

State of the Art in Machine Learning

**Introduction to ML:
requirements & demos**

Table of Content

1. Pedagogical Goals
2. The Chain of *Business Intelligence*
3. AI: Introduction
4. Data
5. Major Concepts for ML Algorithms
6. Tools and Languages
7. Process of a Project
8. Jobs and ML
9. Ethics & Laws
10. Full Demo Iris: EDA + DT + Clustering
11. Digit: Outliers + PCA
12. (API and Web Scraping)

1. Pedagogical Goals

- A. Placing ML in the BI chain
- B. Understanding the stakes of AI / ML
- C. Identify necessary skills for ML
- D. Keys for successful ML projects

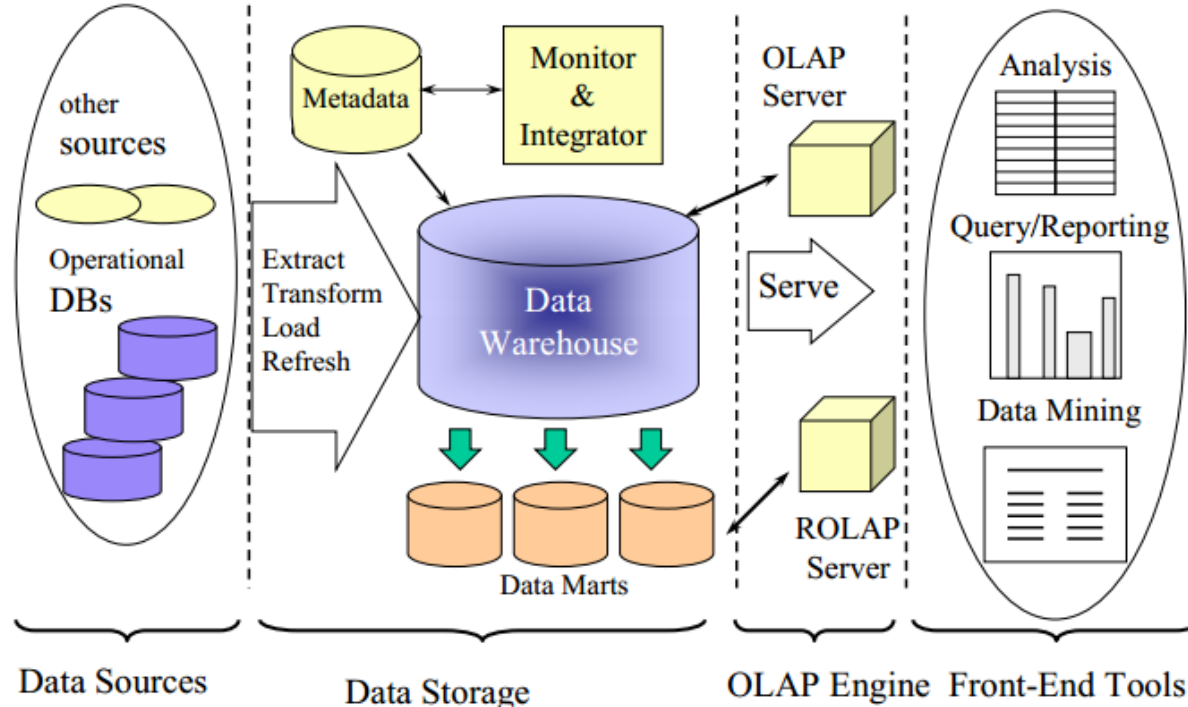
2. The Chain of *Business Intelligence*

- Aim of BI? Aid decision making
- Aid how? Based on reality, numbers
→ Graphs, tables and statistics: **reporting**
- Where to find those numbers? Clients, production, HR, sales, warehouses,...
→ Operational, internal and external sources: **sources**

2. The Chain of *Business Intelligence*

- Issues: sources from everywhere, not centralised, not consistent
- Solutions:
 - *Extract Transform Load*: **cleaning**
 - Centralisation into *Datawarehouses*: **reference**
- Conclusion: transform ***data into information***

2. The Chain of *Business Intelligence*



3. Introduction

Exercises to do in groups of 2-3 people:

Try answering the following questions relative to AI:

1. Come up with a list of applications you use that integrates AI. Can you group some of them together? Can you name these groups?
2. Could you explain, with your own words, how these apps/AI could work? What is needed for these to work?
3. Who does make use of those technologies? Why?

3. Introduction

???



amazon

???



TESLA

???



3. Introduction

Recommendation systems



amazon

Computer vision



Natural Language Processing
(NLP)



3. Introduction

And much more:

- Medical: bioinformatics, cancer recognition
- Finance: risk evaluation and prediction
- Logistics: storage, seasons and events, optimal openings / driving hours
- Waze: traffic jams

3. Introduction



3. Introduction

How can a computer, an AI, learn to recognise images of vehicles or animals, play chess, ...?

- With **a lot** of data
- With algorithms
- Plenty of computational power

3. Introduction

One could make the analogy with a human learning process. For example, we still need examples, preferably in different contexts, before we can use new concepts/things effectively.

MULTIPLICATION TABLE				
1 x 1 = 1	2 x 1 = 2	3 x 1 = 3	4 x 1 = 4	5 x 1 = 5
1 x 2 = 2	2 x 2 = 4	3 x 2 = 6	4 x 2 = 8	5 x 2 = 10
1 x 3 = 3	2 x 3 = 6	3 x 3 = 9	4 x 3 = 12	5 x 3 = 15
1 x 4 = 4	2 x 4 = 8	3 x 4 = 12	4 x 4 = 16	5 x 4 = 20
1 x 5 = 5	2 x 5 = 10	3 x 5 = 15	4 x 5 = 20	5 x 5 = 25
1 x 6 = 6	2 x 6 = 12	3 x 6 = 18	4 x 6 = 24	5 x 6 = 30
1 x 7 = 7	2 x 7 = 14	3 x 7 = 21	4 x 7 = 28	5 x 7 = 35
1 x 8 = 8	2 x 8 = 16	3 x 8 = 24	4 x 8 = 32	5 x 8 = 40
1 x 9 = 9	2 x 9 = 18	3 x 9 = 27	4 x 9 = 36	5 x 9 = 45
1 x 10 = 10	2 x 10 = 20	3 x 10 = 30	4 x 10 = 40	5 x 10 = 50
6 x 1 = 6	7 x 1 = 7	8 x 1 = 8	9 x 1 = 9	10 x 1 = 10
6 x 2 = 12	7 x 2 = 14	8 x 2 = 16	9 x 2 = 18	10 x 2 = 20
6 x 3 = 18	7 x 3 = 21	8 x 3 = 24	9 x 3 = 27	10 x 3 = 30
6 x 4 = 24	7 x 4 = 28	8 x 4 = 32	9 x 4 = 36	10 x 4 = 40
6 x 5 = 30	7 x 5 = 35	8 x 5 = 40	9 x 5 = 45	10 x 5 = 50
6 x 6 = 36	7 x 6 = 42	8 x 6 = 48	9 x 6 = 54	10 x 6 = 60
6 x 7 = 42	7 x 7 = 49	8 x 7 = 56	9 x 7 = 63	10 x 7 = 70
6 x 8 = 48	7 x 8 = 56	8 x 8 = 64	9 x 8 = 72	10 x 8 = 80
6 x 9 = 54	7 x 9 = 63	8 x 9 = 72	9 x 9 = 81	10 x 9 = 90
6 x 10 = 60	7 x 10 = 70	8 x 10 = 80	9 x 10 = 90	10 x 10 = 100



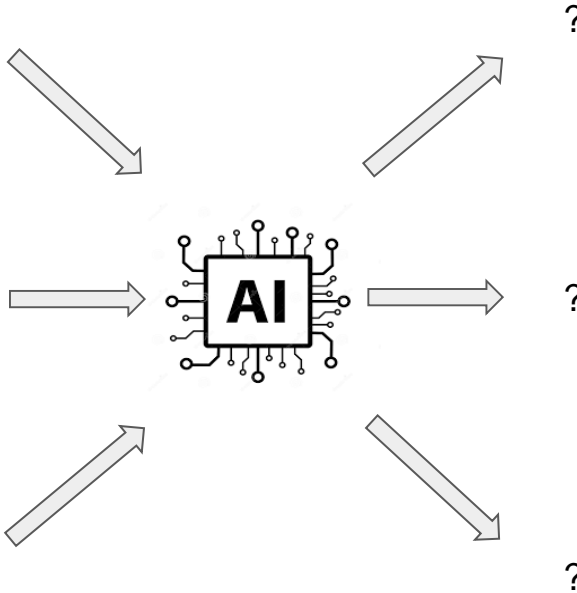
3. Introduction

However, is AI truly *intelligent*?

The next tweet summarises well the similarities between human intelligence and artificial intelligence



3. Introduction

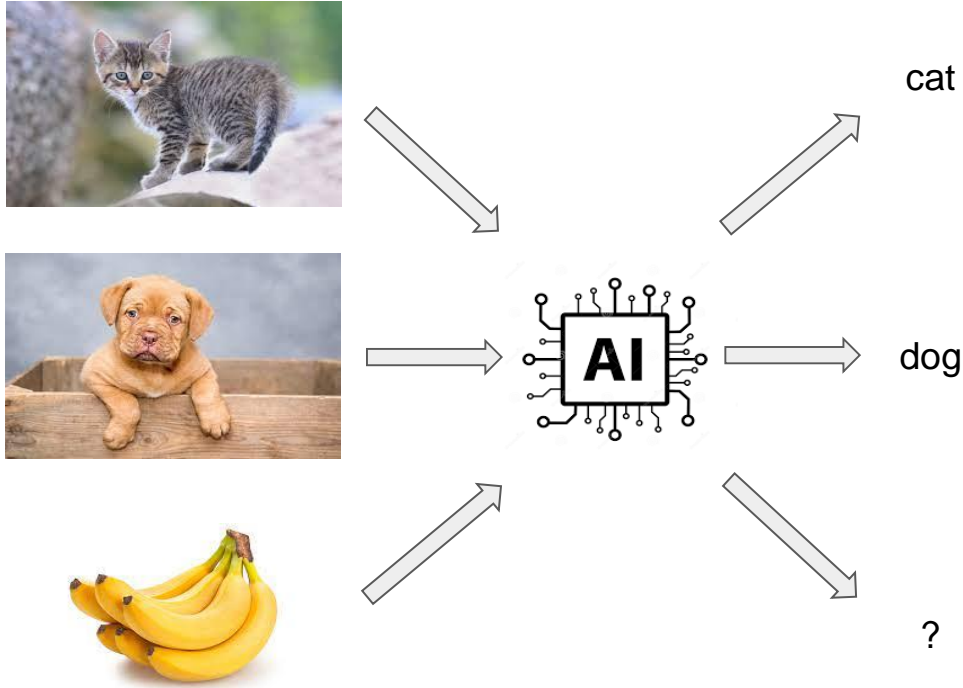


? Suppose we trained an AI to distinguish pictures of cats from pictures of dogs.

? What would happen if we present it with a banana picture?

?

3. Introduction



A priori, if we did our work well, the AI would provide the correct answer for both cat and dog.

However, if we would give it something different from a cat or dog, it will give as answer either cat, dog or both at the same time.

3. Introduction

Currently, AI is used to solve very specific problems. We are not able to produce an AI with a general intelligence similar to that of a human being.

We could train an AI to pass IQ tests but it would fail miserably at distinguishing a rabbit from a carrot.

If we train an AI to differentiate a cat from a motorcycle, we need a lot of pictures of both cats and motorcycles. If we add images of bicycles, we will just need as many. But bikes and motorcycles being very similar objects, this can bring its own complications.

3. Introduction

A human brain consumes 300 to 500 kcal per day (or 15 to 25 watts) compared to 250 watts for a professional GPU (e.g. Tesla V100) but there are some points to keep in mind:

- A single graphic card is not always enough
- You have to be able to cool them down
- Their life span is limited

3. Introduction

In spite of everything, the fantasy of "reproducing the human brain" remains very prevalent in our societies for at least three reasons:

- The media talk about it as an almost magical technology
- Books, comics, series, movies are all about this fantasy
- The lack of training

The comparison between the two intelligences seems to be risky at the moment. We should start by understanding how our brain works but also how some AIs take decisions.

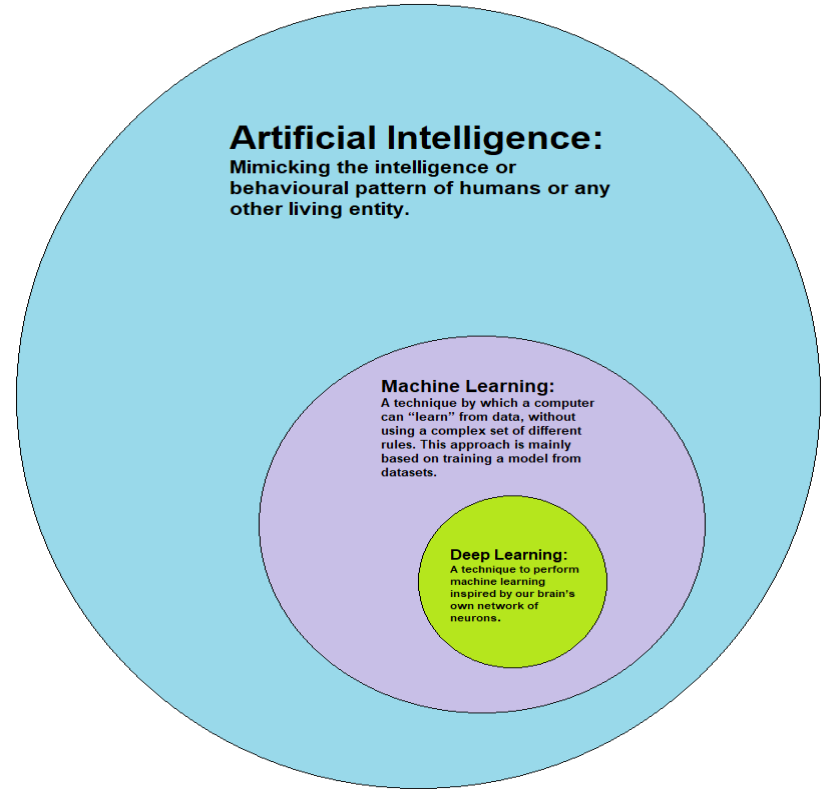
3. Introduction

How is AI related to *Machine Learning*?

AI: general and/or narrow intelligence, ethics, laws

ML: tools and techniques for narrow AI

DL & NN (neural networks): a very different kind of AI compared to conventional ML



4. Data

Activity in groups of 2 / 3:

Try answering the following questions:

- Give examples of data. Are they all the same?
- How are data collected?
- How are they stored?
- What can we do with it?
- Can we do everything with it?

4. Data

Structured Data:

- An Uber Eats delivery
- An Amazon order
- A shopping list
- A class of students and their grades

Unstructured Data:

- Images, videos
- Text, books, tweets, YouTube comments, *etc.*
- Sounds
- Medical data (genome)

4.1 Data Acquisition

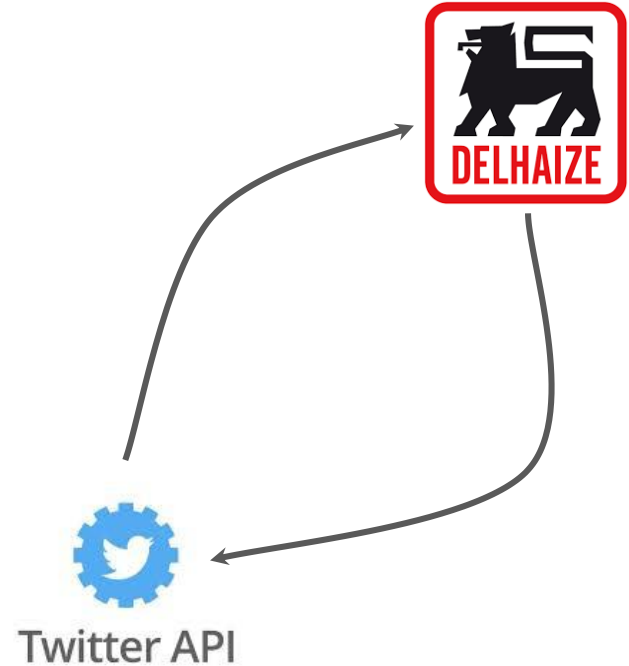
How do we generate, collect or obtain data?

- Manual encoding
- Polls, satisfaction surveys, *etc.*
- When you create an account or log in
- When you comment, like, swipe, post a photo, *etc.*
- It is possible to buy data

4.1 Data Acquisition


Let's say you work for Delhaize and you launch a new advertising campaign on social networks.

How do you get the data to analyse how the campaign has been received?



4.1 Data Acquisition

APIs are services that allow companies, government agencies, NGOs to expose their data to the outside world.

FREE	DEVELOPER	STARTER	PRO	BUSINESS	ENTERPRISE
\$0 /month	\$4 /month	\$7 /month	\$19 /month	\$65 /month	
Get Started	Get Started	Get Started	Get Started	Get Started	Contact Us
1,000,000 Calls per month	2,000,000 Calls per month	3,000,000 Calls per month	5,000,000 Calls per month	10,000,000 Calls per month	High volume calls per month
Realtime weather	Realtime weather	Realtime weather	Realtime weather	Realtime weather	Realtime weather
3 day city and town weather. Daily and Hourly.	5 day city and town weather. Daily and Hourly.	7 day city and town weather. Daily and Hourly.	10 day city and town weather. Daily and Hourly.	14 day city and town weather. Daily and Hourly.	14 day city and town weather. Daily and Hourly.
Search API	Search API	Search API	Search API	Search API	Search API
Astronomy API	Astronomy API	Astronomy API	Astronomy API	Astronomy API	Astronomy API
IP Lookup	IP Lookup	IP Lookup	IP Lookup	IP Lookup	IP Lookup

<https://www.weatherapi.com/pricing.aspx>

4.1 Data Acquisition

Web scraping is another tool that allows to extract information, photos, videos from a web page in an automated way with the help of robots.

Several disadvantages:

- Sites are increasingly protecting themselves from it
- Not always legal
- If the site changes, the scraping does not work anymore

4.1 Data Acquisition

Quotes to Scrape

"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking."

by [Albert Einstein](#) (about)

Tags: [change](#) [deep-thoughts](#) [thinking](#) [world](#)

"It is our choices, Harry, that show what we truly are, far more than our abilities."

by [J.K. Rowling](#) (about)

Tags: [abilities](#) [choices](#)

"There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle."

by [Albert Einstein](#) (about)

Tags: [inspirational](#) [life](#) [live](#) [miracle](#) [miracles](#)

"The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid."

by [Jane Austen](#) (about)

Tags: [aliteracy](#) [books](#) [classic](#) [humor](#)

<https://quotes.toscrape.com/>



1d681e0a040c9f
a8bfc3a2fd1b15f
5280a98b2cb.jpg



1fb3bbc5dbca1e
ac326738dc1f1d
a238e48649cb.jp



3eeb58092c9d71
08634ae1ec7dc7
b903f6a385e3.jp



3fc91589e130f46
a8e9e41ac9543fe
6aa35d85fb.jpg



5a1ac9d08bba40
3f446b77822218
d3498e08d4c.jp



5b5ec5caa03682
872a2a0af67305
6650849a57c6.jp



5f1f09e62587efb
d50025e5b8494b
c7e7f9b5db5.jpg



6f222aa7fd5e0d0
6e0d7bdc476032
10f2140368.jpg



9b59f8a369147fb
79580b6859dd45
54618539399.jpg



9eb241f185f936a
db97aa955f048c
611a9073aa6.jpg



20bdeab733a902
d3dd43377222a
04d418ad2d2d.jp



23e8b7f3798df28
602429f5194176
a21146fa8d8.jpg



43c92c68669ac4
0730e257e807f9
0afda571885.jpg



98ba7223263dfb
01c4d2c96f06d6
78c92ae038ff.jpg



quote	author	tags	born_date	born_place
the person be it gentleman or lady who h...	Jane Austen	aliteracy,books,classic...	1775-12-16 00:00:00.000	in Steventon Rectory, Hampshir...
it is our choices harry that show what we ...	J.K. Rowling	abilities,choices	1965-07-31 00:00:00.000	in Yate, South Gloucestershire, ...
imperfection is beauty madness is geniu...	Marilyn Monroe	be-yourself,inspirational	1926-06-01 00:00:00.000	in The United States
the world as we have created it is a proc...	Albert Einstein	change,deep-thoughts...	1879-03-14 00:00:00.000	in Ulm, Germany
there are only two ways to live your life o...	Albert Einstein	inspirational,life,life,mi...	1879-03-14 00:00:00.000	in Ulm, Germany
try not to become a man of success rath...	Albert Einstein	adulthood,success,val...	1879-03-14 00:00:00.000	in Ulm, Germany
if you want your children to be intelligent ...	Albert Einstein	children,fairy-tales	1879-03-14 00:00:00.000	in Ulm, Germany
a wise girl kisses but doesnt love listens ...	Marilyn Monroe	attributed-no-source	1926-06-01 00:00:00.000	in The United States
love does not begin and end the way we ...	James Baldwin	love	1924-08-02 00:00:00.000	in Harlem, New York, The Unite...
it matters not what someone is born but ...	J.K. Rowling	dumbledore	1965-07-31 00:00:00.000	in Yate, South Gloucestershire, ...
life is like riding a bicycle to keep your ba...	Albert Einstein	life,simile	1879-03-14 00:00:00.000	in Ulm, Germany
the real lover is the man who can thrill yo...	Marilyn Monroe	love	1926-06-01 00:00:00.000	in The United States
anyone who thinks sitting in church can ...	Garrison Keillor	humor,religion	1942-08-07 00:00:00.000	in Anoka, Minnesota, The Unite...
a woman is like a tea bag you never kno...	Eleanor Roosevelt	misattributed-eleanor-...	1884-10-11 00:00:00.000	in The United States
i may not have gone where i intended to ...	Douglas Adams	life,navigation	1952-03-11 00:00:00.000	in Cambridge, England, The Uni...

4.2 Data Storage



4.2 Data Storage

When we talk about data storage, we need at least three elements:

- A database
- A model, a way of seeing and storing data
- A database management system (DBMS), a software to manage databases

4.2 Data Storage

The most common model for database design is the relational model, which is built around a system of tables, relationships between tables and rules to be respected.

It is accompanied by a whole series of different DBMS that support the relational model. They are themselves accompanied by a language (SQL) that allows you to create, query and modify a database.

4.2 Data Storage

ID_Customer	LastName	FirstName	Address
1	De Croo	Alice	Rue des anges, 32
2	Dupont	Eric	Avenue Charles Lestranger, 265
3	Heidegger	Margot	Rue des anges, 15



OrderNum	Date	Customer	Product
1	10/08/2020	1	2
2	02/01/2021	1	2
3	15/05/2021	2	1
4	28/06/2021	3	3



ID_Product	Name	Price
1	Iphone	800€
2	Samsung Galaxy	750€
3	Sony XM4	300€

4.2 Data Storage

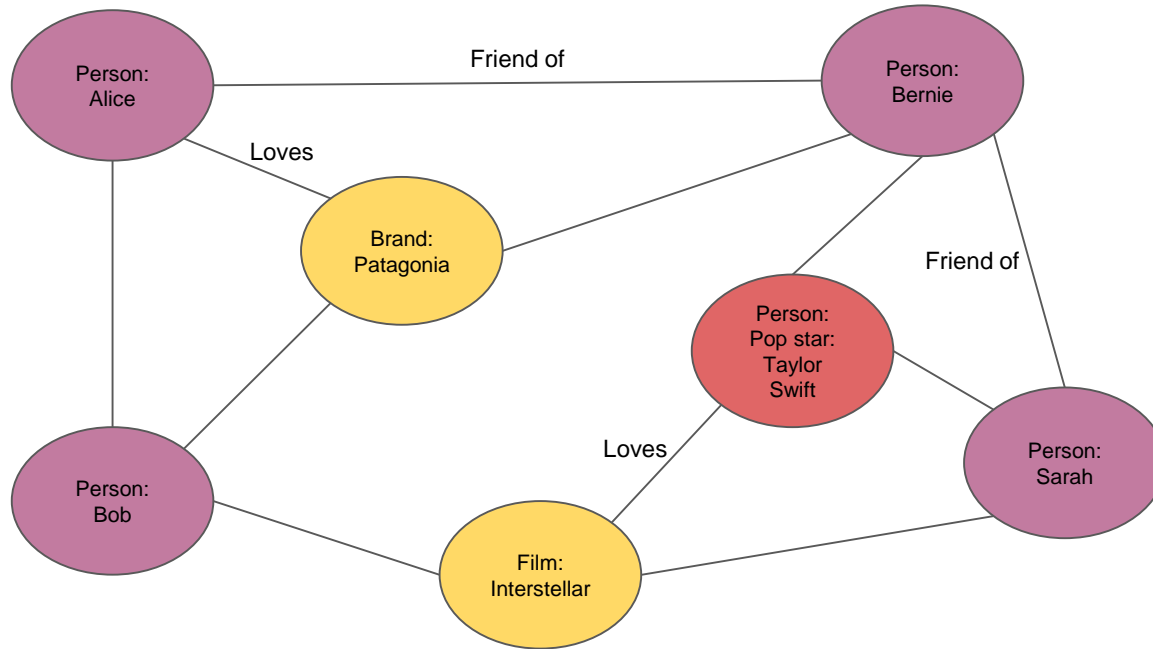
Advantages

- Very structured vision of its data
- Takes up little space on a hard disk
- A common basic language (SQL)

Disadvantages

- How to manage unstructured data?
- Performance problems during heavy operations on large volumes of data

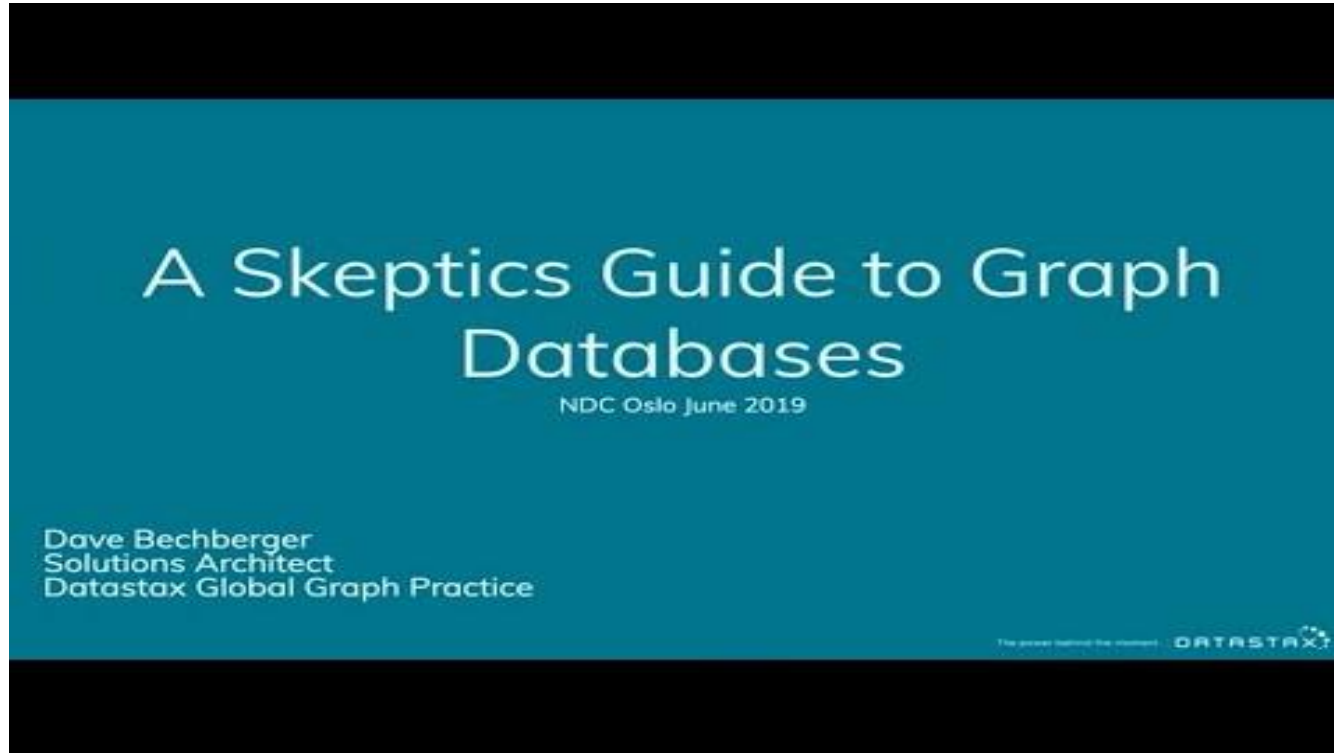
4.2 Data Storage



Next to relational databases, we find oriented databases:

- Documents
- Graphs (see left)
- *etc.*

4.2 Data Storage

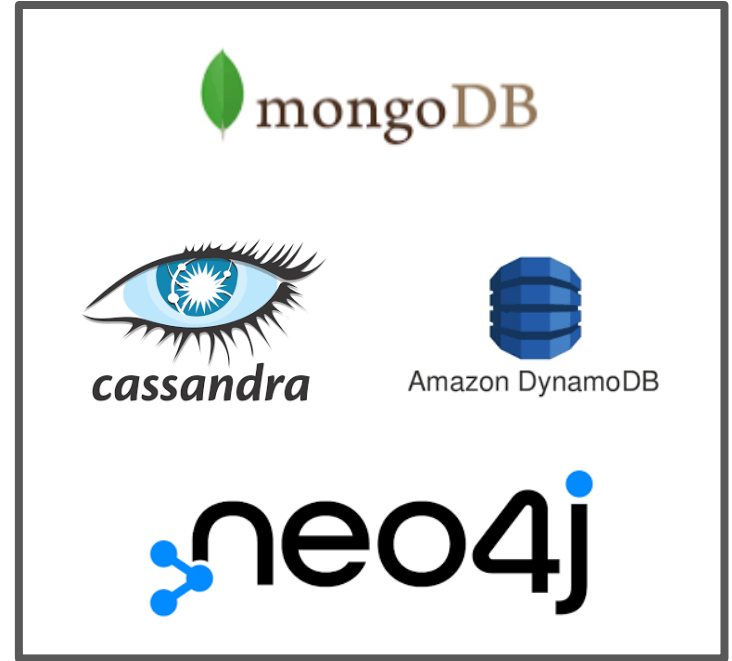


4.2 Data Storage

SGBD SQL



SGBD NoSQL



4.2 Data Storage

Why DBMS?

- Security
- Consistency
- Better quality data
- Easier to share data
- *etc.*



4.3 Not Everything Is AI / ML



4.3 Not Everything Is AI / ML

The use of data goes far beyond the framework of artificial intelligence:

- It has first of all an operational role. If Amazon asks for your address, it is first and foremost to know where to ship your orders. It can also be used to optimise the routes of the delivery drivers for example.
- For legal reasons, companies have to keep their data for a certain period of time.

4.3 Not Everything Is Destined for AI / ML

Another possible use consists in highlighting certain information contained in the data with the help of graphics (**reporting**):

- Allows you to make information contained in your data intelligible
- Allows you to track your Key Performance Indicators (KPI) in real time
- Allows you to help your decision making
- *etc.*

4.3 Not Everything Is AI / ML

Tableau or Power BI are tools that allow you to produce interactive dashboards in order to present data, present analysis, etc.

Example:

- [Bicycle accidents in London](#)
- [Most Wanted](#)

5. Algorithms

5.1 Introductory Exercise

For groups consisting of 2-3 people:

Two problems from everyday life are proposed to you.

Try to solve them and, above all, note the steps and elements that seem important to you to solve these problems.

Several approaches are possible, feel free to list them.



5.1 Introductory Exercise

Choosing your car

Problem:

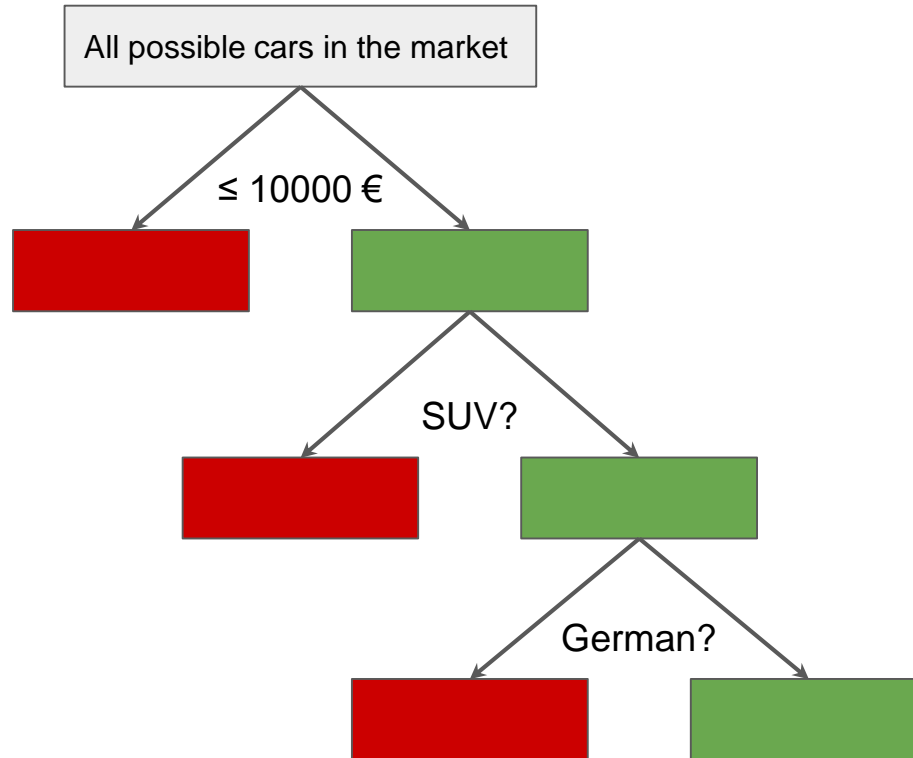
- In the automotive market, buyers can quickly get lost when making their next car purchase
- How would you go about choosing your next car from this almost endless catalog of cars?

Optimise your harvest

Problem:

- You are a farmer or a novice in gardening and you don't know when to plant your seeds in order to maximise your harvest
- How would you go about determining the best planting period?

5.1 Introductory Exercise

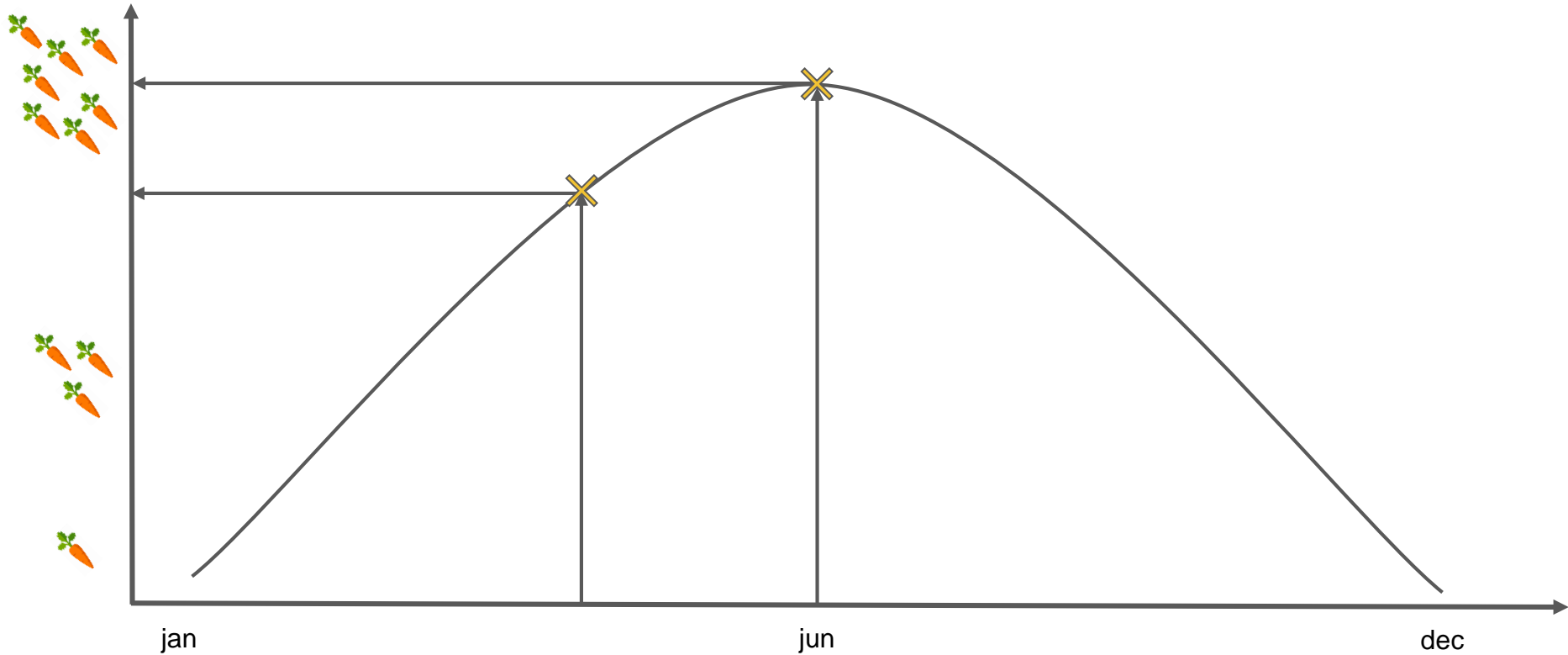


5.1 Introductory Exercise

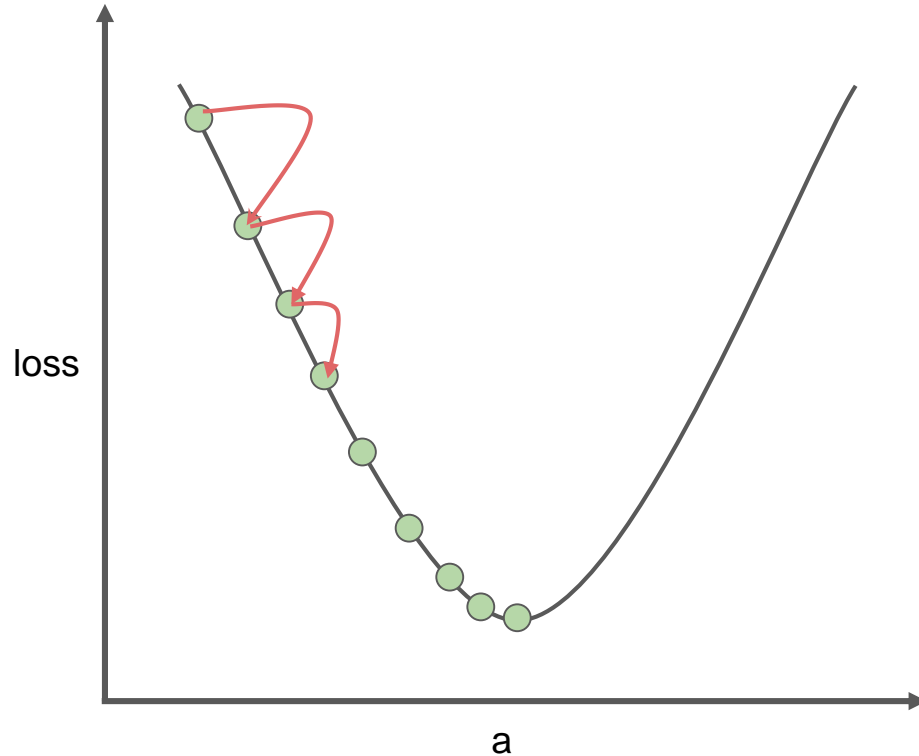
Decision trees are very common algorithms for various reasons:

- Easy to understand
- Easy to build (sequences of binary decisions)
- Very efficient despite their apparent naivety
- They can be used for various problems. Estimating the price of a house, a car, choosing your next car. They are also used in some video games.

5.1 Introductory Exercise



5.1 Introductory Exercise



The gradient descent algorithm is an optimisation algorithm that minimises what is called a cost function (loss).

In our example, our cost function could be the number of carrots lost or not grown.

5.2 Basic Concepts

As a reminder, a computer needs a lot of examples for its learning. This is what we call a dataset.

It is not enough to give it examples. It is also necessary to give it a method to learn (an algorithm).

However, there are many different algorithms which answer different situations.

5.2 Basic Concepts

A dataset is composed of:

- Samples: the observations, the rows
- Features: the explanatory variables, the columns, the X
- Label (target): optional, a target value that we want to determine, y

Two main categories of learning:

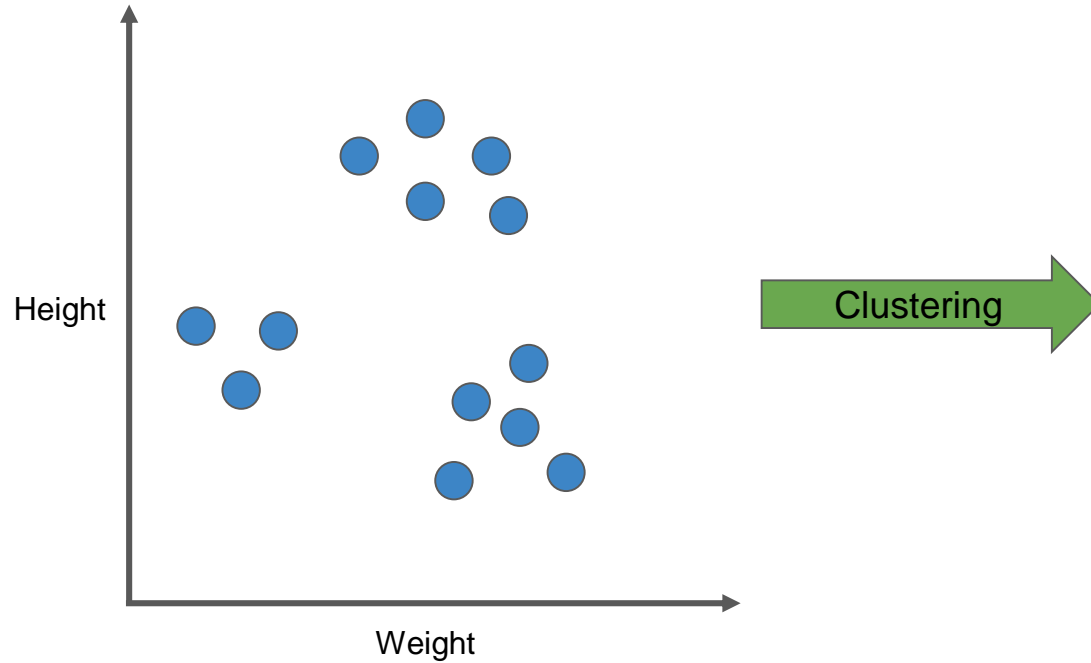
- Unsupervised: no target
- Supervised: one or more targets
- (Reinforcement: exploring possible paths iteratively)

5.2 Basic Concepts

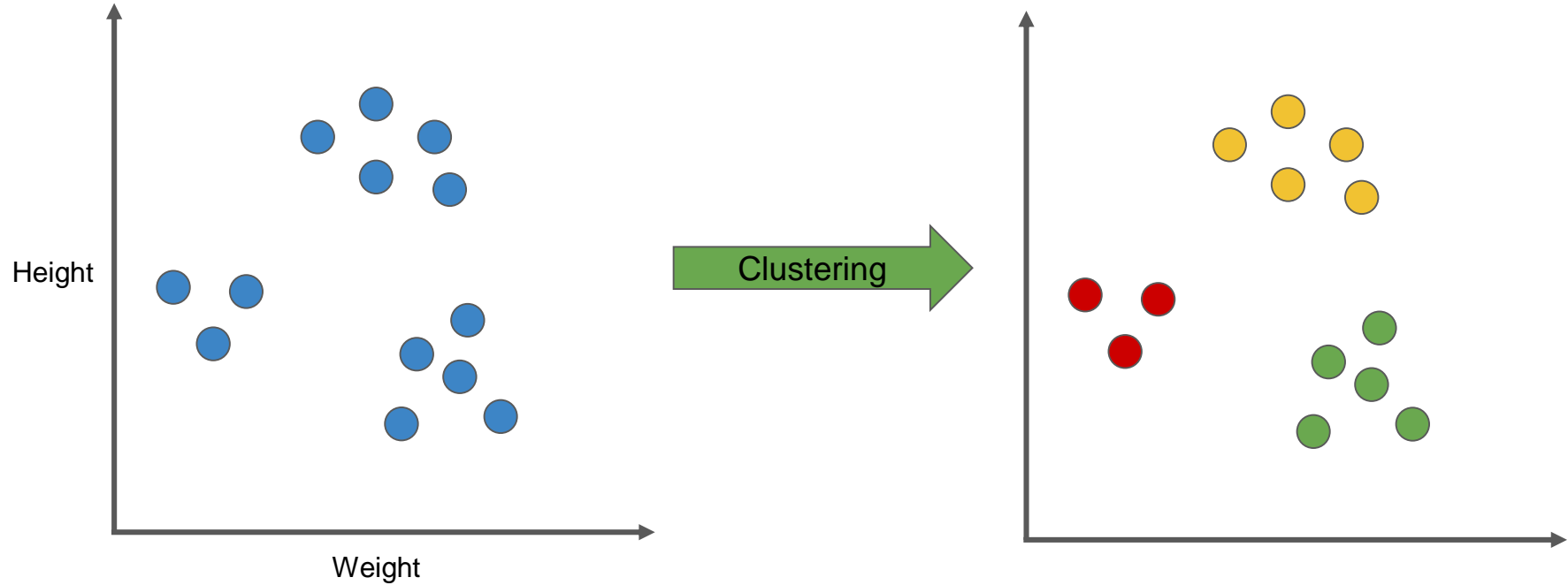
3 main categories of problems:

- Clustering: unsupervised, we group similar individuals together
- Regression: supervised, we try to determine the value of a continuous quantitative variable (e.g. the price of a house)
- Classification: supervised, we try to determine the value of a discrete variable (e.g. dog or cat)

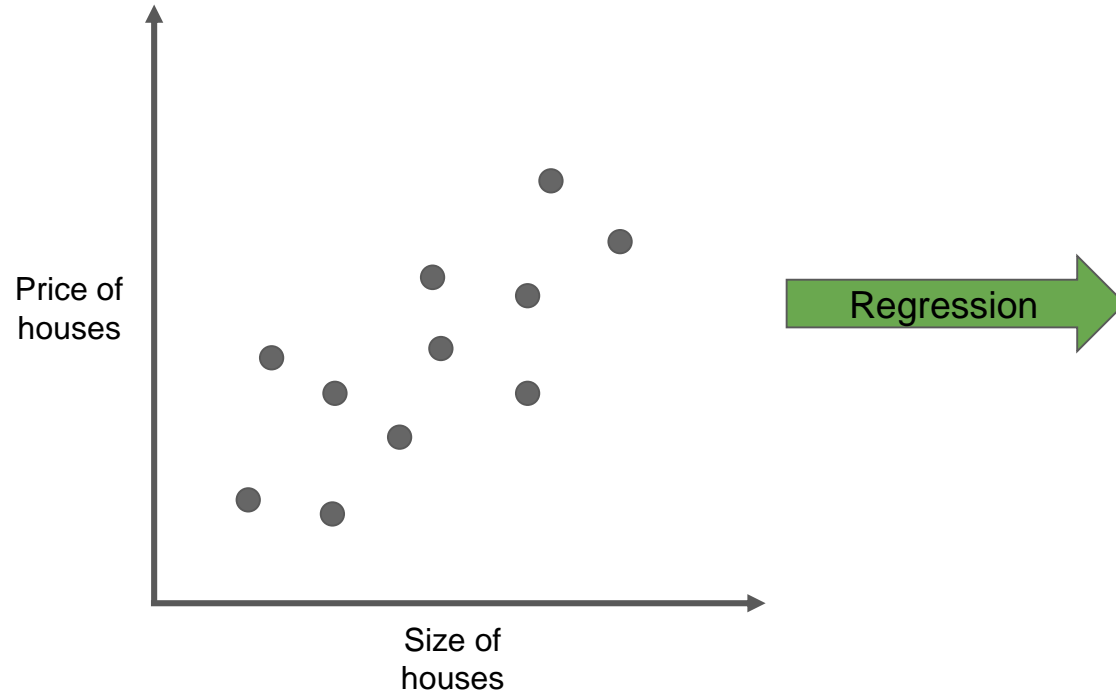
5.2 Basic Concepts



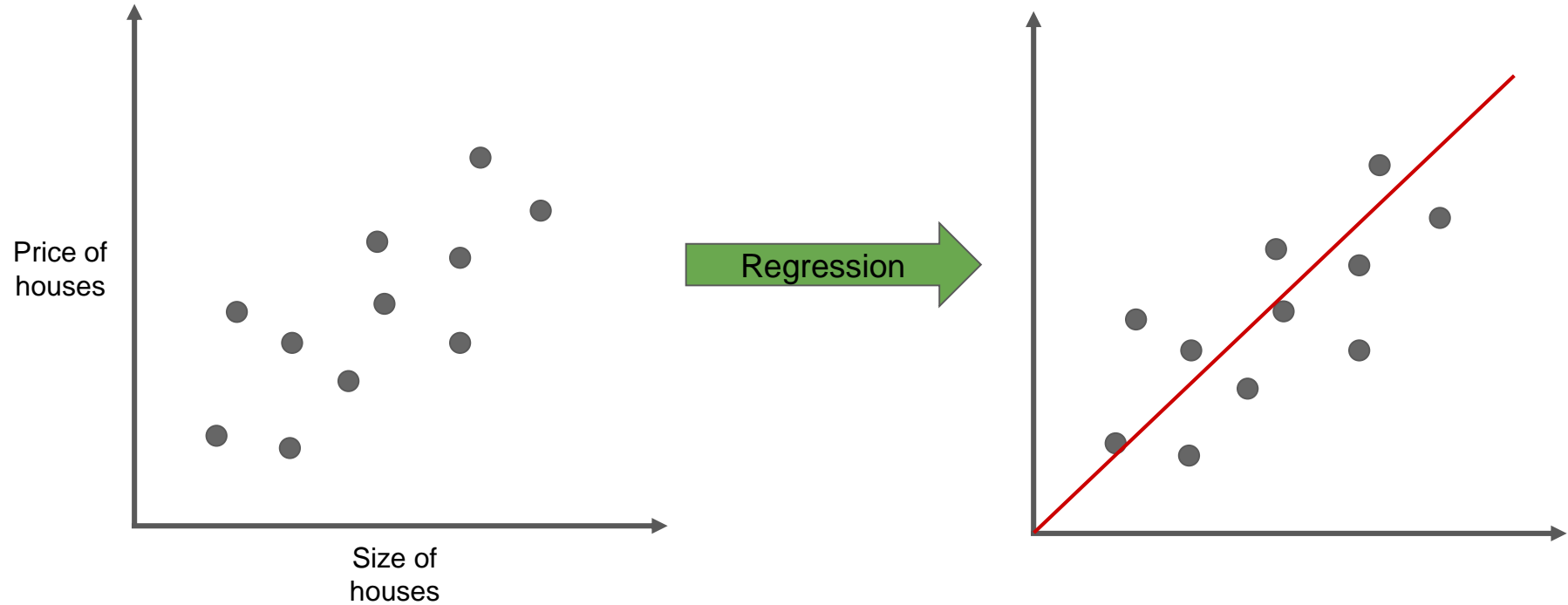
5.2 Basic Concepts



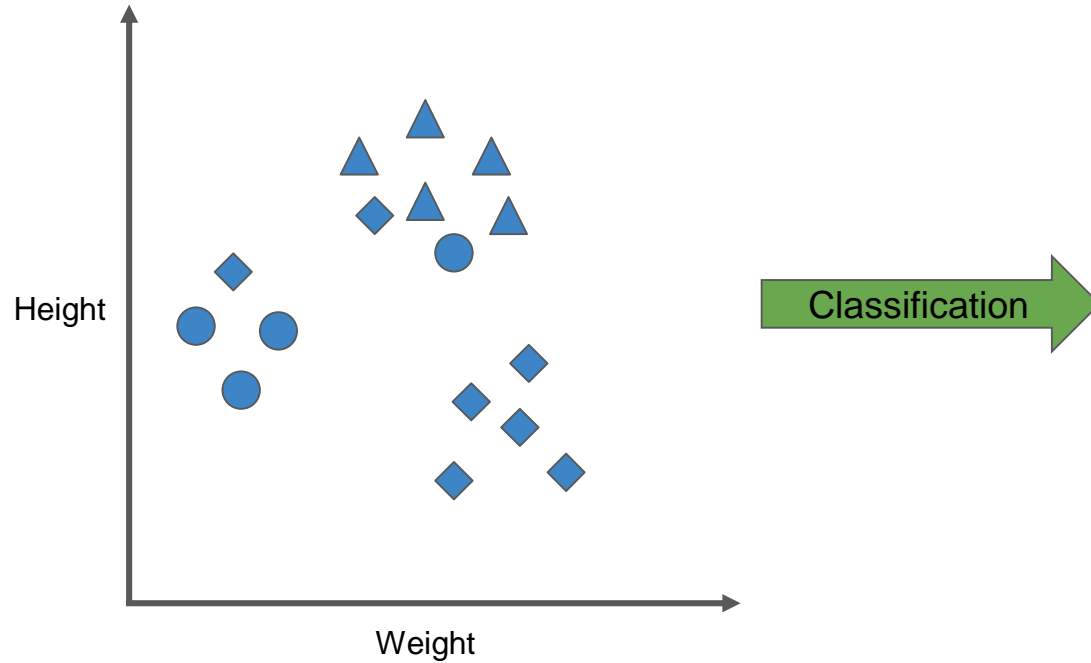
5.2 Basic Concepts



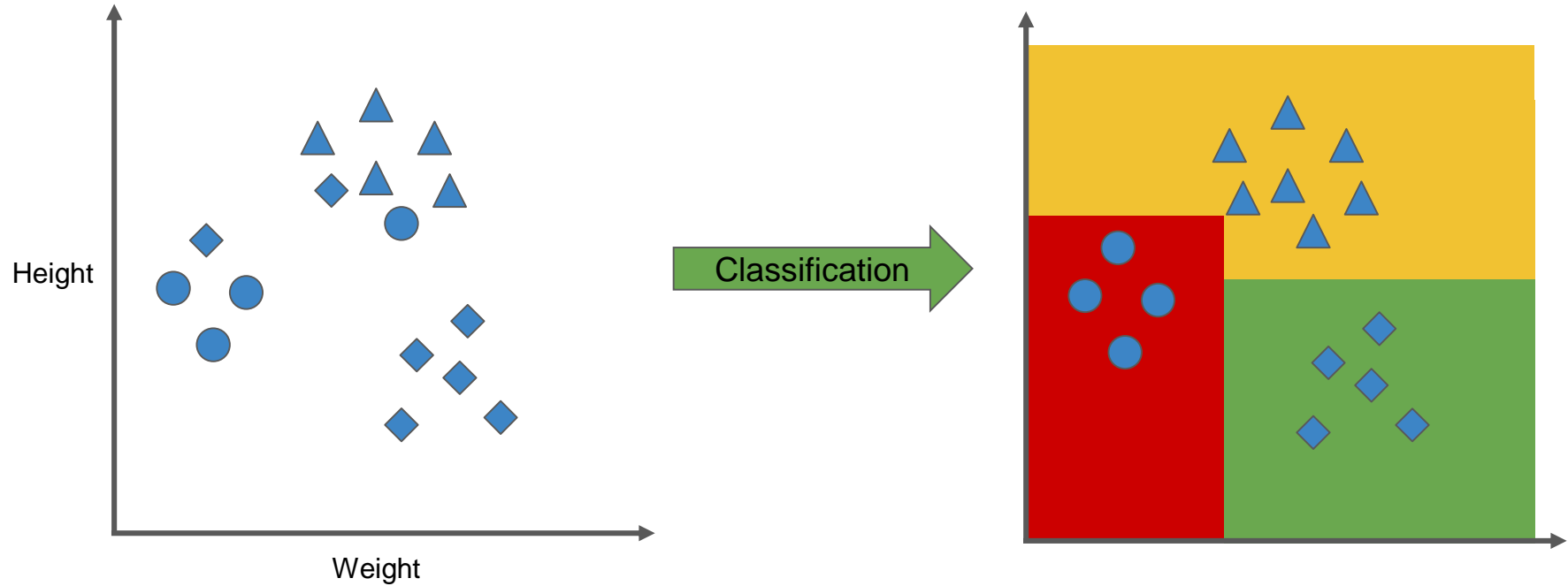
5.2 Basic Concepts



5.2 Basic Concepts



5.2 Basic Concepts



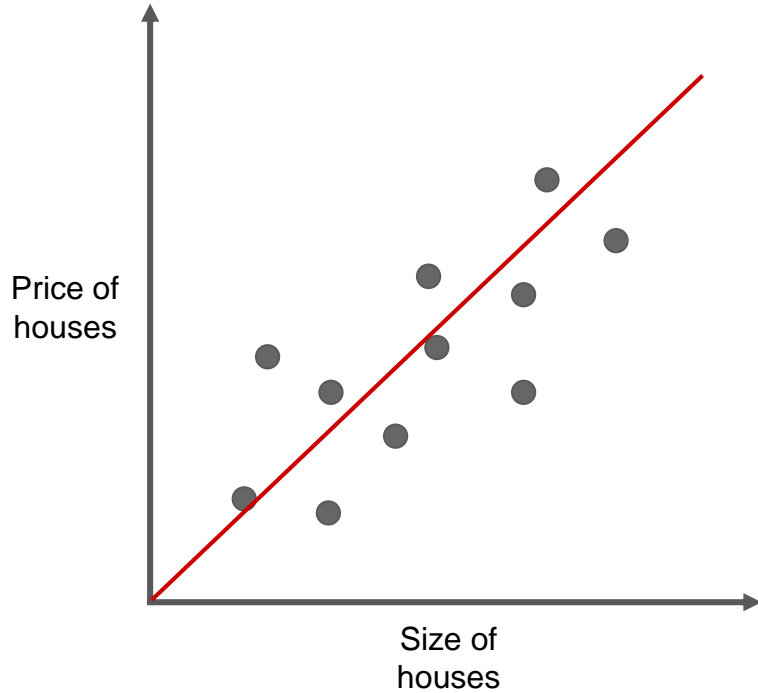
5.2 Basic Concepts

In concrete terms, what does it mean to train an algorithm?

Two examples to better understand:

- Linear regression and more generally linear models
- Decision trees

5.3 Linear Regression

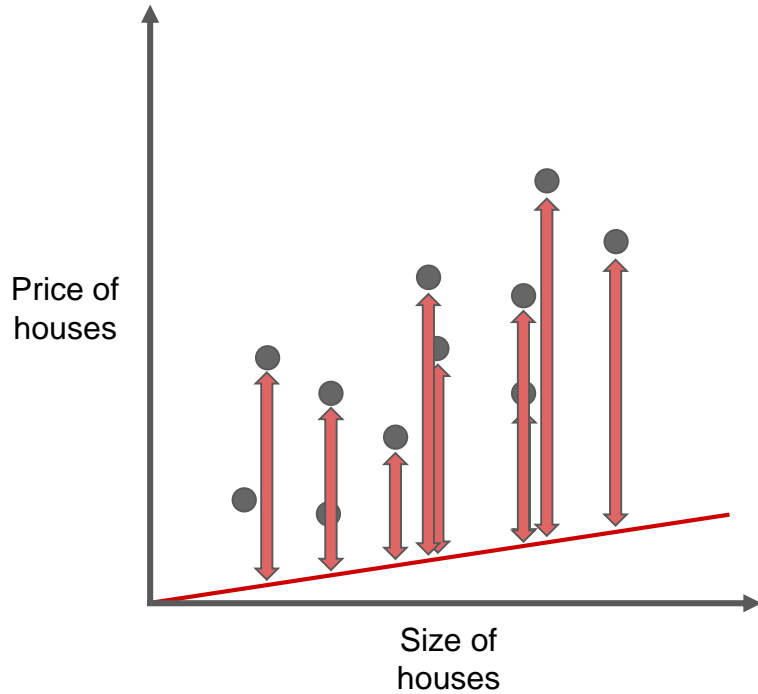


We have drawn a line between our points

$$f(x) = ax + b$$

- a = coefficient
- b = constant
- x = variable (size of our house)
- $f(x)$ = to make it simple, our target

5.3 Linear Regression

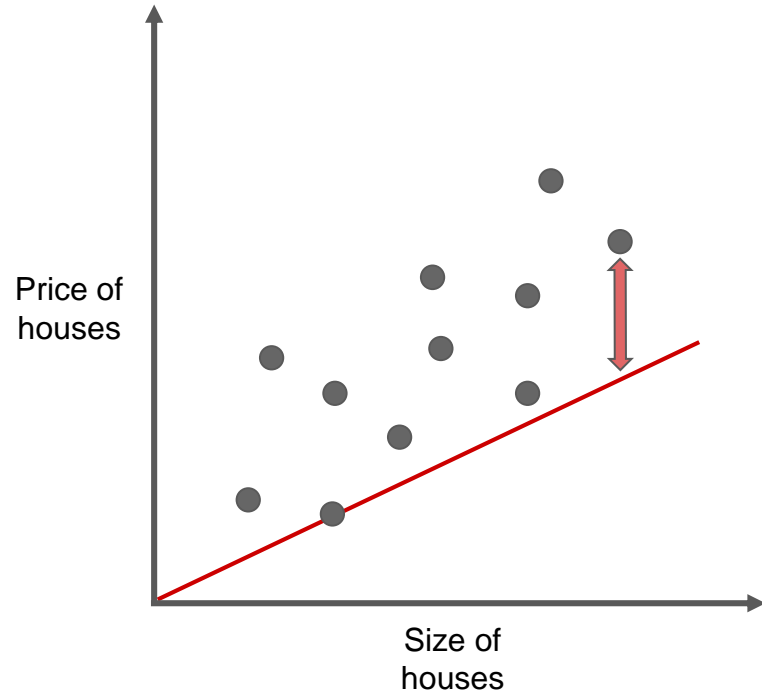


How to find the right coefficient?

By trial and error (*):

- We choose the first time a coefficient randomly
- We look at the amount of error
- We modify our coefficient

5.3 Linear Regression

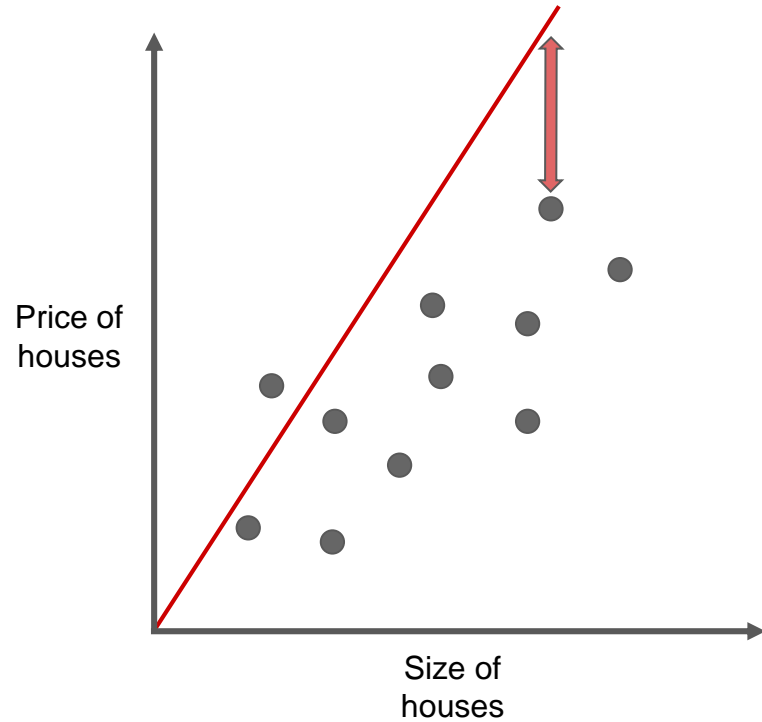


At each iteration, we calculate our errors and modify our coefficient

5.3 Linear Regression



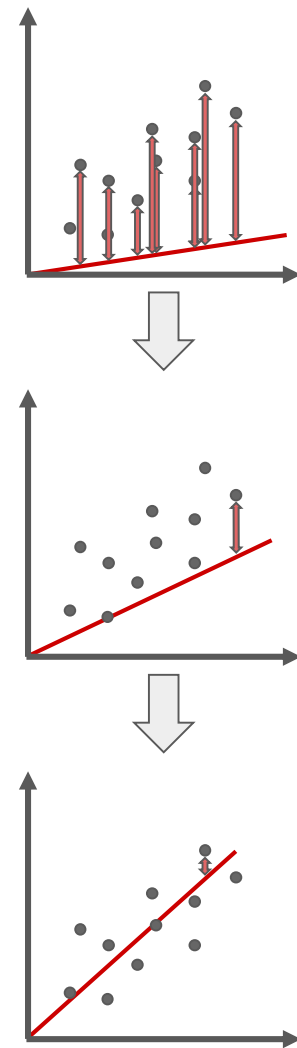
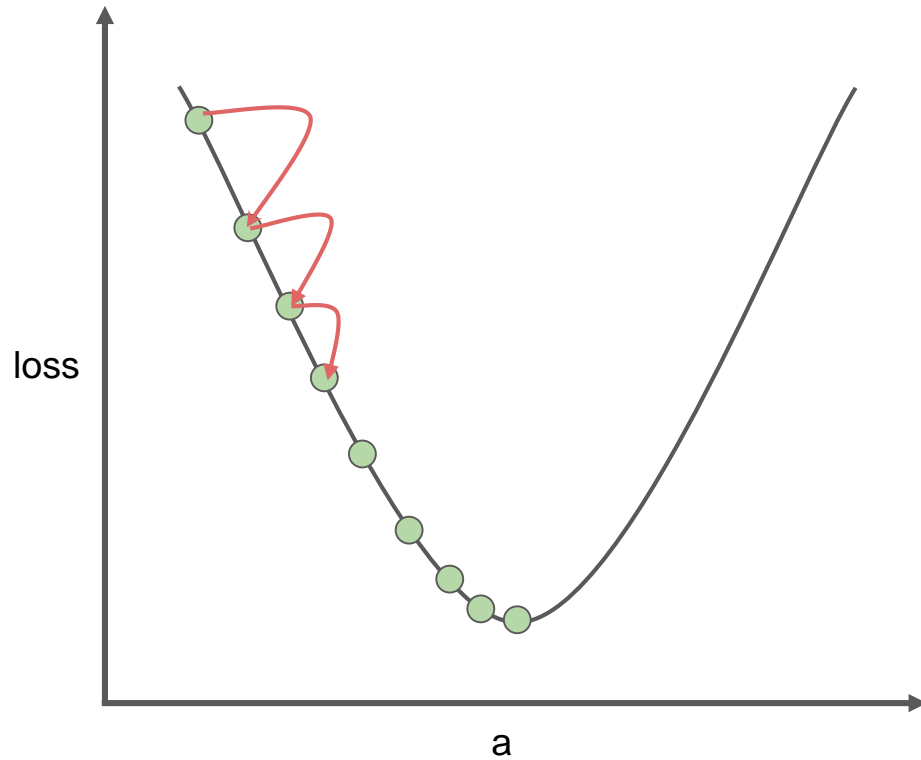
5.3 Linear Regression



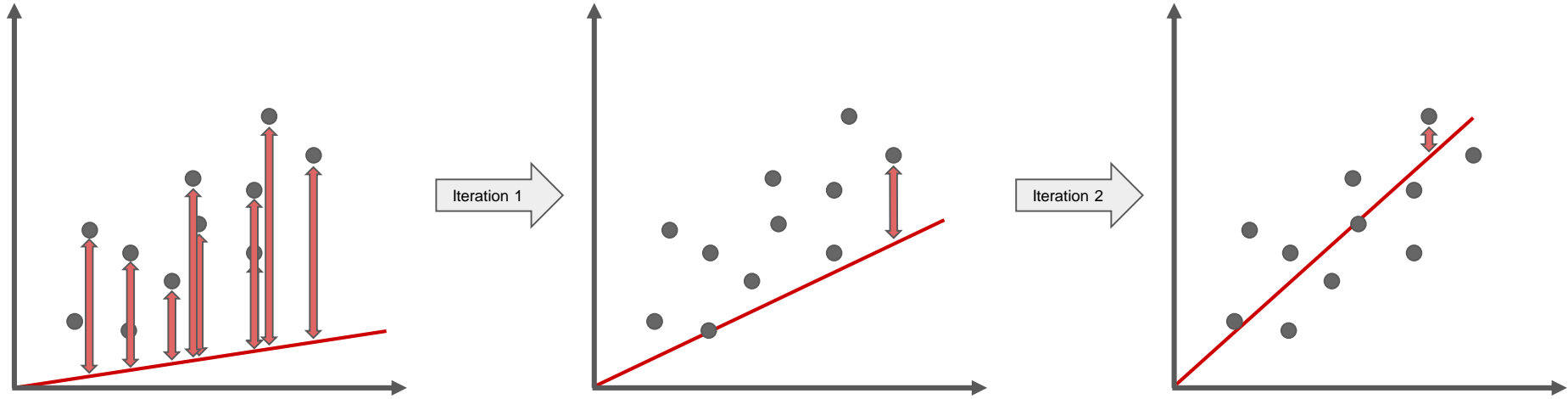
We went too far.

The training is over

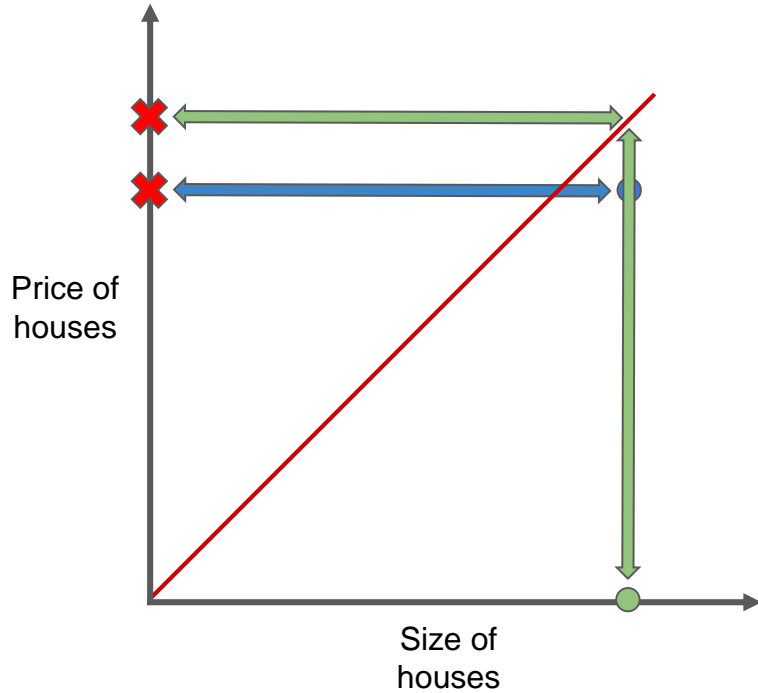
5.3 Linear Regression



5.3 Linear Regression



5.3 Linear Regression



Now we can predict the price of a house.

In blue, the real value, and, in green, the prediction of our model

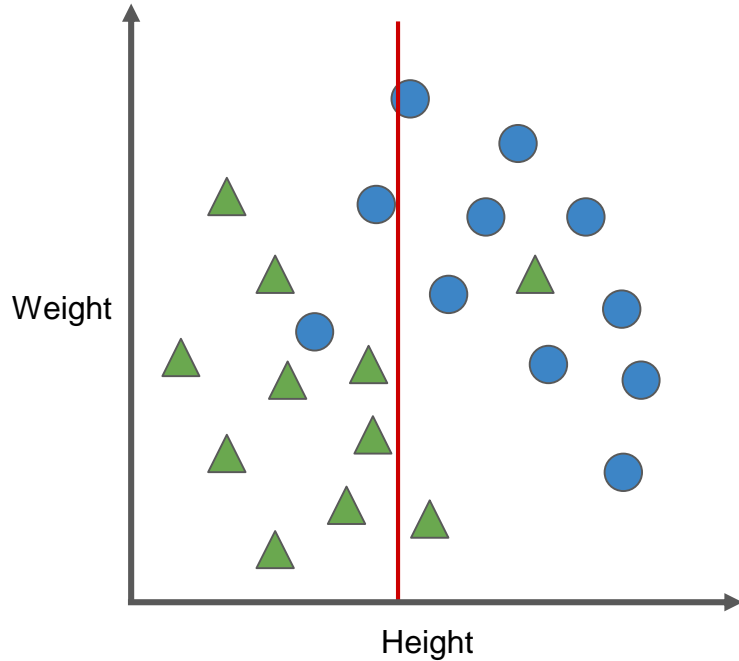
5.3 Linear Regression

If we have more than one variable?

$$f(x) = a_1x_1 + a_2x_2 + b$$

price = a_1 * house size + a_2 * garden size + b

5.4 Logistic Regression

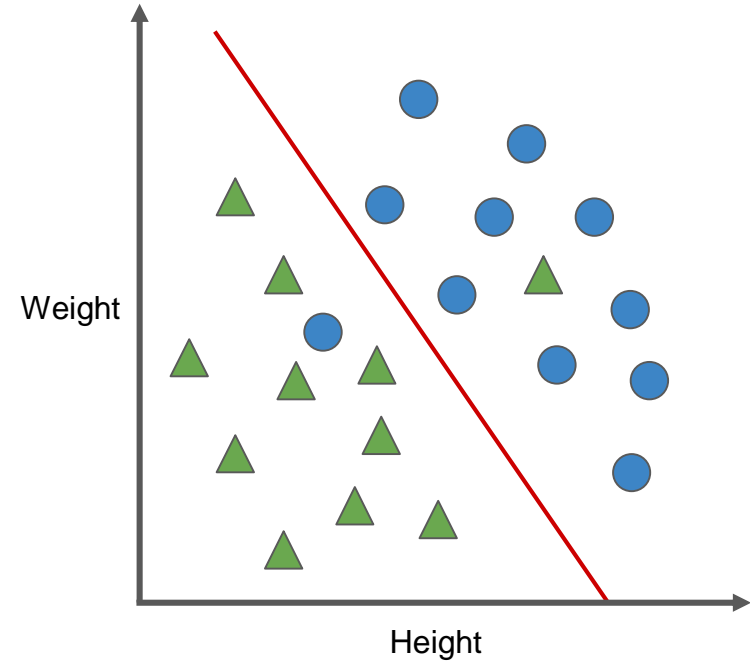
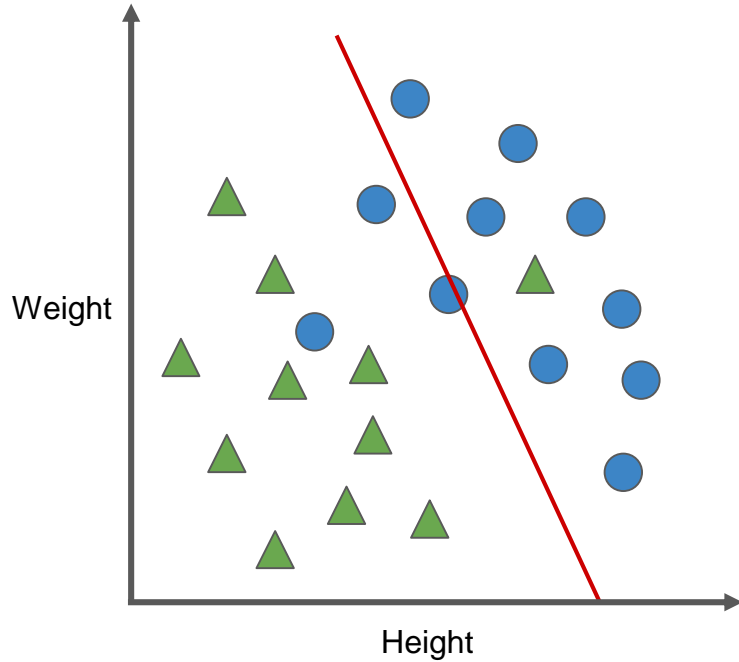


→ Same principle as for linear regressions
→ Classification

With linear regression we minimise the errors.

With logistic regression we try to maximise the distance between our classes.

5.4 Logistic Regression



5.5 Decision Trees

- For regression or classification
- Non linear
- Very simple to interpret

5.5 Decision Trees

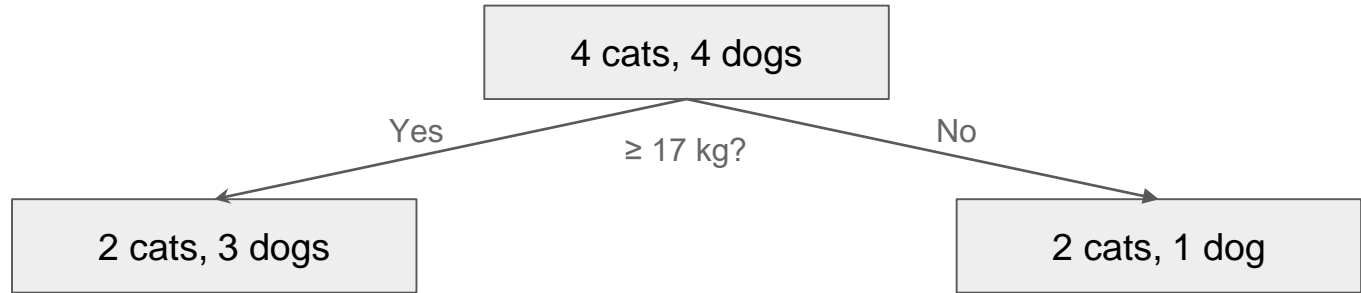
Weight (kg)	Target (class)
13	Cat
17	Dog
26	Dog
32	Dog
18	Dog
8	Cat
15	Dog
20	Cat

We have two possible classes and only one feature to discriminate them.

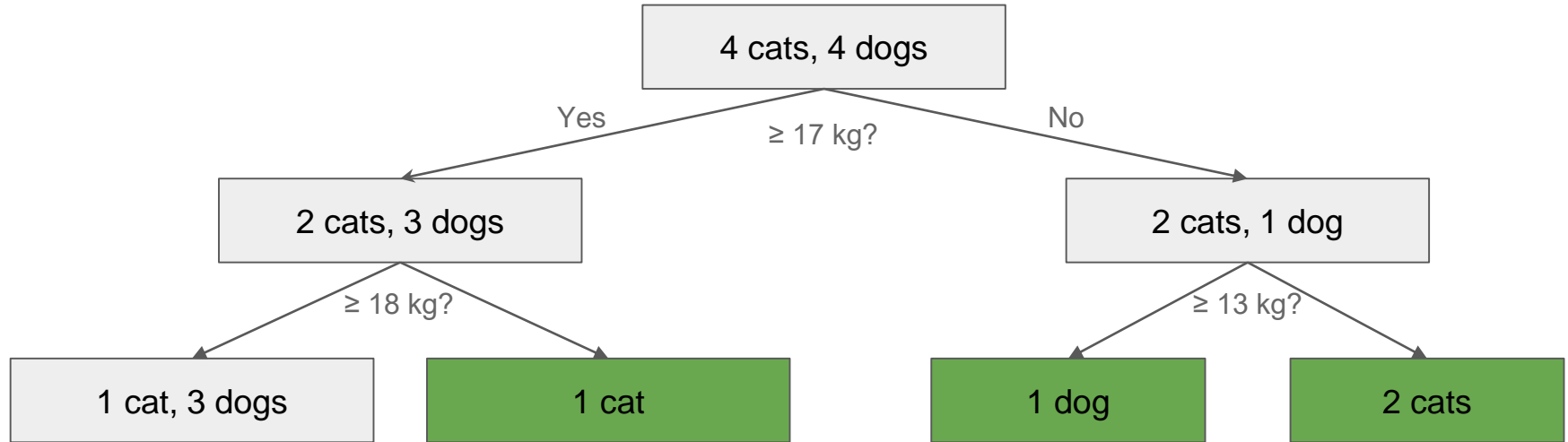
→ Decision trees are built by asking a succession of questions.

We take a random value to separate our data in two.

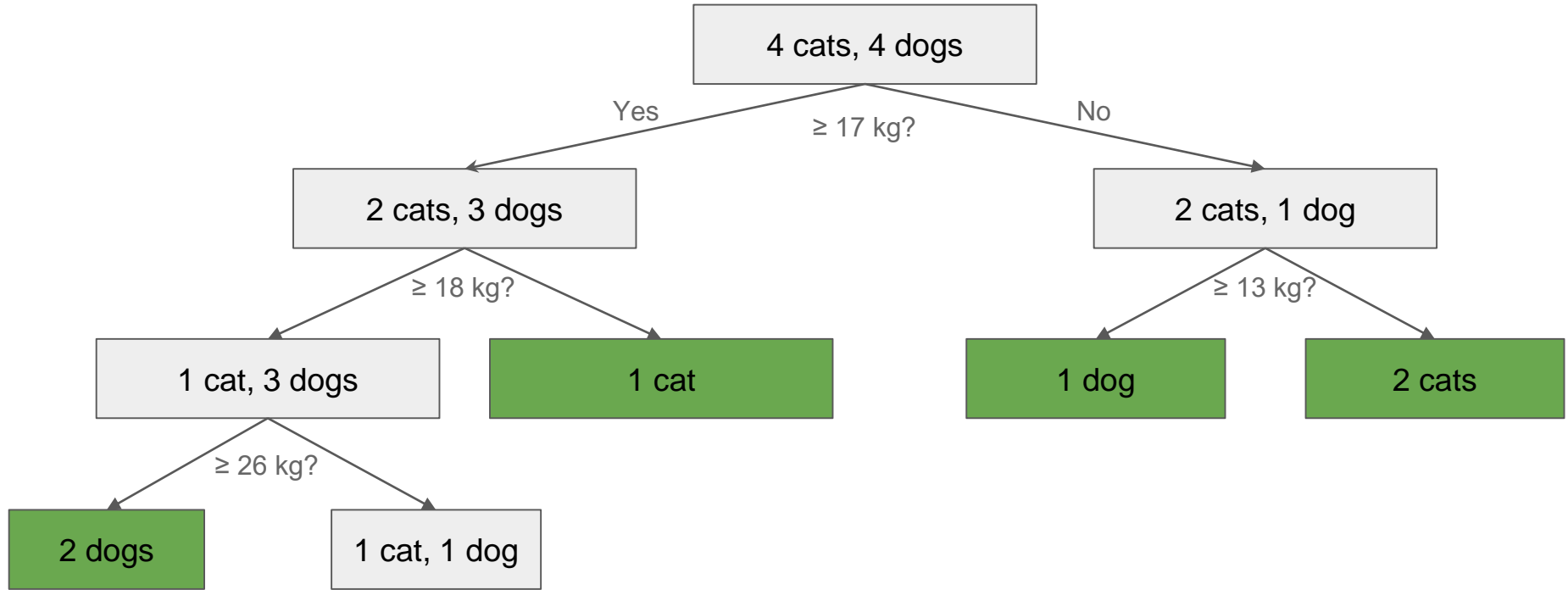
5.5 Decision Trees



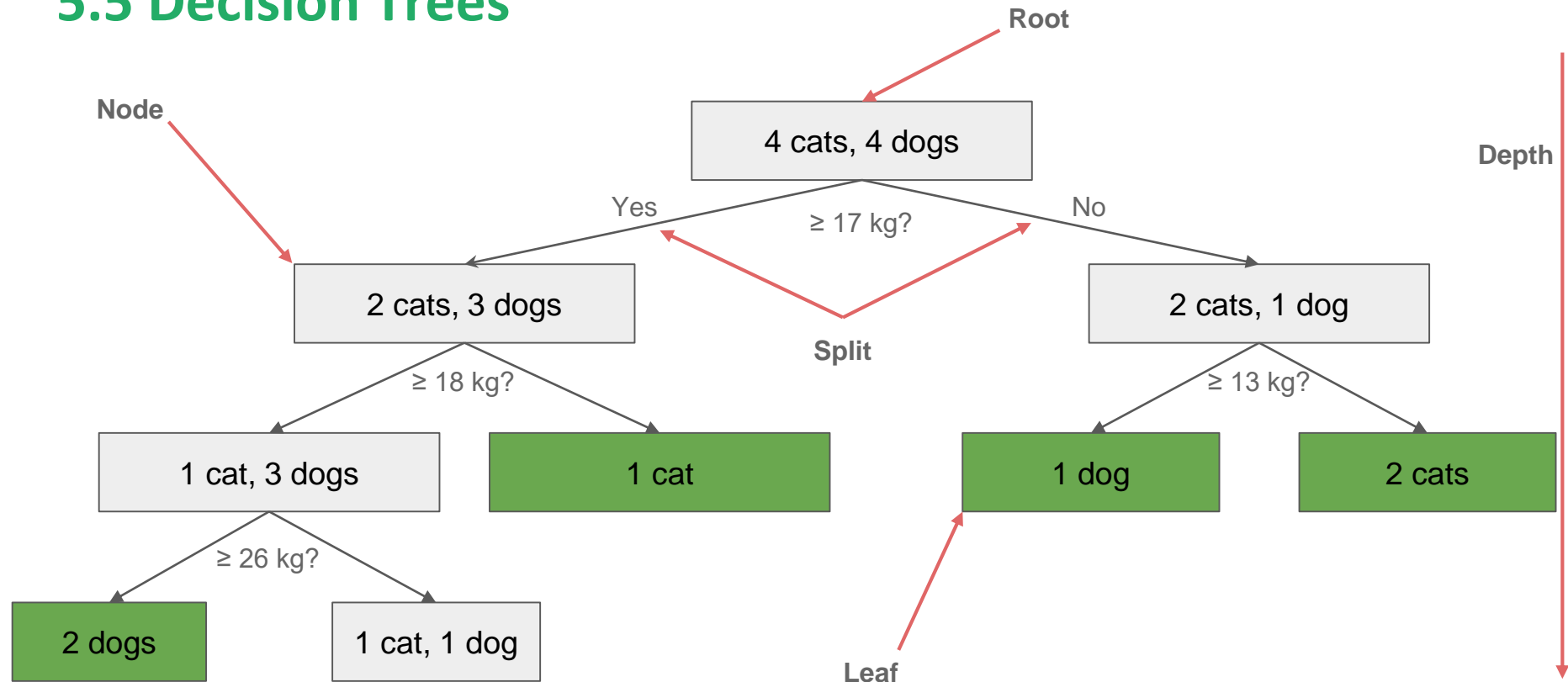
5.5 Decision Trees



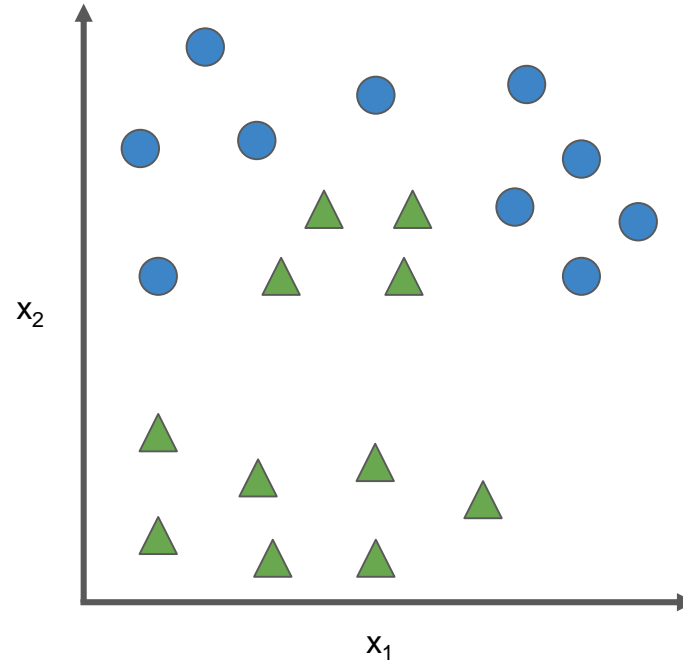
5.5 Decision Trees



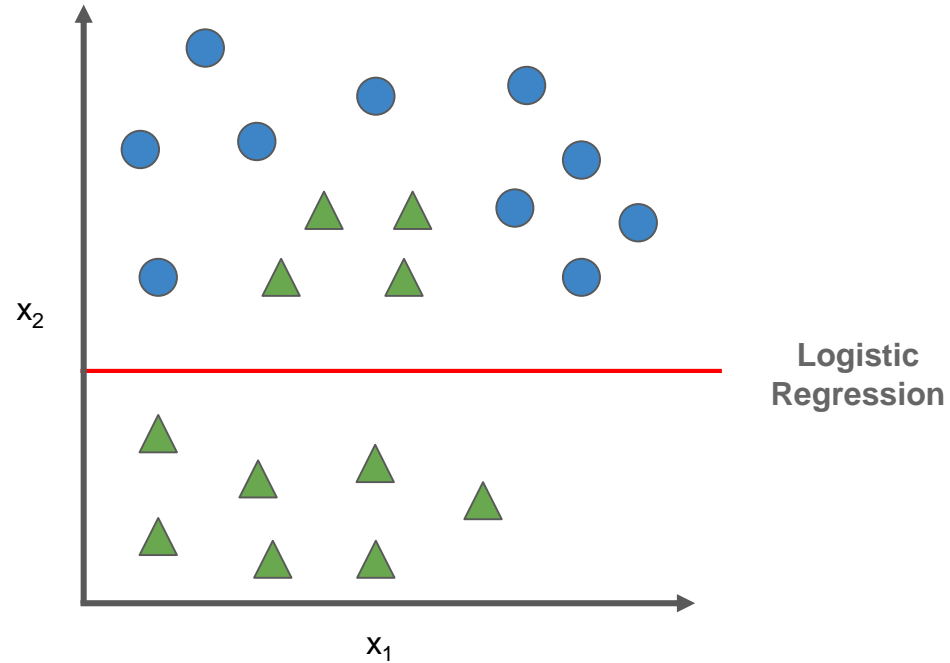
5.5 Decision Trees



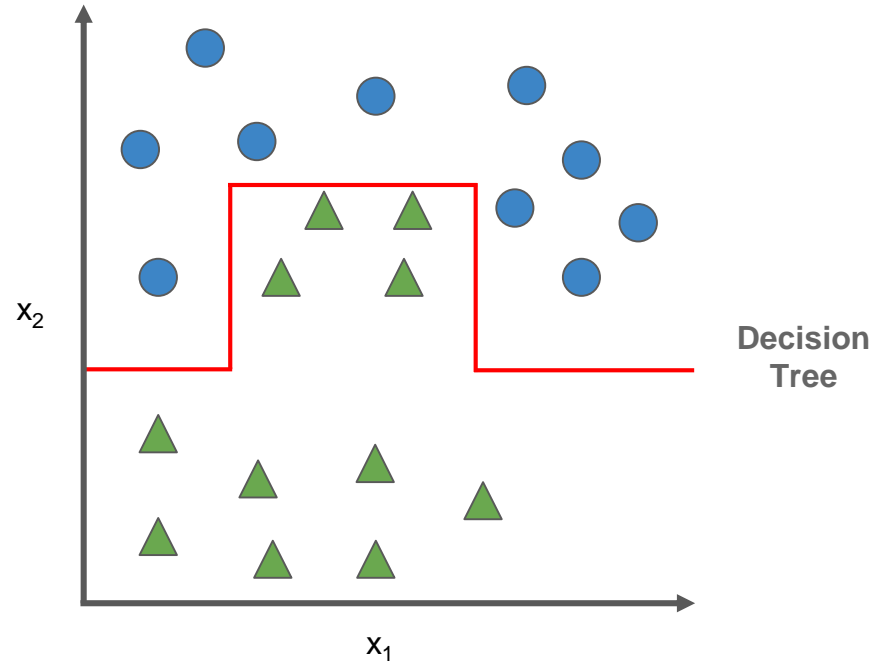
5.5 Decision Trees



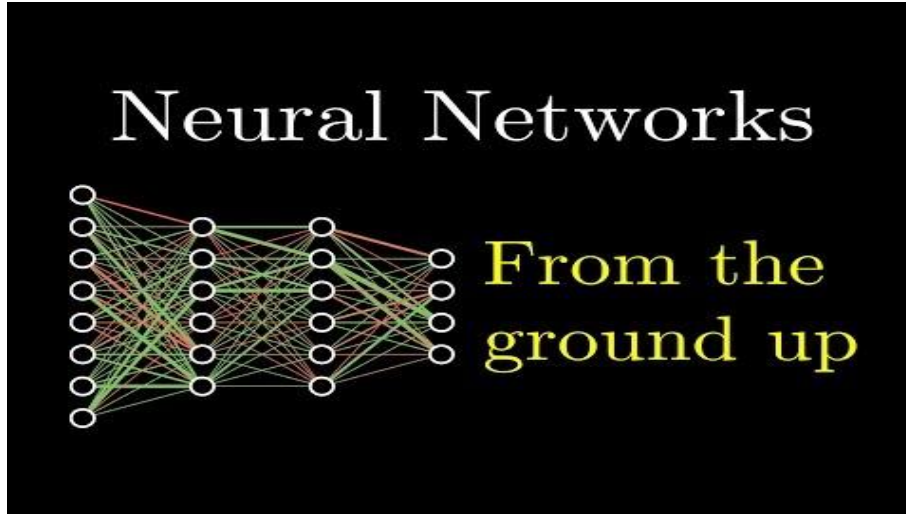
5.5 Decision Trees



5.5 Decision Trees



5.6 ML, Neural Nets, Deep Learning



5.6 ML, Neural Nets, Deep Learning

Let's try to understand the intuition behind some deep learning architectures:

- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)
- Long Short Term Memory (LSTM)
- Generative Adversarial Network (GAN)
- *etc.*

5.6 ML, Neural Nets, Deep Learning

Convolutional neural networks are mainly used to process images.

They can be used for:

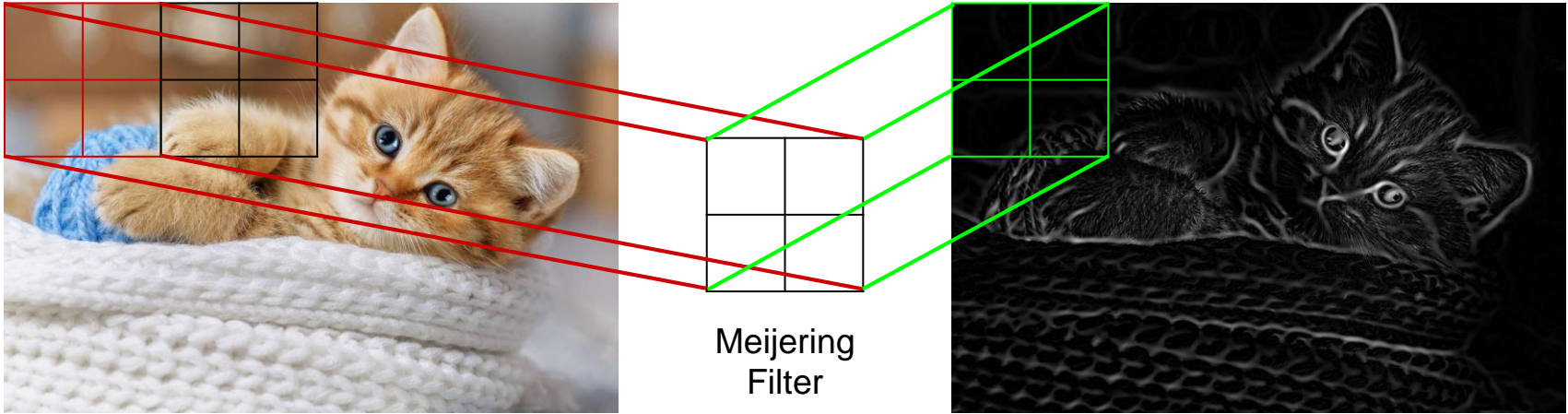
- Classification
- Segmentation
- Denoising
- etc.



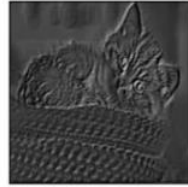
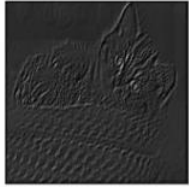
5.6 ML, Neural Nets, Deep Learning

A convolution designates the fact of scanning an image with a filter.

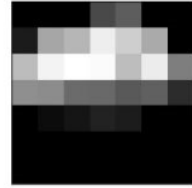
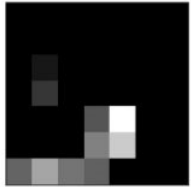
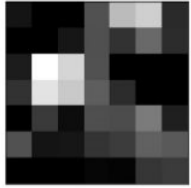
This filter is a small neural network that tries to "summarise" the image.



5.6 ML, Neural Nets, Deep Learning



5.6 ML, Neural Nets, Deep Learning



5.6 ML, Neural Nets, Deep Learning

RNNs are networks capable of managing sequences:

- Letters, words
- Images
- Sound, words
- ...



5.6 ML, Neural Nets, Deep Learning

⇒ You need a memory to predict the movement of an object

Where are they used?

- Voice recognition
- [Autonomous car](#)
- Text generation
- [Music generation](#)

LSTM, BI-LSTM are enhanced RNN

5.6 ML, Neural Nets, Deep Learning

GANs are recent architectures (2014)

The idea is simple, we put two networks in competition:

- A generator
- A discriminator

The objective of the generator is to produce information that the discriminator cannot identify as false.

5.6 ML, Neural Nets, Deep Learning

Where are they used?

- Deepfakes
- [Fake faces](#)
- Generation of fashion collections
- 3D modeling (Architecture, chemistry, pharmacy)
- ...



5.6 ML, Neural Nets, Deep Learning

Where are they used?

- Deepfakes
- [Fake faces](#)
- Generation of fashion collections
- 3D modeling (Architecture, chemistry, pharmacy)
- ...



5.6 ML, Neural Nets, Deep Learning

Reinforcement learning (RL) is a special case since the neural network has to manage itself:

- Without data
- Without rules

Data are inherent to the environment.

Only one thing is provided: **a reward**.

5.6 ML, Neural Nets, Deep Learning

Difficult to implement in reality:

- [Simulation](#)
- [Reality](#)

5.6 ML, Neural Nets, Deep Learning

Recent buzz in AI: Generative AI



Stable Diffusion Explained & Code

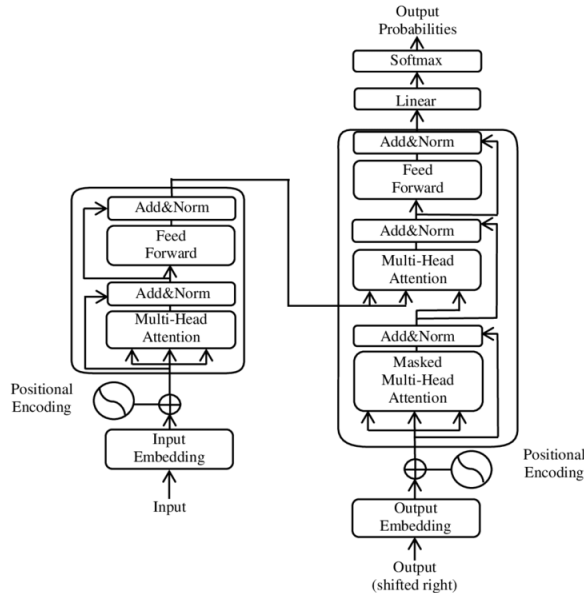
```
def produce_latents(text_embeddings, height=512, width=512, num_inference_steps=50, guidance_scale=7.5):
    if Latents is None:
        latents = torch.randn((text_embeddings.shape[0], height // 8, width // 8))
        latents = latents.to(device)

    scheduler.set_timesteps(num_inference_steps)
    latents = latents * scheduler.sigmas[0]

    with autocast('cuda'):
        for i, t in tqdm(enumerate(scheduler.timesteps)):
            # expand the latents if we are doing classifier-free guidance
            latent_model_input = torch.cat([latents] * 2)
            sigma = scheduler.sigmas[t]
            latent_model_input = latent_model_input / ((1 + sigma ** 2) ** 0.5)
```

5.6 ML, Neural Nets, Deep Learning

Recent buzz in AI: Generative AI



arXiv > cs > arXiv:1706.03762

Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7)]

Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Comments: 15 pages, 5 figures

Subjects: **Computation and Language (cs.CL)**, Machine Learning (cs.LG)

Cite as: [arXiv:1706.03762](https://arxiv.org/abs/1706.03762) [cs.CL]
(or [arXiv:1706.03762v7](https://arxiv.org/abs/1706.03762v7) [cs.CL] for this version)
<https://doi.org/10.48550/arXiv.1706.03762>

Submission history

From: Llion Jones [view email]

[v1] Mon, 12 Jun 2017 17:57:34 UTC (1,102 KB)

[v2] Mon, 19 Jun 2017 16:49:45 UTC (1,125 KB)

[v3] Tue, 20 Jun 2017 05:20:02 UTC (1,125 KB)

[v4] Fri, 30 Jun 2017 17:29:30 UTC (1,124 KB)

[v5] Wed, 6 Dec 2017 03:30:32 UTC (1,124 KB)

[v6] Mon, 24 Jul 2023 00:48:54 UTC (1,124 KB)

[v7] Wed, 2 Aug 2023 00:41:18 UTC (1,124 KB)

5.6 ML, Neural Nets, Deep Learning

Recent buzz in AI: Generative AI

arXiv > cs > arXiv:2001.08361

Computer Science > Machine Learning

[Submitted on 23 Jan 2020]

Scaling Laws for Neural Language Models

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei

We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the dependence of training speed on model size. These relationships allow us to determine the optimal allocation of a fixed compute budget. Larger models are significantly more sample-efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.

Comments: 19 pages, 15 figures

Subjects: **Machine Learning (cs.LG)**, Machine Learning (stat.ML)

Cite as: arXiv:2001.08361 [cs.LG]

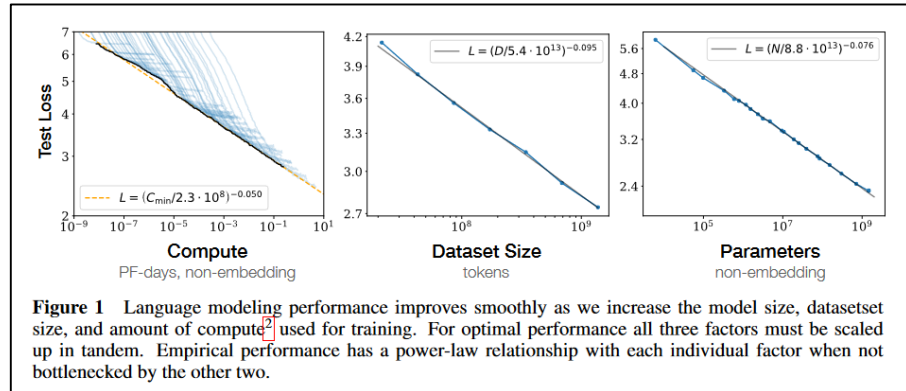
(or arXiv:2001.08361v1 [cs.LG] for this version)

<https://doi.org/10.48550/arXiv.2001.08361> 

Submission history

From: Samuel McCandlish [\[view email\]](#)

[v1] Thu, 23 Jan 2020 03:59:20 UTC (1,520 KB)



5.7 AI, ML Conclusions

First conclusions for ML:

- Is a narrow / specific artificial ("kind of") intelligence
- **ML = statistics + programming: (mainly) for prediction**
- Different algorithms: different use cases:
 - Tabular / structured vs unstructured, images, texts, sounds
 - Amount of data, labeled? Relation between data?
- Require lots of data → **data oriented**
 - Either after successful BI-chain has been established
 - Or 'plugged' closer to operational databases

6. Tools and Languages

What tools are used to train an AI?

- Databases
- Programming languages
- Libraries

6. Tools and Languages

A programming language allows to give a serie of instructions that the computer is able to interpret, to execute and eventually to return a result.

There are many of them and each one has its specificities.

```
def get_cnn(initializer='glorot_uniform') -> tf.keras.Model:
    """
    Build Simple CNN architecture for siamese Network.
    We Tried a lot of different architectures with Optuna for HP optimization but it seems really
    hard to find other viable architectures.

    2 Convolution layers and one Dense

    Args:
        | initializer (str): kernel initializer Used for layers. Defaults to glorot_uniform

    Returns:
        | tf.keras.Model: CNN model
    """
    inputs = Input(IMG_SHAPE)

    x = Conv2D(64, 4, activation='relu', kernel_initializer=initializer)(inputs)
    x = MaxPool2D()(x)

    x = Conv2D(64, 4, activation='relu', kernel_initializer=initializer)(x)
    x = MaxPool2D()(x)

    x = Flatten()(x)
    # x = GlobalAvgPool2D()(x)
    outputs = Dense(128, activation='relu', kernel_initializer=initializer)(x)

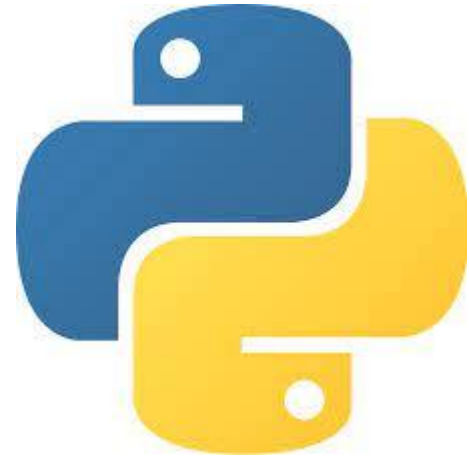
    cnn = keras.Model(inputs, outputs)
    return cnn
```


6. Tools and Languages

In the field of artificial intelligence, one language stands out from the rest, Python.

Why?

- (Relatively) easy to learn
- Lots of well-supplied and free libraries
- Big and active community
- Multitasking (~ universal)



6. Tools and Languages

However, there are other languages that can also be used (depending on the case, the company, the training, *etc.*).

- R, a language with very extensive statistical libraries
- C ++, a lower level language, but more powerful than Python. Very useful for embedded applications
- Julia, a rather young language that aims to combine the ease of use found in Python with the performance of C++



6. Tools and Languages

- Libraries should be seen as toolboxes
- They are already implemented code that you can reuse
- For example it is not necessary to recode a decision tree or a logistic regression

Machine learning



Deep learning

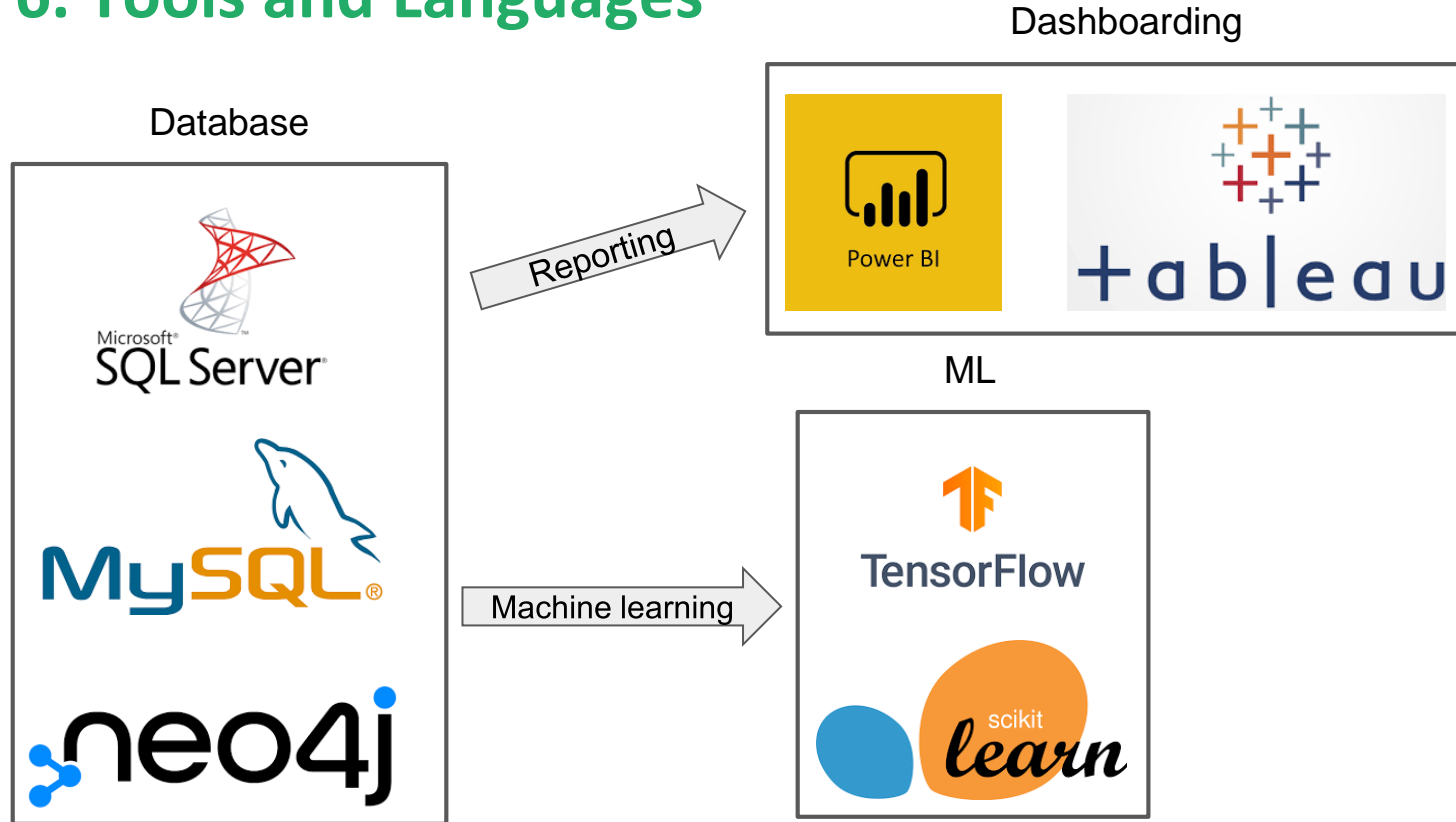


6. Tools and Languages

How to correctly choose the right tool with this multitude of languages, libraries, algorithms, *etc.*? There are several elements to take into account:

- The amount of data
- The nature of the problem (classification, regression, *etc.*)
- The complexity of the problem (2 classes vs 1000 classes)
- The available computing power (cloud servers or smartphone)
- *etc.*

6. Tools and Languages



6. Tools and Languages

There is a stone missing from the building.

How do we integrate our AI into a service, a website, an application?

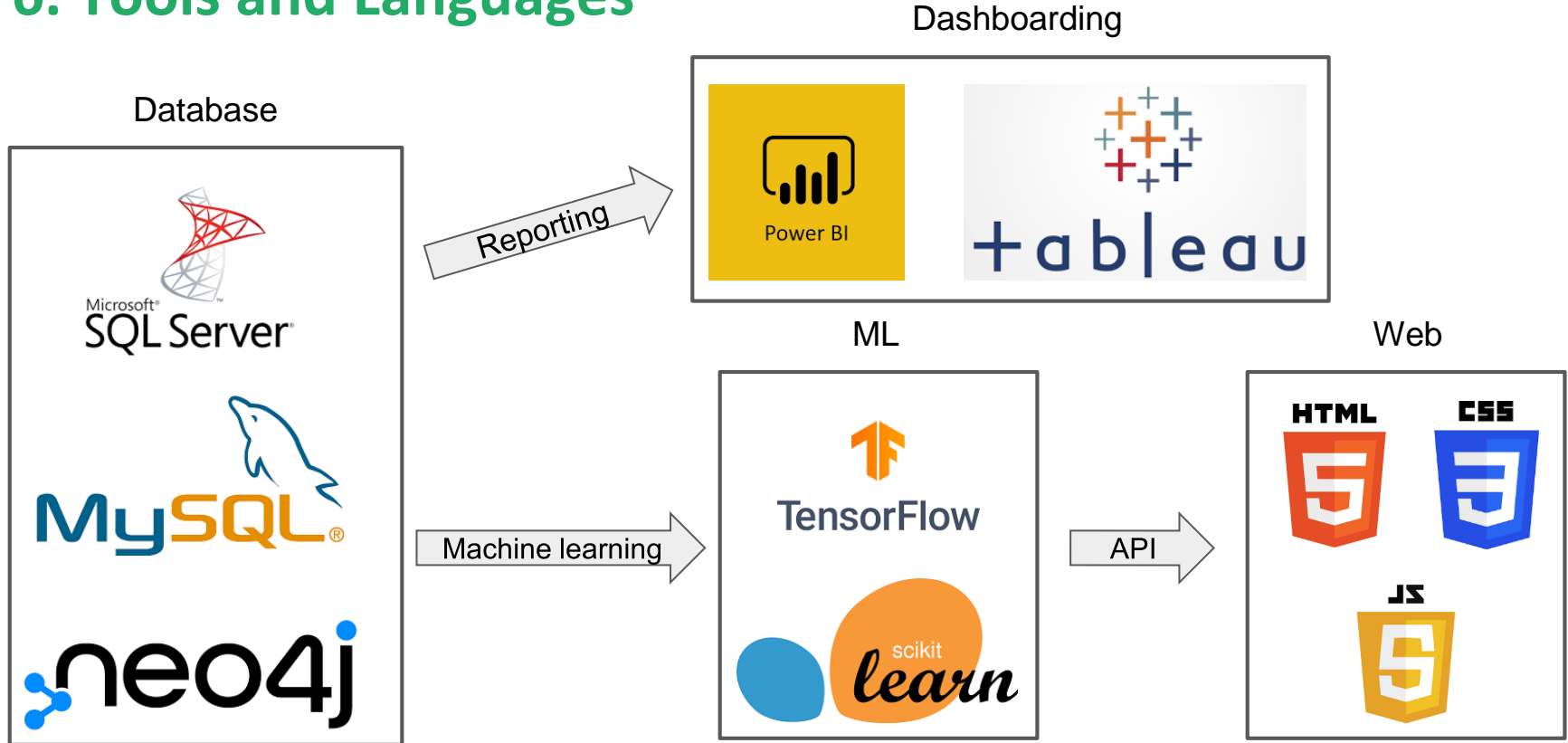
We can't ask someone to take a piece of Python code, enter his data and execute the code himself.

To do this we need web technologies like HTML, CSS, Javascript, Angular, etc.



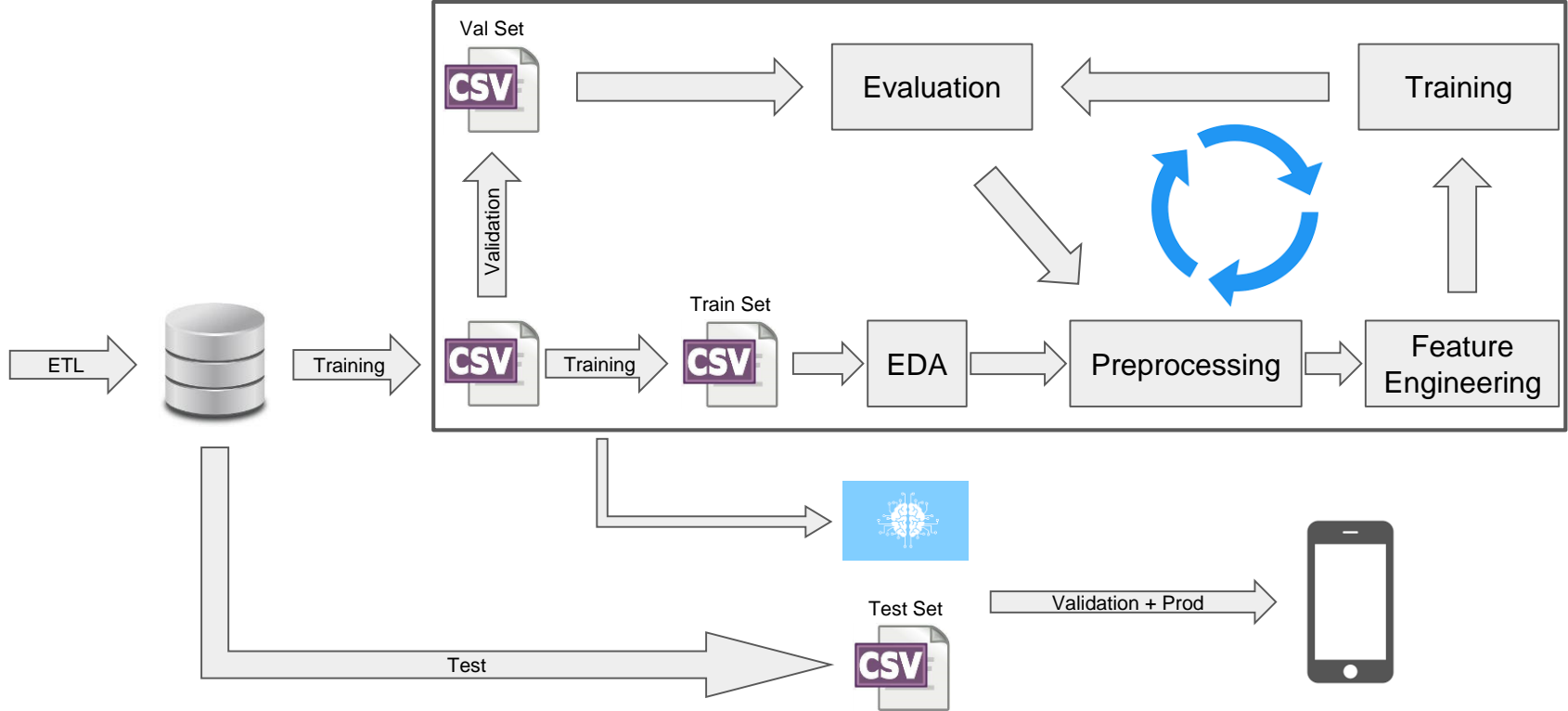
```
C:\> Invite de commandes - conda activate base  
  
(base) C:\Users\romain\Desktop>python main.py  
J'estime votre maison à 200000 €
```

6. Tools and Languages



7. Process of a Project

odoo



7. Process of a Project

1. Identification of the problem

- What is the problem?
- Can it be solved with a machine learning solution?
- Is it the most adequate solution?
- Do we have the means to do it?

7. Process of a Project

2. Audit on the data

- Do we have the necessary data? In sufficient quantity? Labeled?
 - If not, can we collect them, how, how much?
- What is their nature?
- Where are they?
- Can we use everything?

7. Process of a Project

3. ETL

- **E**xtract: extract the necessary data from the different available sources
- **T**ransform: this part consists in cleaning your data and homogenising them
- **L**oad: load all the data in a datawarehouse for example

→ There are different tools dedicated to this task: Python, Talend, Pentaho, SSIS, etc.

7. Process of a Project

4. Constitution of datasets

After the constitution of your dataset, it is **IMPERATIVE** to subdivide it into two sub datasets:

- Train Set
- Test Set

You must at all costs keep a dataset aside that your model will never see!

7. Process of a Project

5. **E**xploratory **D**ata **A**nalysis (EDA)

This is about analysing your data for a variety of reasons:

- Understand what they are talking about specifically
- Identify potentially interesting variables
- Identify modifications to be made for the next step
- Identify some potential algorithms
- etc.

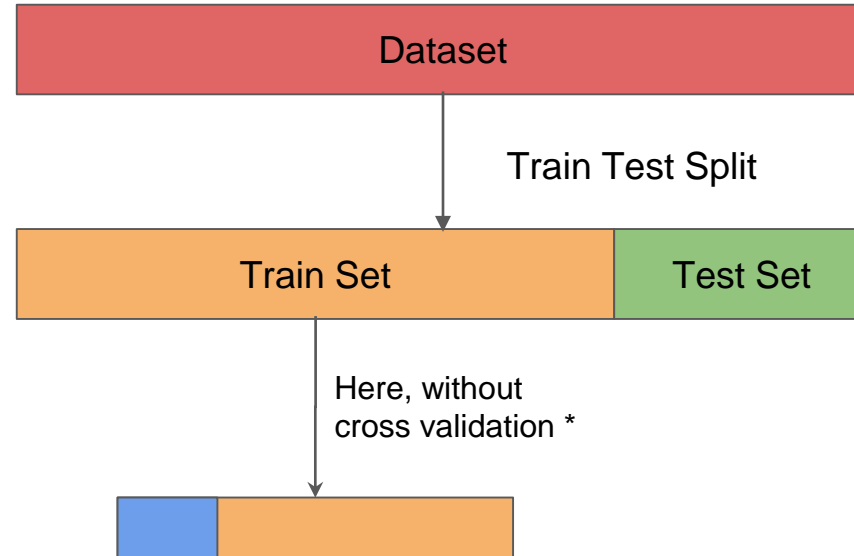
7. Process of a Project

6. Constitution of a validation set

At this stage, you have a choice of several algorithms and it is impossible to know in advance which one will do the best.

This is what your validation set is for.

7. Process of a Project



* Better with cross validation, if possible; depending on ML-Algo

7. Process of a Project

7. Preprocessing

With the next step, this is the heart of the work.
Poor quality data can only lead to bad results.

"Garbage in, garbage out"

- What to do with text variables, categories?
- What if I encounter missing values?
- Are there encoding errors?
- etc.

7. Process of a Project

8. Features engineering

We could have merged this step with the previous one, but this step consists of:

- Selecting the variables
- Discretising some variables
- Extracting variables
- etc.

7. Process of a Project

9. Selecting a model and putting it into production

Some points of attention during deployment:

- Are the calculations done on cloud servers?
 - What happens without an Internet connection?
- Are the calculations performed on the users' devices?
 - Do we need to refactor the code?
- An hybrid solution?

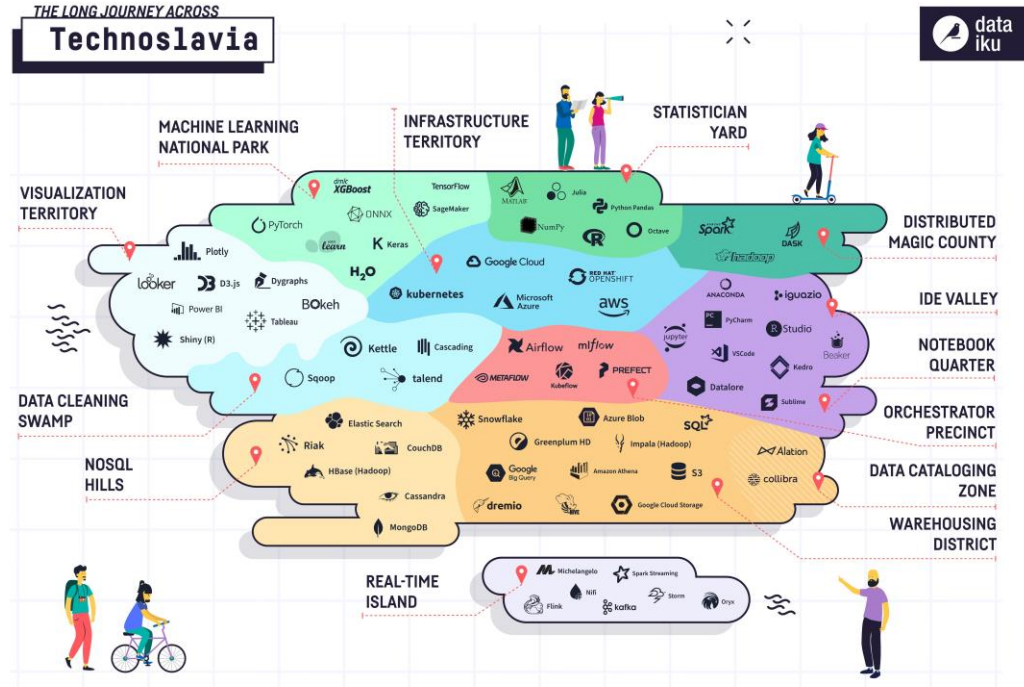
7. Process of a Project

10. Monitoring

Once in production, you have to keep an eye on your model and often re-train it for two main reasons:

- Data drift
- Appearance of new classes

7. Process of a Project



What does the future of the data world look like?

An empire in which a single language dominates all others?
Or a unified world where multiple republics of data technologies coexist and complement each other?
At Dataiku, we believe in the latter. And in this world (which we've dubbed Technoslavia), data teams need a reassuring guide; a tool that is always on the forefront of the latest technologies and that also unites all profiles, from data scientist to analyst to IT.



8. Jobs and ML

Depending on what constitute the project with ML, different skilled labours are possible to implement and/or maintain and/or deploy those ML results:

- Academics / Scientists / Engineers / ... in big tech:
 - Invention of **new** algorithms
 - Profile caricature: PhD in mathematics (or equivalent)

8. Jobs and ML

- Data Scientists (= expert in AI/ML outside academia):
 - Good programming & good mathematics/statistics skills
 - Application and deep understanding of well known ML/DL algorithms
 - Ability to keep up with new techniques/tools/algorithms
 - Can implement **structured data** with **ease**
 - DL: **specialisation** into **one unstructured data** (no deep understanding in *all of them*)
 - Full project pipeline (+ potentially full BI-chain*)
 - More interpretation of results, less implementation into Apps

8. Jobs and ML

- Data Analysts:
 - Mastering of BI-chain (operational DB → reporting)
 - Some dabble in simple ML (structured data) projects: decision trees, linear regressions, ...
 - Some can be trained into Data Scientists (! math/stats)
- ML Engineers:
 - Overlap(*) with Data Scientists, but is not as comfortable with the BI-chain
 - \pm = Devs \Leftrightarrow data
 - Less interpretation, more implementation into Apps

9. Ethics and Laws

AI raises a lot of questions.

The idea in this chapter is not to answer them but to initiate an awareness through some examples.

Let's start with a game:

- [Moral machine](#)

9. Ethics and Laws

Try answering the following questions:

- Does your position with regards to AI hold under all circumstances?
 - Identify borderline cases
 - Suggest these edge cases to your colleagues
- Outside of autonomous cars, think of events, applications, cases that might also raise these kinds of questions?

9. Ethics and Laws

[In September 2021 on Facebook](#), after watching videos with black people, Facebook's algorithm asks if they want to see more videos of monkeys...

Twitter and Google have also encountered this issue.

The question which arises is how to succeed in not reproducing the discrimination the society is facing?

9. Ethics and Laws

Recommendation algorithms suggest the content that is supposed to appeal to you the most. When it comes to Netflix, Spotify, this may seem anecdotal.

However, when it comes to politics, especially during elections, Twitter tends to favour certain ideas, to lock users in a bubble.

When Facebook bans accounts, deletes posts, [etc.](#), is this always desirable?

9. Ethics and Laws

These algorithms are based on data and are therefore sensitive to it.

This means that it is totally possible to alter their behaviour by providing them with bad data (data poisoning).

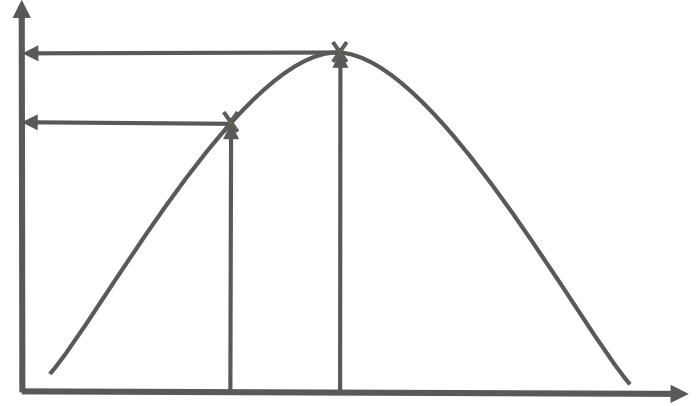
This is exactly what some political parties could do by using robots.

9. Ethics and Laws

These recommendation algorithms call into question individual autonomy.

Since Netflix offers you the best possible choice, in reality you no longer have a choice.

These algorithms tend to make you choose the optimal choice, which implicitly means taking away the action of choosing.



9. Ethics and Laws

What is the current status?

- [Regulations](#) are being put in place
- Research is *slowly* developing in universities
- Emerging [libraries](#) are providing tools to help solve these problems

But unfortunately, there are regular steps [backwards](#)

9. Ethics and Laws

In addition to this, there are other regulations such as the [General Data Protection Regulation](#) (GDPR) that govern the collection and use of data:

- You can ask to delete your data
- Ethnic censuses are prohibited
- Asking or seeking to know the health status (HIV, pregnancy, etc.) of a person is prohibited. Except for research purposes
- Companies can only collect the data they need

9. Ethics and Laws

The privacy shield is a device that allows data of European citizens to be sent to the United States.

On July 16, 2020, this agreement was invalidated by a [decision](#) of the Court of Justice of the European Union.

However, [since March 2022](#), new talks regarding its successor are under way.

9. Ethics and Laws



EU guidelines:

<https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>

Bonus: Deeper into the Analytics

A.1 Model Selection

Define a model:

⇒ At a minimum it is a classification algorithm, a regression, etc.

→ In general it is a succession of processes including an algorithm:

- Data transformations
- A training of our algorithm
- A result

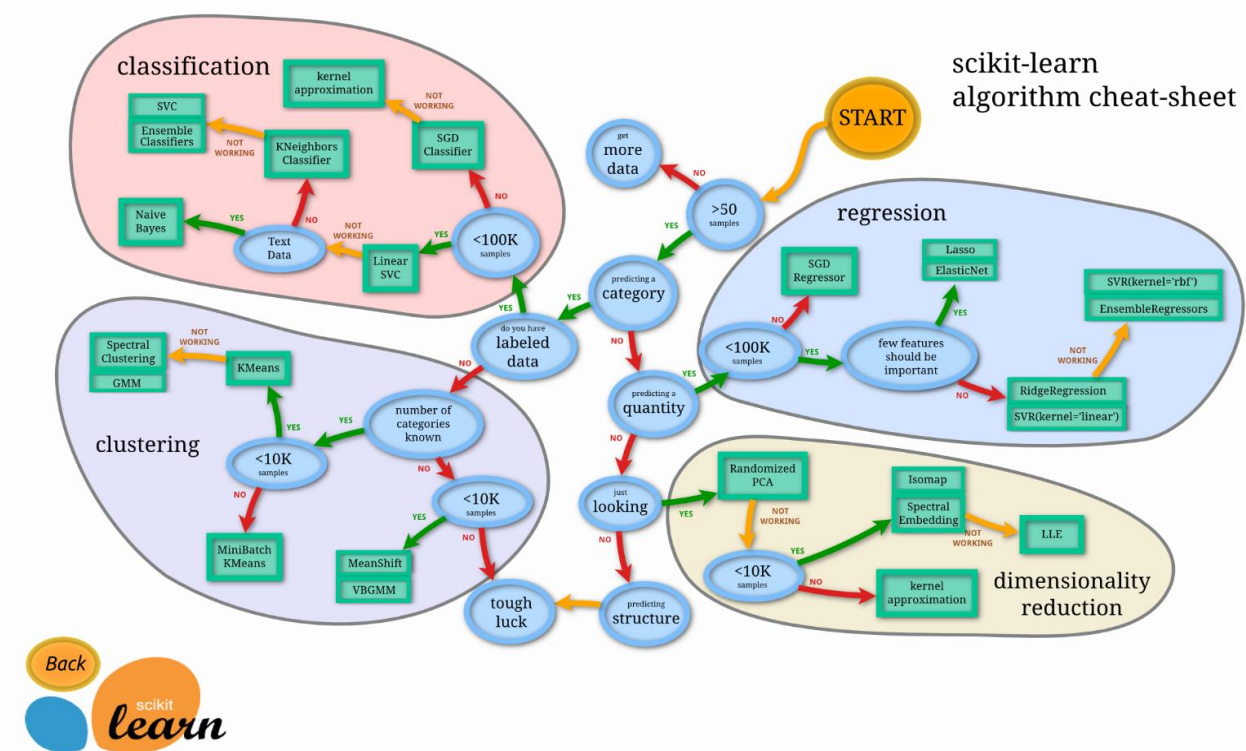
A.1 Model Selection

KNN, Logistic Regression, Linear Discriminant Analysis, Decision Tree, SVM, Ridge, Lasso, Naive Bayes, etc.

How to choose a model?

- **Identify** your problem (Clustering, Regression, Classification)
- **Empiricism**, trial and error method
- [Cheat Sheet](#)
- With a little experience, we know more or less which models to choose

A.1 Model Selection



A.1 Model Selection

`sklearn.model_selection.train_test_split`

In machine learning we **systematically** separate our dataset into two sub-datasets:

- The train set: is used as a basis for training a model
- The test set: is used to evaluate our model **AFTER** having trained it correctly

This allows us to make sure that our model manages to **generalise well** and did **not learn by heart** the train set

→ Shuffle the dataset before splitting

A.1 Model Selection

Our models often have *hyperparameters*:

- Those hyperparameters are defined by the Devs / Data Scientists and not altered during optimisation
- For example: DT's max_depth, min_sample_leaf, ...
- For example: in PolyReg: polynomial order N
- "Different hyperparameters define different versions of the same underlying (family of) model/algorithm"



Parameters are the things that are optimised during the training phase:

- For example: LinReg (PolyOrderNbr = 1): a, b in $y \sim f(x) = ax + b$
- For example: DT: the chosen criteria for each node (column & threshold)

A.1 Model Selection

Can we optimise these 'hyperparameters' (HP) too? In conventional ML, yes.

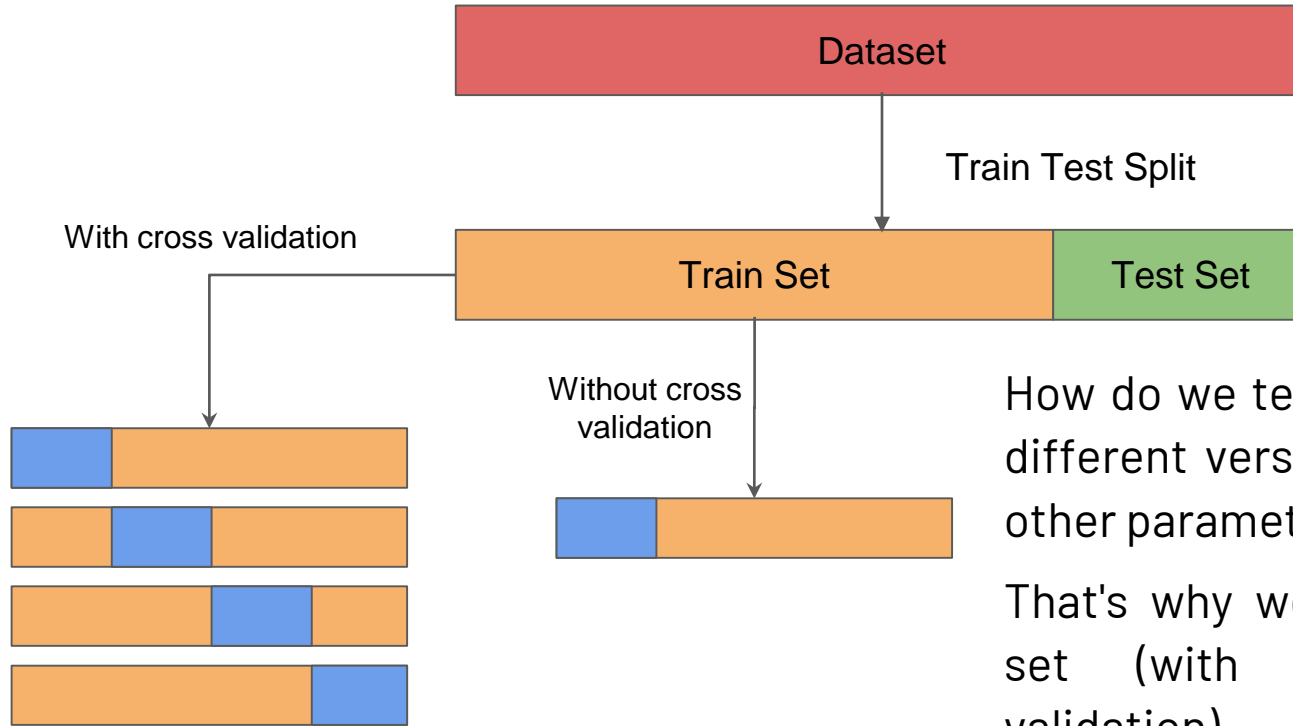
- GridSearchCV allows you to test a model with multiple HP values
- RandomizedSearchCV works the same way but does not test all possibilities

We often declare

- A range of possible hyperparameters (not $-\infty$ to $+\infty$),
- Or keywords like 'f1_score', 'Euclidean metric',...

depending on the chosen algorithm.

A.1 Model Selection



How do we test and compare those different versions, while optimising other parameters?

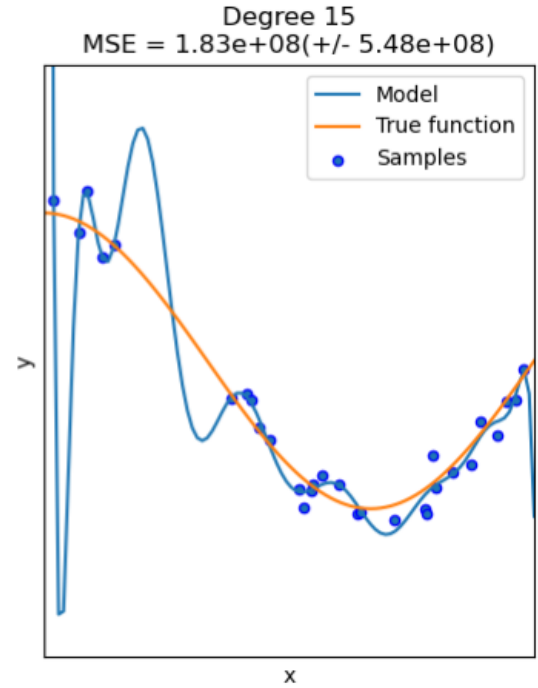
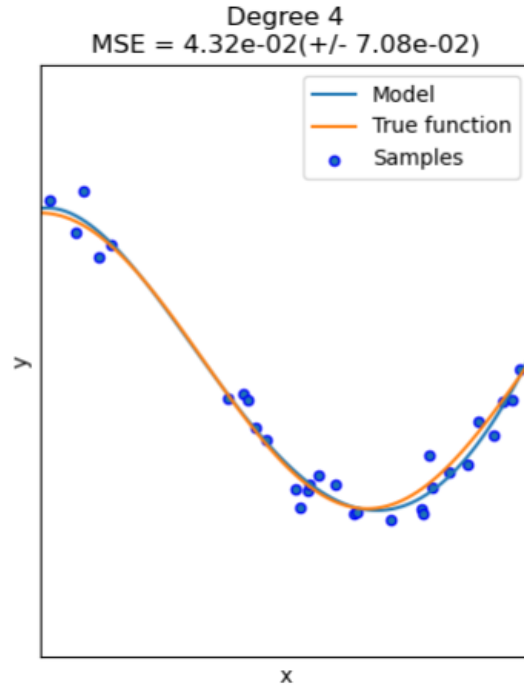
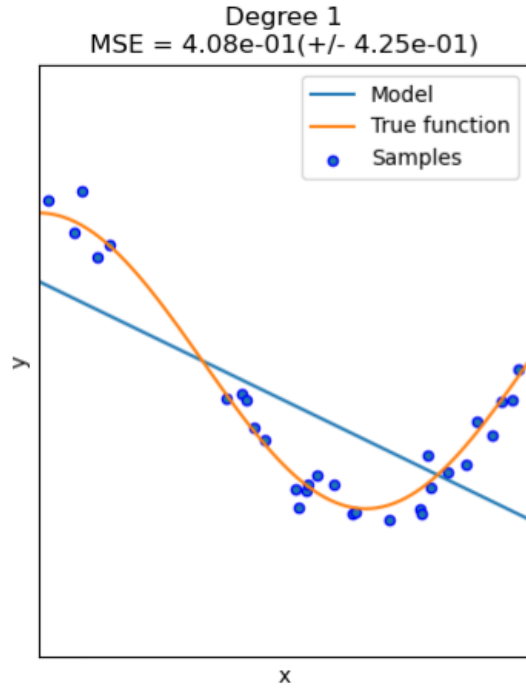
That's why we need the validation set (with or without cross validation).

A.1 Model Selection

There are two cases where our model does not generalise well:

- Overfitting: model with very **good** performances on the **train set** but **bad** performances on the **validation set**
- Underfitting: poor performances on the **train set and validation set**

A.1 Model Selection



A.2 Metrics

So far we have been talking about the "performances" of a model, but what exactly is it?

It depends on:

- The problem (i.e. classification, regression, etc.)
- The chosen metric
- The type of metric (score vs loss)
 - Maximise scores
 - Minimise losses

The choice of the metric is **crucial** and is made before the choice of the model

A.2 Metrics

sklearn.metrics.accuracy_score

Score between 0 and 1

Total number of correct predictions / Number of observations

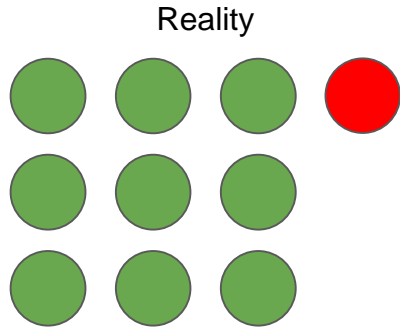
$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

A.2 Metrics

Accuracy can be deceiving.

Let's say you have to develop a cancer screening test.

Out of a sample of 100 individuals, you know that 10 have cancer



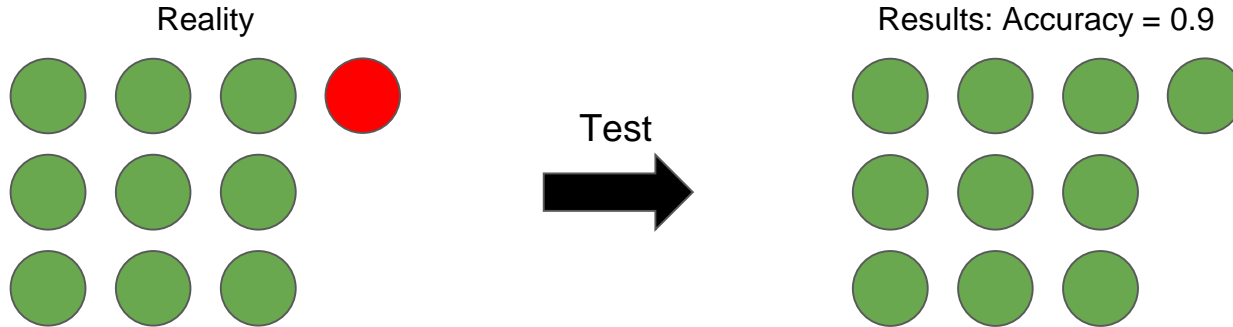
What is the potential problem with accuracy?

A.2 Metrics

Accuracy can be deceiving.

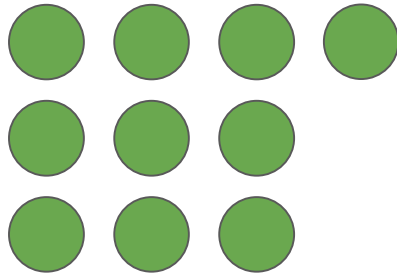
Let's say you have to develop a cancer screening test

Out of a sample of 100 individuals, you know that 10 have cancer



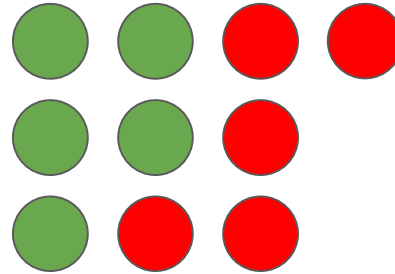
A.2 Metrics

Results 1: Accuracy = 0.9



- The test misses all the patients
- To clear all doubts you need to take **100 X-Rays**

Results 2: Accuracy = 0.6



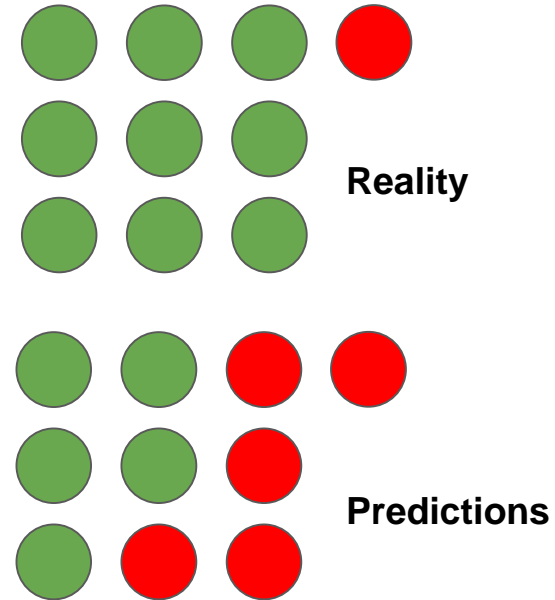
- The test detects all the patients + some of the non patients
- To remove all doubts, you only have to take **50 X-Rays**

A.2 Metrics

Confusion matrix

Allows to compare predictions to reality

Pred \ Actual	Healthy	Cancer
Healthy	5	0
Cancer	4	1



A.2 Metrics

`sklearn.metrics.precision_score`

Precision (not to be confused with accuracy) is a measure of exactitude.

It is a more micro measure, a `precision_score` per class.

In other words, when my test tells me this class, what is the probability that this is really the case.

Pred \ Actual	Healthy	Cancer
	Healthy	Cancer
Healthy	5	0
Cancer	4	1

A.2 Metrics

precision_score healthy class = 1

- $5/5 = 1$
- The test predicted 5 healthy people and these 5 people are really healthy

precision_score cancer class = 0.2

- $1/5 = 0.2$
- The test predicted 5 cancers but only one person actually has cancer

Pred \ Actual	Healthy	Cancer
Healthy	5	0
Cancer	4	1

A.2 Metrics

`sklearn.metrics.recall_score`

Recall or sensitivity is a measure of completeness.

Probability of finding all individuals in a class.

In other words, the probability of not missing individuals of a class.

Pred \ Actual	Healthy	Cancer
	Healthy	Cancer
Healthy	5	0
Cancer	4	1

A.2 Metrics

recall_score healthy class = 0.56

- $5/9 = 0.56$
- The test found 5 of the 9 really healthy people

recall_score cancer class = 1

- $1/1 = 1$
- The test found 1 of the 1 person(s) really cancerous

Pred \ Actual	Healthy	Cancer
	Healthy	Cancer
Healthy	5	0
Cancer	4	1

A.2 Metrics

sklearn.metrics.f1_score

This is a metric that summarises precision and recall.
In more statistical terms it is an harmonic mean:

$$\mathbf{f1_score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

$$\mathbf{f1_score} \text{ healthy class} = 2 * (1 * 0.56) / (1 + 0.56) = 0.71$$

$$\mathbf{f1_score} \text{ cancer class} = 2 * (0.2 * 1) / (1 + 0.2) = 0.33$$

A.2 Metrics

sklearn.metrics.mean_squared_error (MSE)

→ Difference between reality and prediction squared divided by the number of observations

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

A.2 Metrics

$y_{\text{test}} = [1, 5, 6, 10, 11]$ (ex. units in meter)

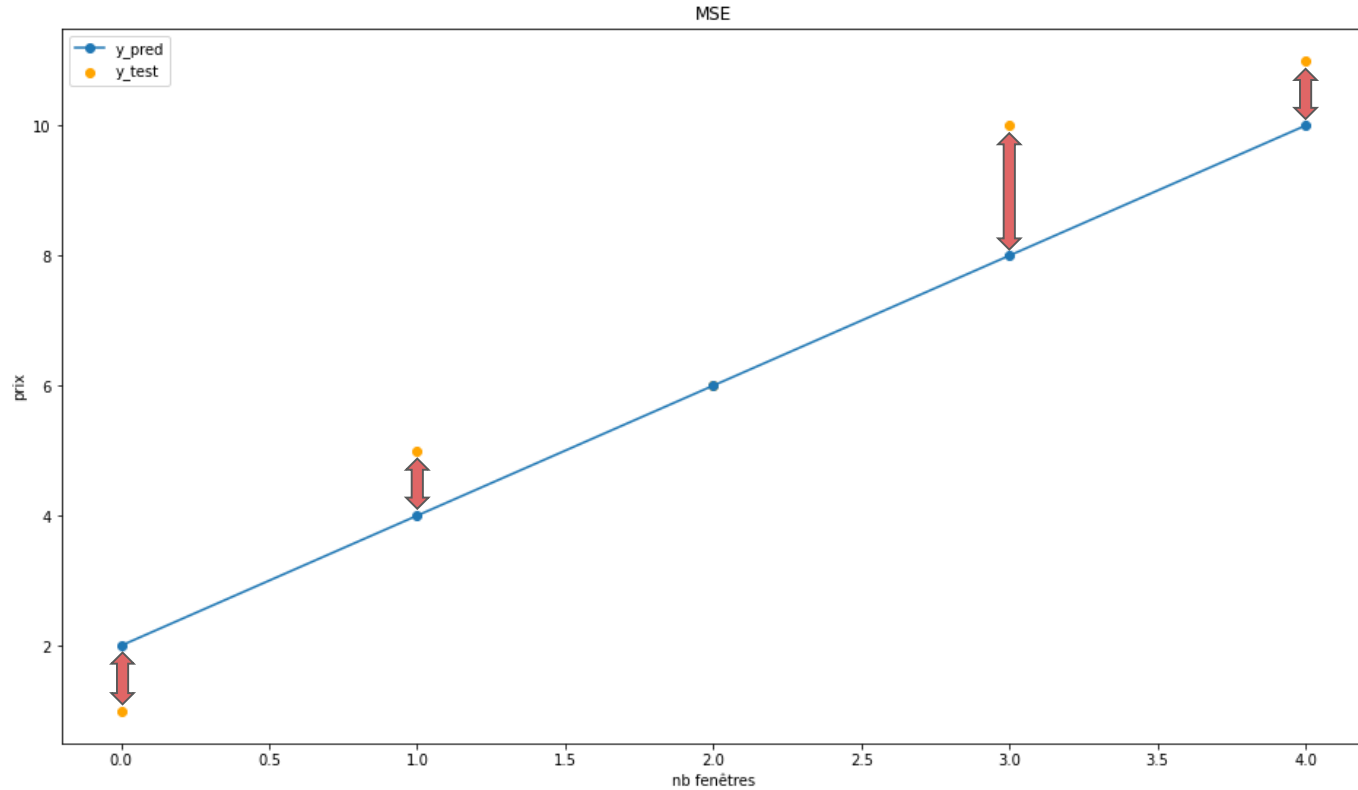
$y_{\text{pred}} = [2, 4, 6, 8, 10]$

$$\begin{aligned} \text{MSE} &= ((1 - 2)^2 + (5 - 4)^2 + (6 - 6)^2 + (10 - 8)^2 + (11 - 10)^2) / 5 \\ &= 1.4 \quad (\text{units in meters}^2) \end{aligned}$$

just take the square root to find the original scale.

Root Mean Squared Error (RMSE) = 1.18 (units in meters)

A.2 Metrics



A.2 Metrics

sklearn.metrics.r2_score

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

With

- The squared error in the numerator
- The variance in the denominator

A.2 Metrics

The smaller your squared error compared to the variance, the closer the r^2 is to 1

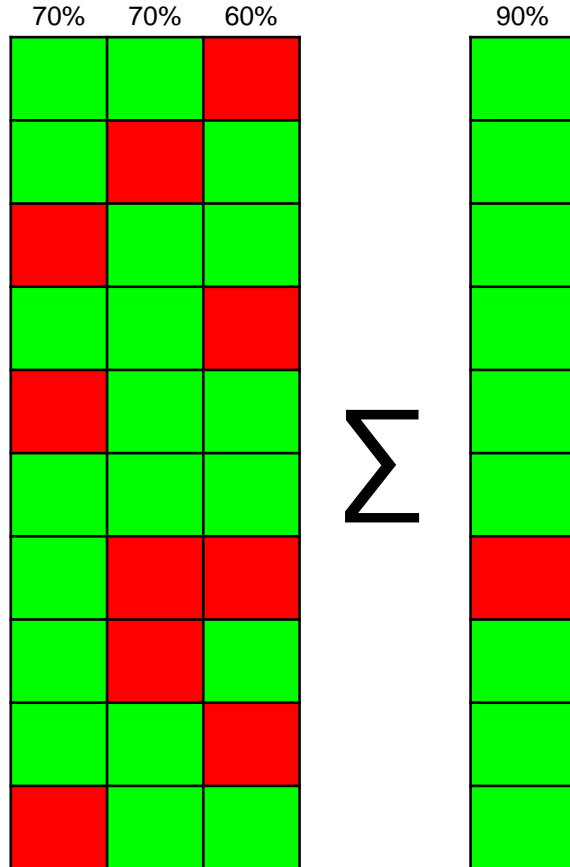
$$\mathbf{r^2_score} = 1 - (2 / 20) = 0.9$$

⇒ My model explains 90% of the variation in my dataset

⇒ The higher the R^2 , the more the difference between your predictions is linked to the variance and not to the errors of your model

→ Is the difference between my measurements related to the errors of my model or to the variance of my distribution?

A.2 Wisdom of the Crowd



None of the 3 models exceed 70% of good classifications.

If we combine the results we reach 90%.

Wisdom of the crowd

A crowd of enlightened people will be right more often than a single expert.

Applicable to machine learning

Warning: the performance of each model must be at least 51%.

Otherwise convergence to 0.

A.2 Wisdom of the Crowd

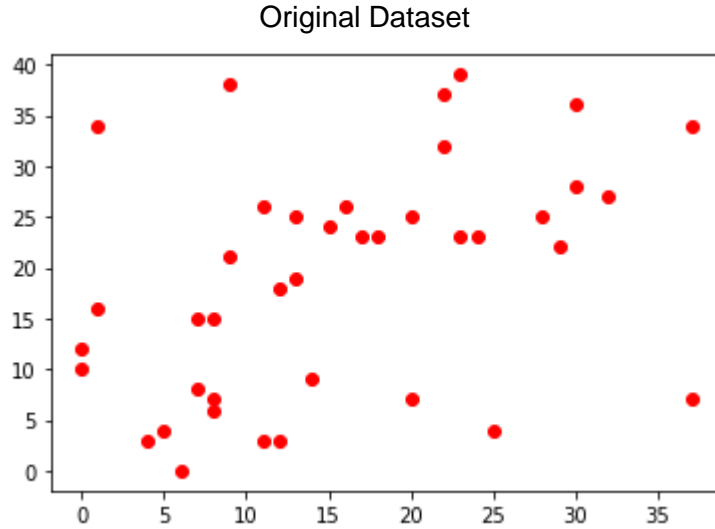
`sklearn.ensemble.Bagging`

Bootstrap **A**ggregating

The idea here is to train a crowd of the same model but using **bootstrapping**.

Bootstrapping: sampling method where individuals are randomly drawn and replaced after.

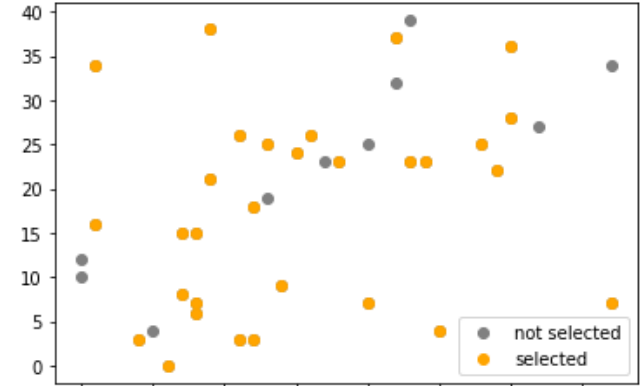
A.2 Wisdom of the Crowd



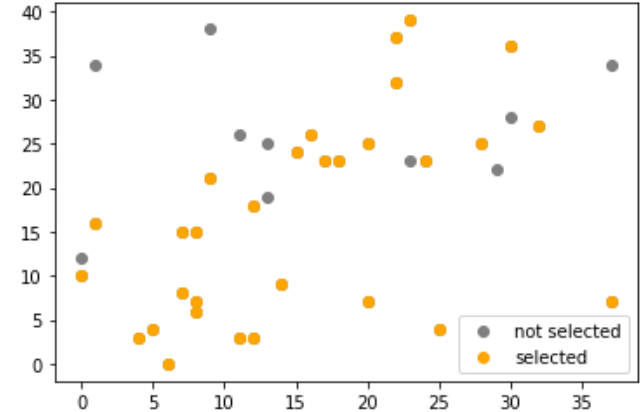
Bootstrapping



Model 1



Model 2



A.2 Wisdom of the Crowd

The models therefore share part of the dataset information.

The idea is to "**average**" the result of all the models.

One of the best known models: **sklearn.ensemble.RandomForest**

- The idea is to train a set of decision trees

A.2 Wisdom of the Crowd

sklearn.ensemble.Bagging

sklearn.ensemble.AdaBoost

A succession of weak models is trained.

On the basis of the errors of the previous one, we modify the weight of the individuals.

More weight is given to the errors.

A.2 Wisdom of the Crowd

XGBoost

Star of the boosting: XGBoost (e**X**treme **G**radient **B**oosting):

- Much more optimised than sklearn trees (more hyperparametrisations)
- sklearn compatible
- Computing on GPU
- Also [available](#) in other languages such as C++, C, Julia, Ruby, etc.

A.2 Wisdom of the Crowd

Bagging:

- **Parallel** training
- Crowd of **experts**
- Individually, the models are **overfitting**
- Crowd reduces overfitting/variance

Boosting:

- **Successive** training
- Crowd of **weak** models
- Taken individually, the models are **underfitting**
- Reduces underfitting/bias

A.2 Wisdom of the Crowd

sklearn.ensemble.stacking

We train **different** models

Then, we train a model **on top** of these models that must determine who is right.

The models must be different and not make the same mistakes.