data type) and not a real value

Suppose that instead of the price value in dollars, we want to classify houses as **expensive** *(class* $y = 1$*)* or **cheap** *(class* $y = 0$*)*

Size $[\text{feet}^2]$	Number of bedrooms	Number of floors	Age of home $[\text{year}]$	Price $[\text{class}]$
2104	5	1	45	1
1416	3	2	40	0
1534	2	1	30	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
x_1	x_2	x_3	x_4	y

The point $\boldsymbol{x}$ is classified to class $y = 1$ if *the probability of* $\boldsymbol{x}$ *to belong to class 1 is* ≥ 0.5

Supervised learning: problem statement

The aim is to **learn an unknown function** f given a dataset $\mathcal{D}$

- The function is searched in the hypothesis space $\mathcal{H}$, where $h \in \mathcal{H}$ is a specific function
- We want to find a function h that approximates f well, on the **whole domain** $\mathcal{X}$

What does $h \approx f$ mean?

- We need to define a **measure of error** or **cost (loss) function**

Supervised learning: problem statement

Pointwise error measures

Pointwise error measures $\ell(\boldsymbol{x}; \boldsymbol{\theta})$ are based on a single point $\boldsymbol{x}$. Examples are:

- **Quadratic error:** $\ell\big(f(\boldsymbol{x}), h(\boldsymbol{x}; \boldsymbol{\theta})\big) = \big(f(\boldsymbol{x}) - h(\boldsymbol{x}; \boldsymbol{\theta})\big)^2$ → used for regression

- **Binary error:** $\ell\big(f(\boldsymbol{x}), h(\boldsymbol{x}; \boldsymbol{\theta})\big) = \mathbb{I}\{f(\boldsymbol{x}) \neq h(\boldsymbol{x}; \boldsymbol{\theta})\}$ → used for classification

Supervised learning: problem statement

Global error measures

These error measures take into account all N observations. It is important to distinguish between **in-sample error** (train error) and **out-of-sample error** (validation or test error).

In-sample error

Error that the model makes on the observed N **data available**, which were used to estimate it

$$E_{\text{in}}\big(h(\boldsymbol{\theta})\big) \equiv J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \ell\big(f(\boldsymbol{x}), h(\boldsymbol{x};\boldsymbol{\theta})\big)$$

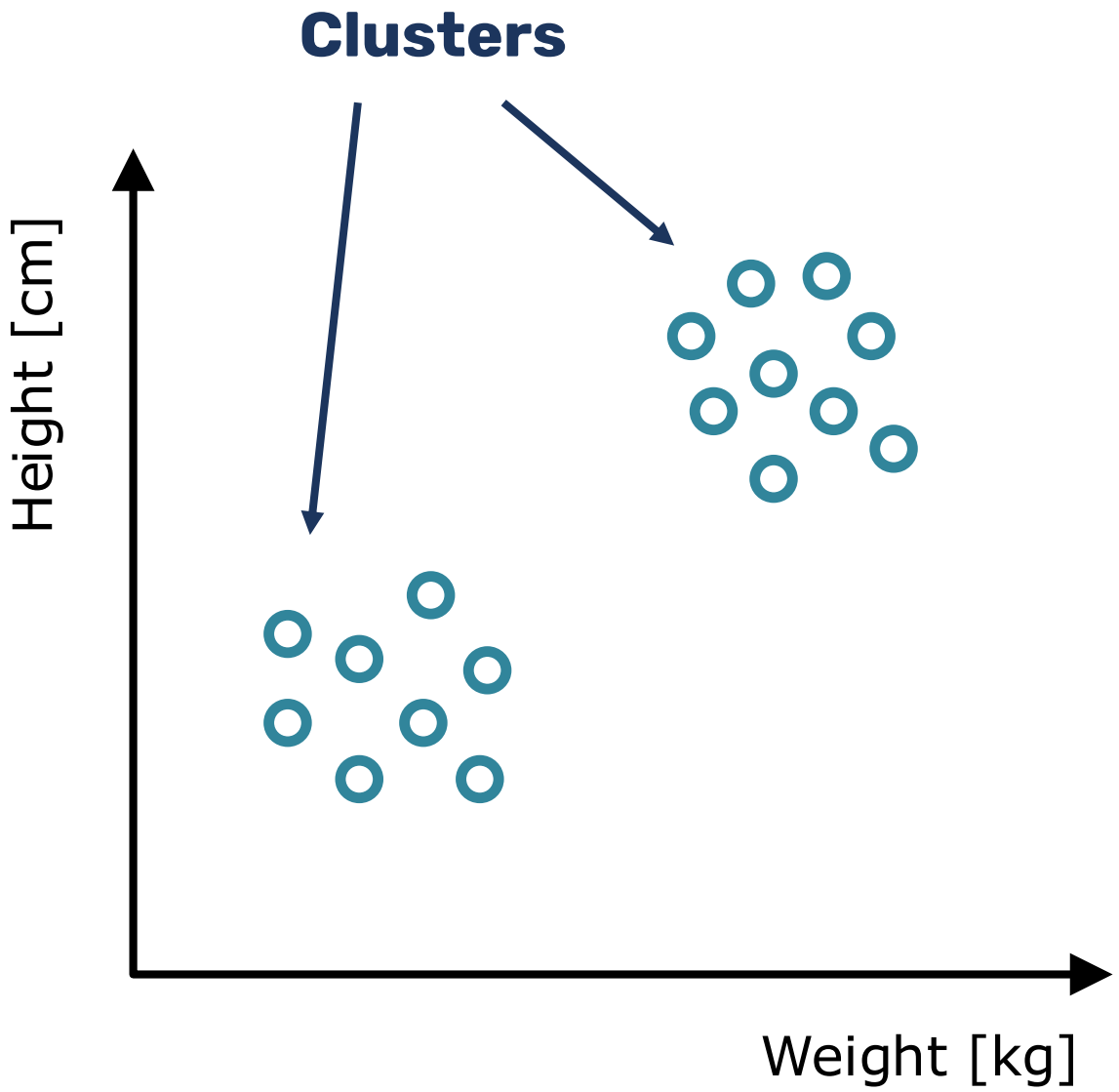
Out-of-sample error

Error that the model makes on the entire domain of f (therefore also **data that I have not observed**)

$$E_{\text{out}}\big(h(\boldsymbol{\theta})\big) = \mathbb{E}_{\boldsymbol{x}}\big[\ell\big(f(\boldsymbol{x}), h(\boldsymbol{x};\boldsymbol{\theta})\big)\big]$$

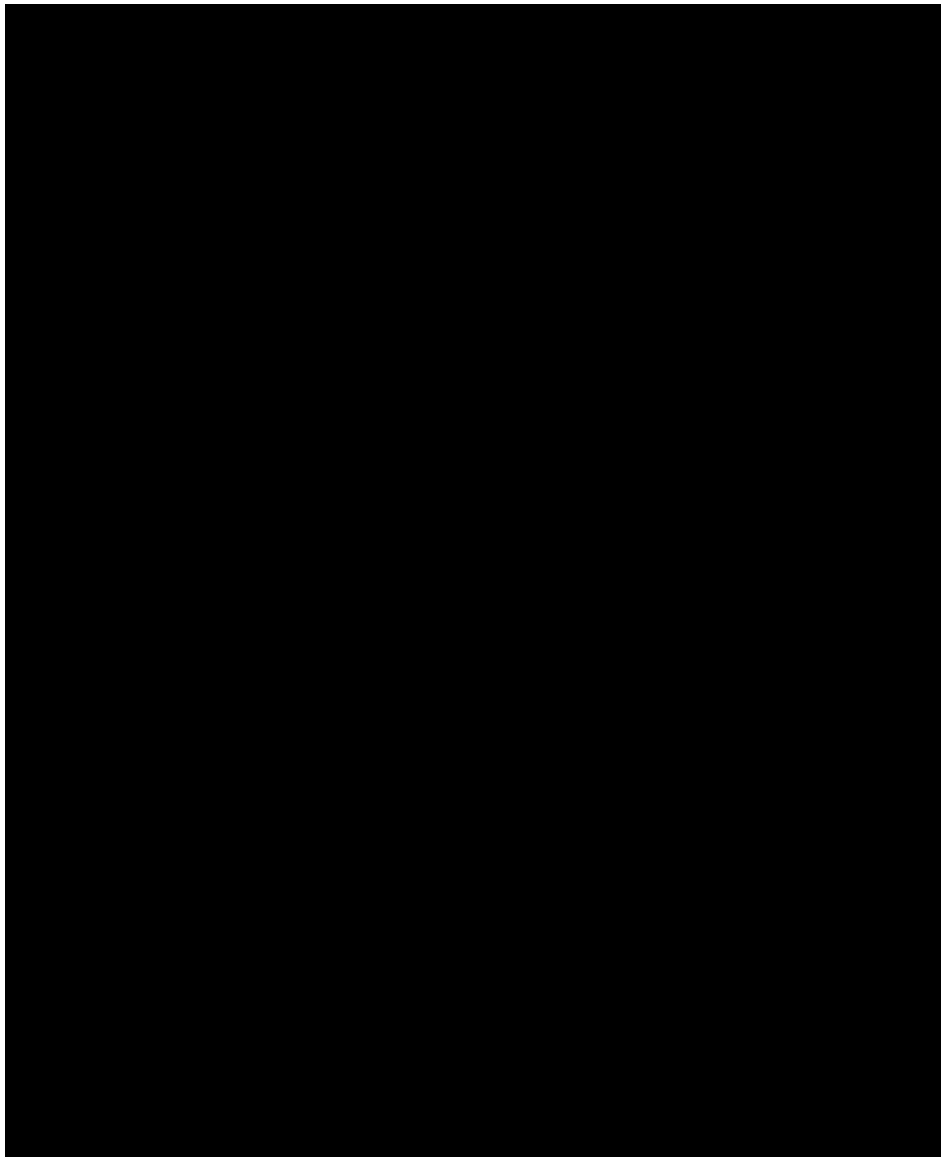
Unsupervised learning

- Instead of **(input, output)** we get **(input, ?)**
- Here there is no a function f to learn
- Find properties of the inputs $\boldsymbol{x} \in \mathbb{R}^{d \times 1}$
- High-level representation of the input
- Elements in the **same cluster** have **similar properties**



Reinforcement learning

- Instead of **(input, output)** we get **(input, output, reward)**
- The algorithm tries to learn what action to take, in order to maximize the reward
- This is called a policy
- Applications in control, robotics, A/B testing



Business problems as data science examples - revisited

Supervised

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

Unsupervised

- Market segmentation **Clustering**
- Market basket analysis **Co-occurence grouping**
- Language models (word2vec) **Similarity matching**
- Social network analysis **Link prediction**
- Low-order data representations **Data reduction**

Supervised or unsupervised

- Movies recommendation **Similaritiy matching**

Outline

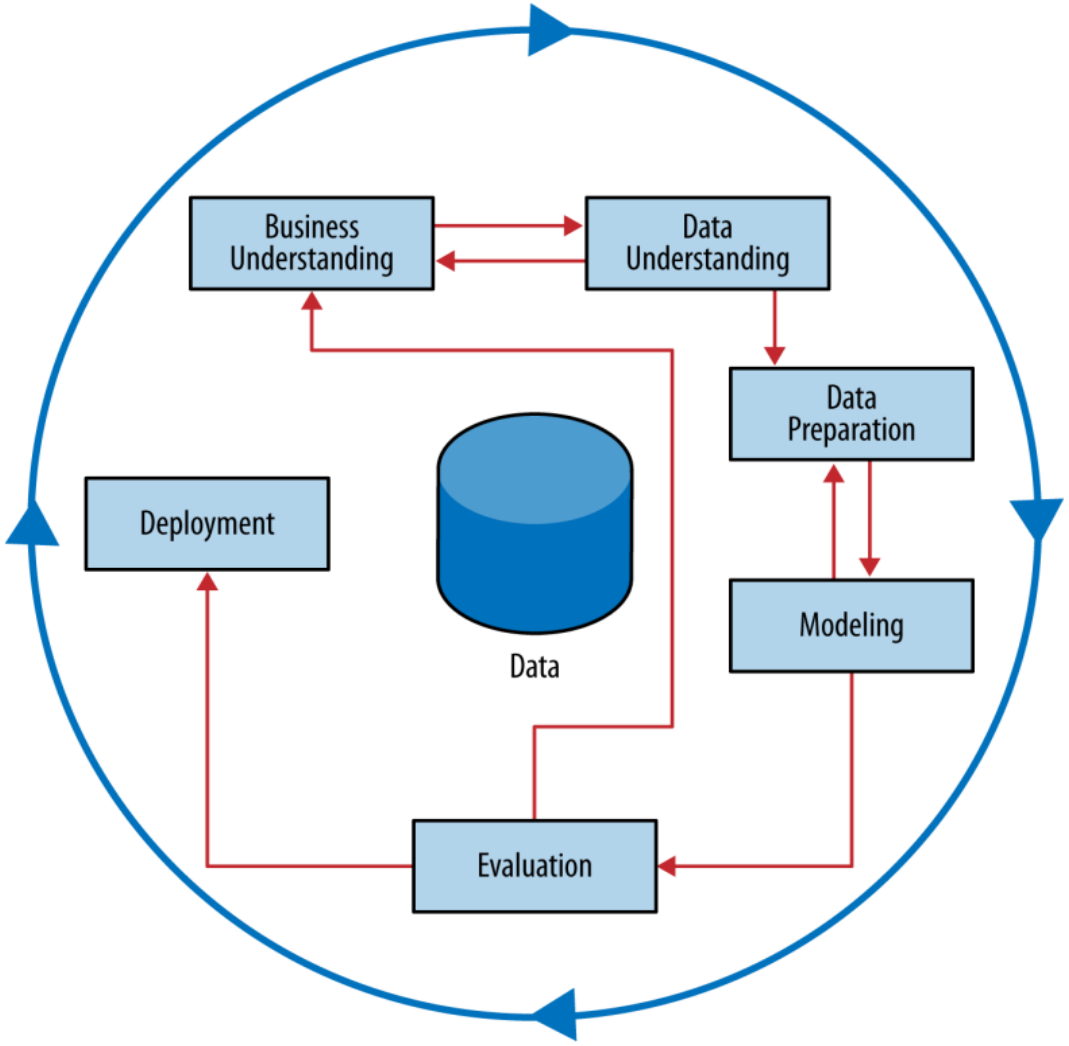
1. The data-driven company
2. Data types and usage
3. From business problems to data-driven tasks
4. Supervised and unsupervised problems
5. **Data science and machine learning processes**

CRISP-DM process

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

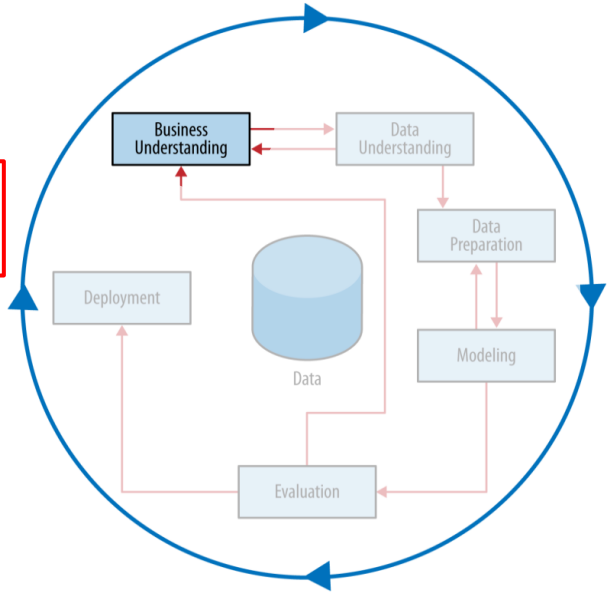
Iteration is the rule rather the exception:

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



CRISP-DM: Business understanding

Cast the business problems into one or more data science problems



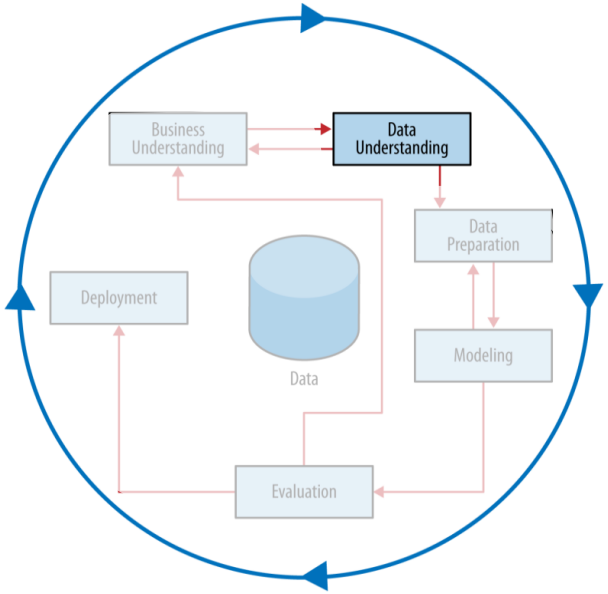
1. Classification
2. Regression
3. Probability estimation
4. Similarity matching
5. Clustering
6. Profiling
7. Link prediction
8. Data reduction
9. Causal modeling

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 science models?**

CRISP-DM: Data understanding

Identify the available and needed data



Costs/benefits of acquiring each source of data

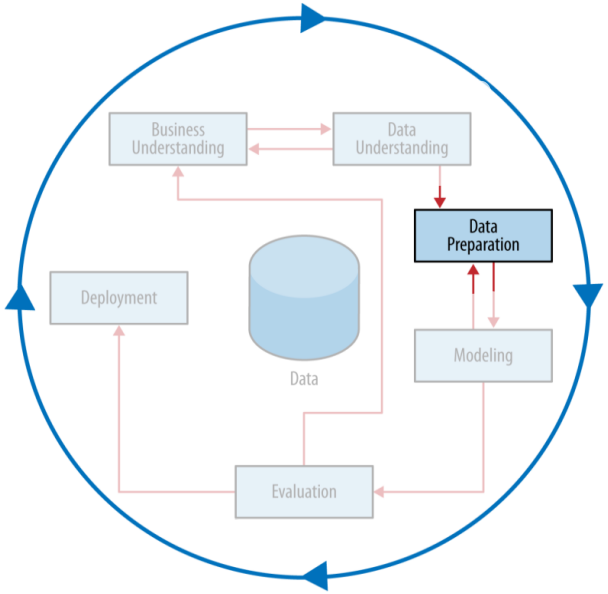
Are the data at disposal **related to the business** problem?

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

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

CRISP-DM: Data preparation

Clean and prepare data for use with algorithms



Usually, the algorithms we employ require **data in a different format** with respect to the available one

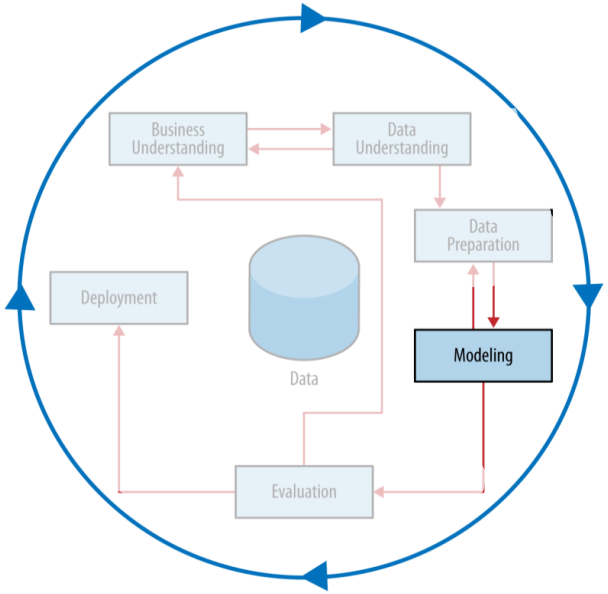
- Convert string to numbers, infer missing data, import data from excel files, …

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

Pay attention to not use historical data that **will not be available** when your model will be used

CRISP-DM: Modeling

Estimate a mathematical model to extract pattern from data



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

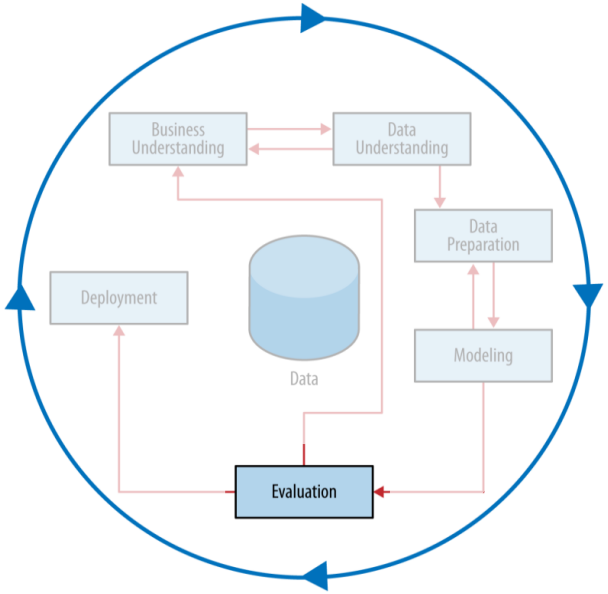
The aim is to find a model that performs well **on unseen data**

The type of the model has to be chosen based on:

- What data mining **task** we want to solve
- Performance **measures**
- Availability of **libraries** for deployement

CRISP-DM: Evaluation

Assess the validity of the results



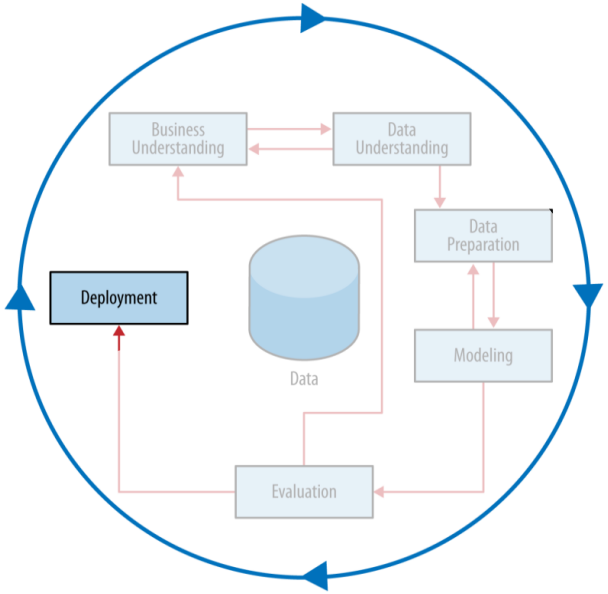
We could find patterns that exist only in the particular dataset that we have at disposal **(overfitting)**

The devised solution and the model's decisions should 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 the existing technology

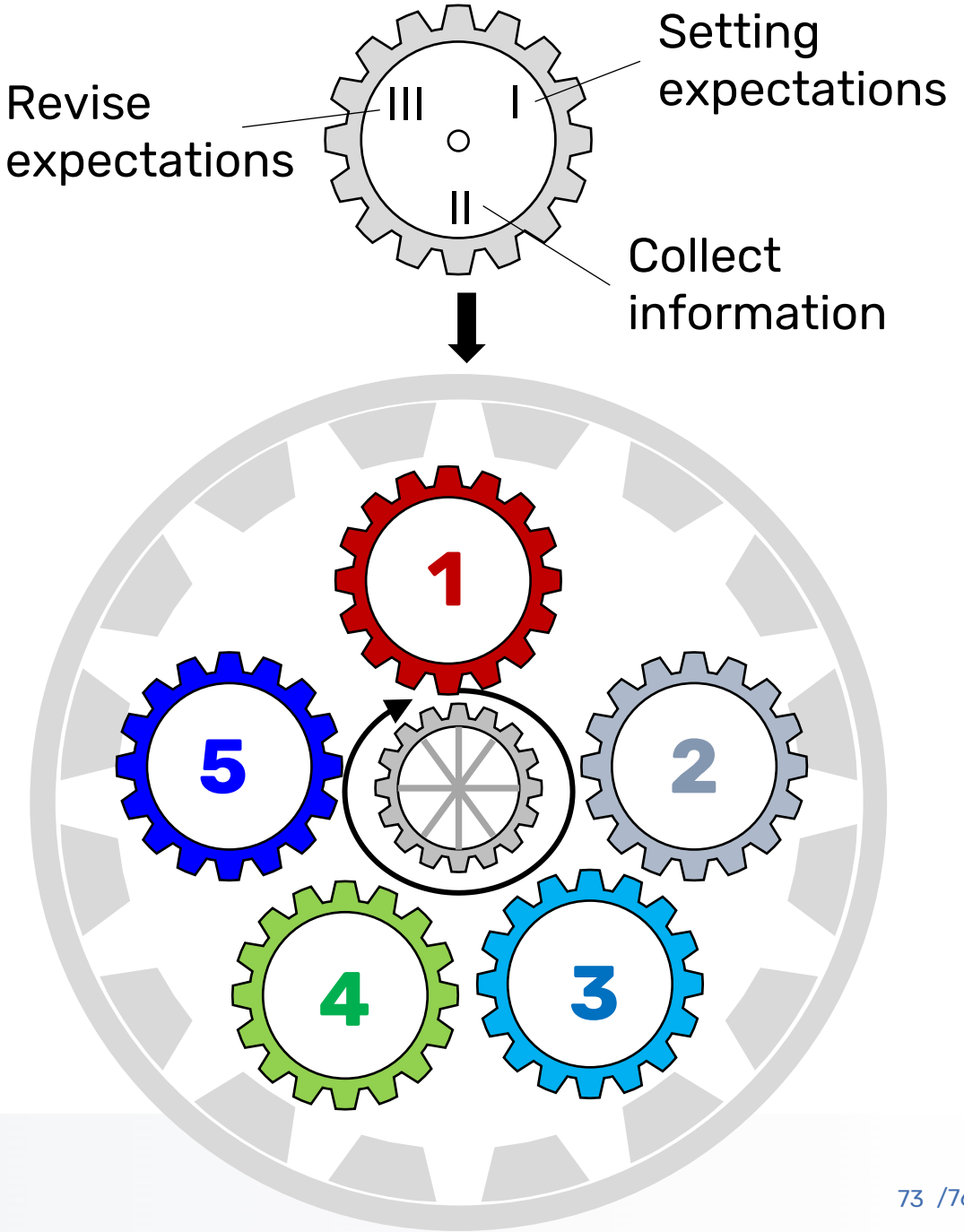
This step can require a quite **investment in time**. Usually the data science team builds a prototype that is then passed to the development team

For this reason, it is suggested to **involve a member of the development team** in the early phases of the data science project

Deployment can involve not only the final model, but also **previous phases** (data collection, model building, evaluation)

Epicyclic of data analysis

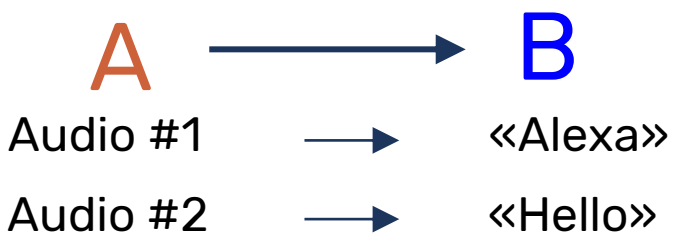
1. **Stating the question**
2. **Exploratory data analysis**
3. **Model building**
4. **Interpretation**
5. **Communication**



Workflow of a machine learning project

Build an home assistant device

1. Collect data
2. Train model
   - Iterate many times until good enough
3. Deploy model
   - Get data back
   - Maintain \ update model



Workflow of a data science project

Manufacturing line

Mix clay → Shape mug → Add glaze → Fire kiln → Final inspection

1. Collect data
2. Analyze data
   - Iterate many times to get good insights
3. Deploy model
   - Deploy changes
   - Re – analyze new data periodically

Clay batch #	Supplier	Mixing time (minutes)
001	Supplier 1	35
034	Supplier 1	22
109	Supplier 2	28

Mug batch #	Humidity	Temperute in kiln (F)	Temperute in kiln (hours)
001	0,002%	1410	22
034	0,003%	1520	24
109	0,002%	1420	22

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.
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.
6. Leo Breiman, *"Statistical Modeling: The Two Cultures"*, Statistical Science, vol. 16. no. 3, 2001

UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione