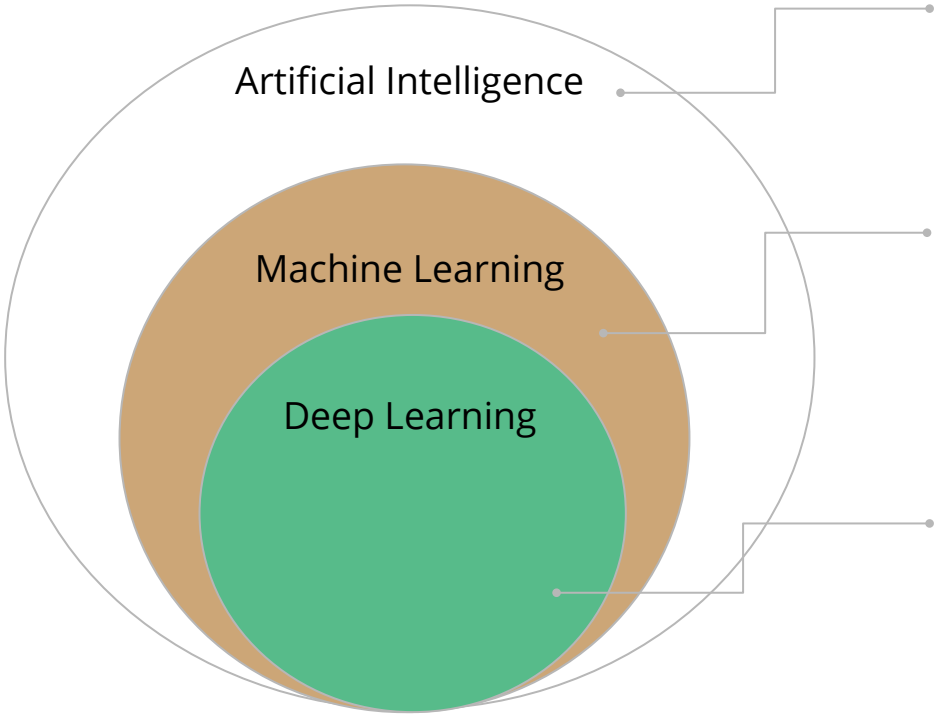
Why in **Machine Learning** & **Deep Learning** not all models are good

Luis González Jaime

esLibre 2019

21/06/2019

A little bit of history



Artificial Intelligence
Mimicking the intelligence or behavioural pattern of humans or any other living entity

Machine Learning
A technique by which a computer can “learn” from data. This approach is mainly based on training a model from datasets.

Deep Learning
A technique to perform machine learning inspired by our brain’s own network of neurons.

Artificial Intelligence

In 1956, John McCarthy, picked the name 'Artificial Int
field.

In **Dartmouth Workshop**, he introduced the term in an

"The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. [...] To solve kinds of problems now reserved for humans, and improve themselves".

Source: https://en.wikipedia.org/wiki/Dartmouth_workshop

Advances in few decades

1996-1997 - Kasparov vs. Deep Blue

Deep Blue versus Garry Kasparov was a pair of six-game chess matches between world chess champion Garry Kasparov and an IBM supercomputer called Deep Blue.

2005 - Autonomous vehicles

In the second competition of the *DARPA Grand Challenge*, was the first time **5** vehicles successfully completed the 212 km.

The vehicle *Stanley*, from **Stanford Racing Team** won the first place.

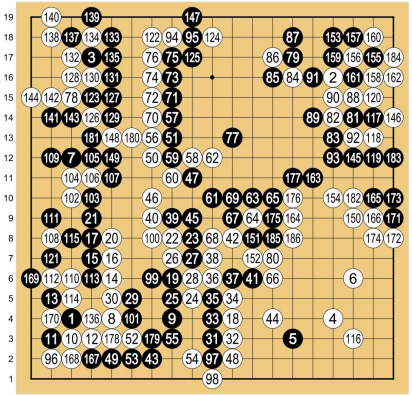
Sources: https://en.wikipedia.org/wiki/Deep_Blue_versus_Garry_Kasparov / https://en.wikipedia.org/wiki/DARPA_Grand_Challenge

Advances in few decades

2016 - AlphaGo

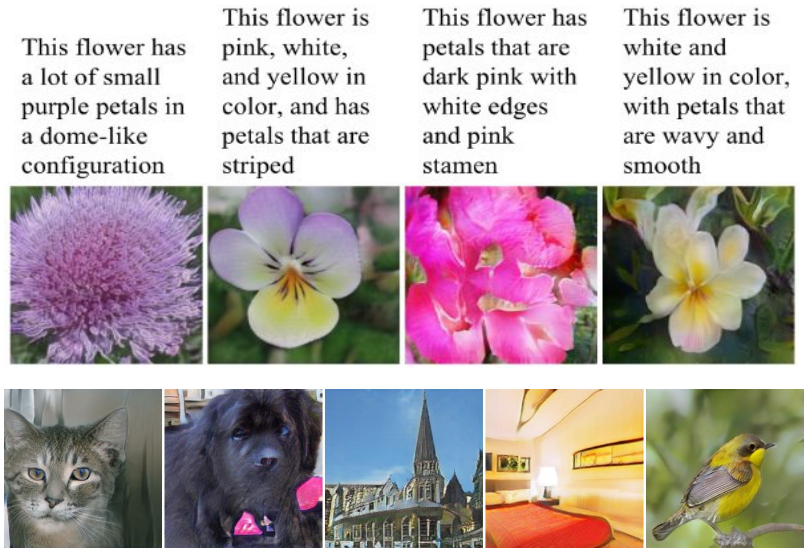
AlphaGo is a computer program that plays the board game **Go**.

In March 2016, the first time a computer Go program has beaten a 9-dan professional without handicaps.



Lee Sedol (B) vs AlphaGo (W) - Game 1

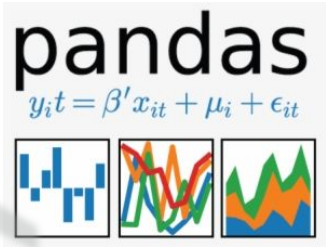
2018 - Generating Images from Text Descriptors



Sources: https://en.wikipedia.org/wiki/AlphaGo / https://github.com/hanzhanggit/StackGAN-v2

Open-source Technologies

- Keras [MIT]
- TensorFlow [Apache License 2.0]
- Pandas [New BSD License]
- DeepVariant [BSD-3 Clause]
- Pythorch [BSD]
- OpenCV [BSD License]
- Scikit-learn [New BSD License]



TensorFlow

PYTORCH

Not so good (or unexpected) results

2017 - Facebook's AI robots shut down after they start talking to each other in their own language

A piece of the conversation between the robots:

```
Bob: i can i i everything else . . . . . .

Alice: balls have zero to me to me to me to me to me to
me to me to me to

Bob: you i everything else . . . . . .

Alice: balls have a ball to me to me to me to me to me
to me to me
```

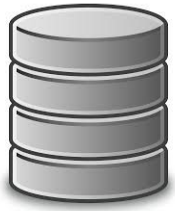
2015 - Amazon scraps secret AI recruiting tool that showed bias against women

The company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way. That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes.

Sources:
https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html /
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

We have our database...

I hear I can apply Machine Learning and Deep Learning...



Our Database

Definition of Machine Learning

Although there is not just one accepted definition, we will take the definition from **Arthur Samuel (1959)**:

> **Machine Learning**: *"Field of study that gives computers the ability to learn without being explicitly programmed, relying on patterns and inference instead"*.

Approaches and considerations

- **Supervised / Unsupervised**
  - Unsupervised (*no labels*), semi-supervised (*some samples have labels*), fully-supervised (*all data labelled*).
- **Clustering / Dimensionality reduction**
- **Discrimination / Detection**
- **Representativeness of data**: all classes are well represented, some classes are not represented well (unbalanced dataset).
- **Others**: Reinforcement learning, recommender systems.

Machine Learning vs Deep Learning

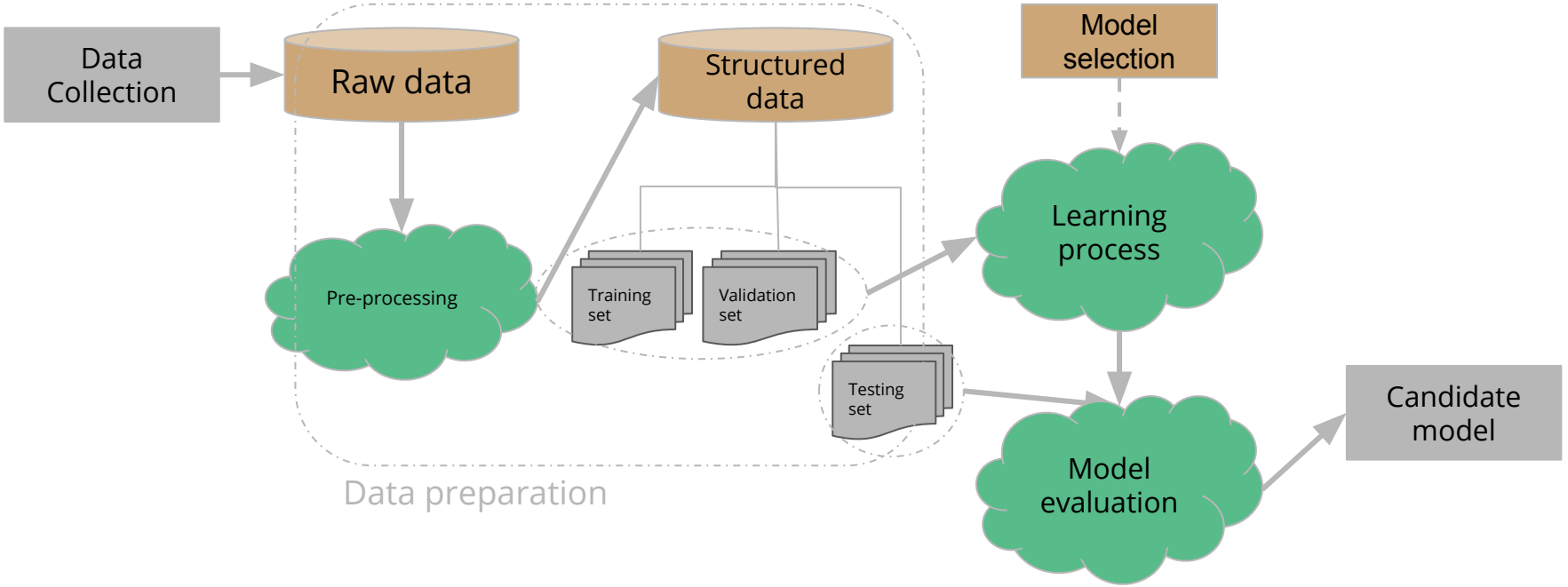
Machine Learning

- Machine Learning is a subset of Artificial Intelligence
- Uses types of automated algorithms which learn to predict future decisions and model functions using the data fed to it
- Usually, there are a **few thousand data** points used for the analysis
- The **Output** usually numerical value, like a score or a classification

Deep Learning

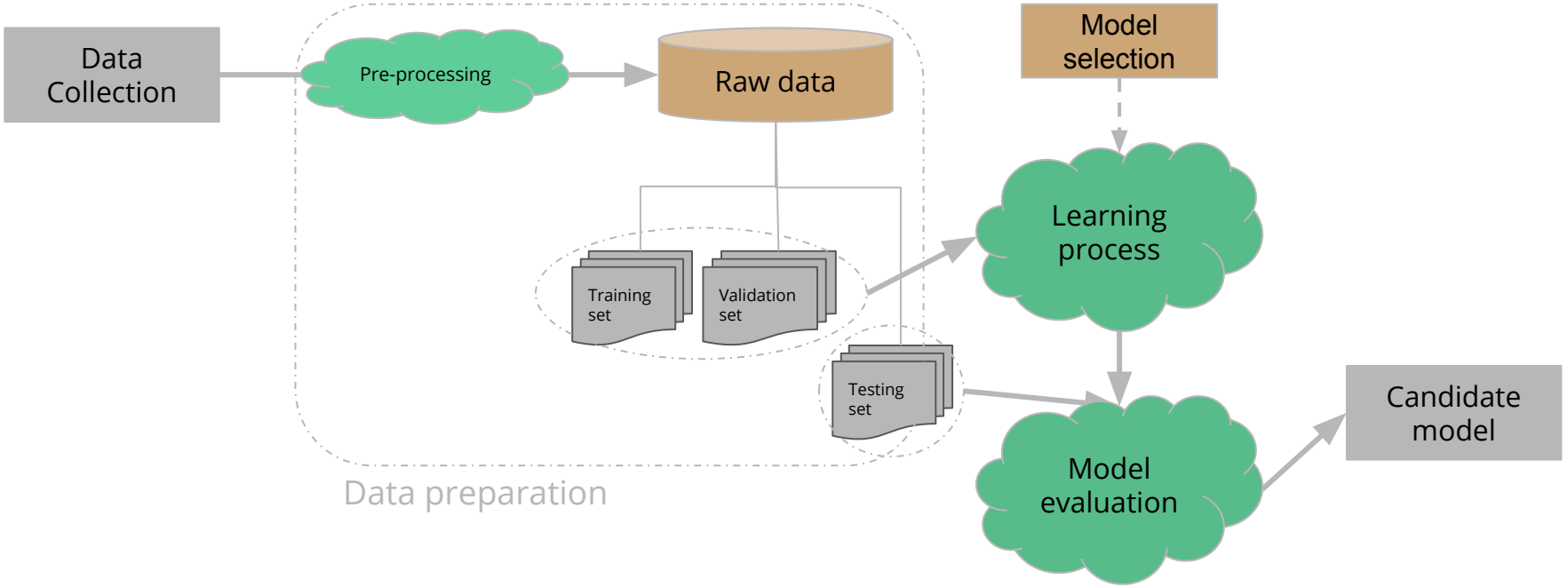
- Deep Learning is a subset of Machine Learning
- Interprets data features and its relationship using neural networks which pass the relevant information through several stages of data processing
- There are **few million data points** used for the analysis.
- The **Output** can be anything from a score, an element, free text, an image..

Machine Learning & Deep Learning process



Data collection → Data preparation → Choose a Model → Train the Model → Evaluate the model → (Candidate model)
(Candidate model) → Parameter tuning → Make predictions

Machine Learning & **Deep Learning** process



Data collection → Data preparation → Choose a Model → Train the Model → Evaluate the model → (Candidate model)
(Candidate model) → Parameter tuning → Make predictions

Considerations when choosing a model

The applicability of the techniques is, a priori, very broad, there is no method that is the panacea. Various reasons make machine learning systems very specific to the problem to be solved:

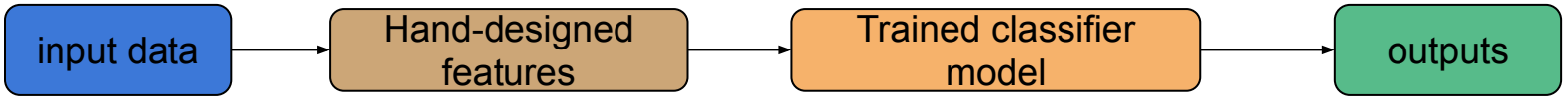
1. **The nature of the data**: characters, writings, symbols, drawings, biomedical images, three-dimensional objects, signatures, fingerprints ...
2. **The system requirements**, especially in response time. It makes some methods superior but not applicable in practice.
3. **Economic factors**: a system equipped with different sensors and very powerful processing equipment can give very satisfactory results but can not be assumed by the users.

These factors make an adequate system for one problem unacceptable for another, which allows the study and development of new techniques.

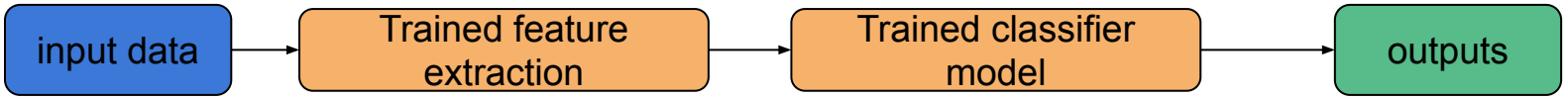
Classification of Artificial Intelligence systems

Rule-based systems (expert systems)

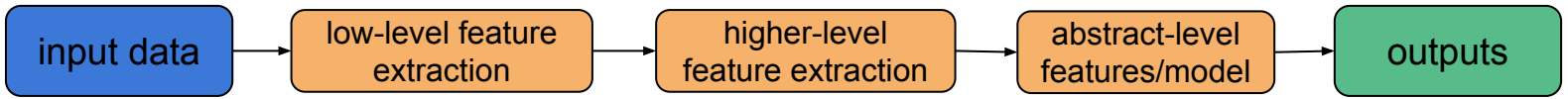
Classical machine learning



Representation learning

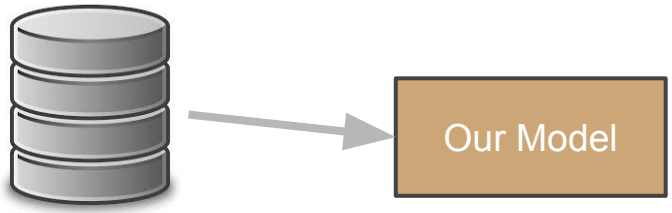


Deep learning



We are ready! Let's rock!

We have **our database** and selected **our model.** What should we do?

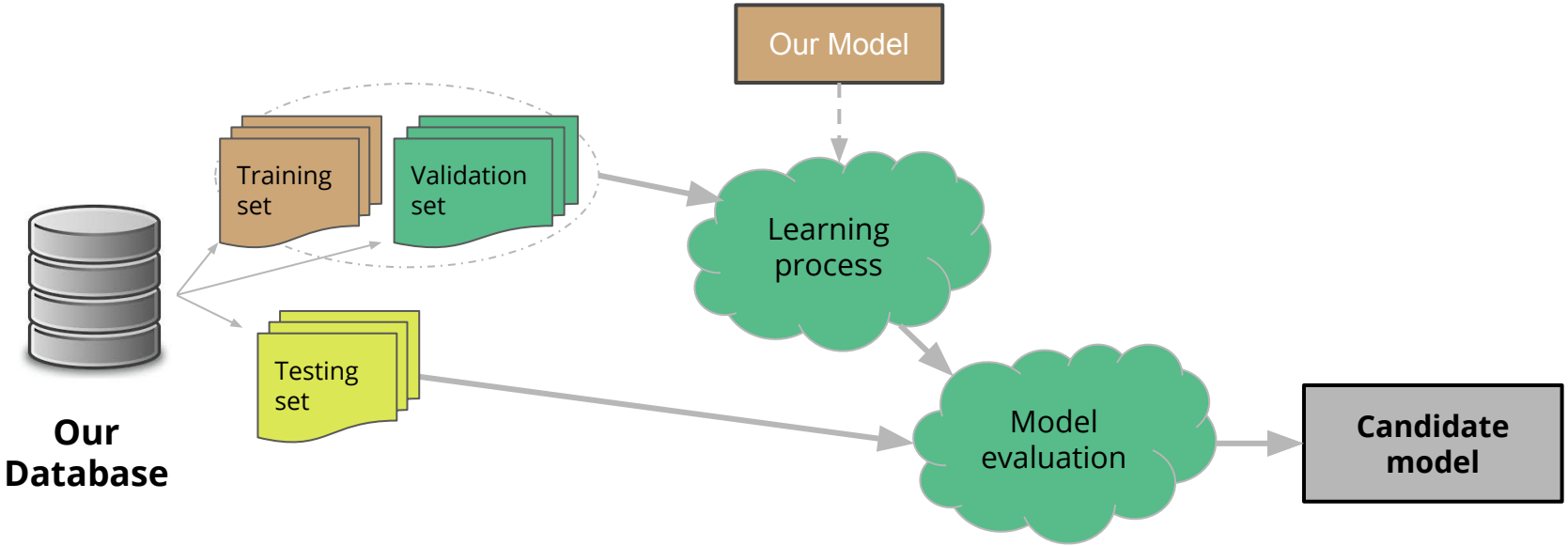


Our Database

We put everything into the model!!

We are done!

We have our **candidate model**!



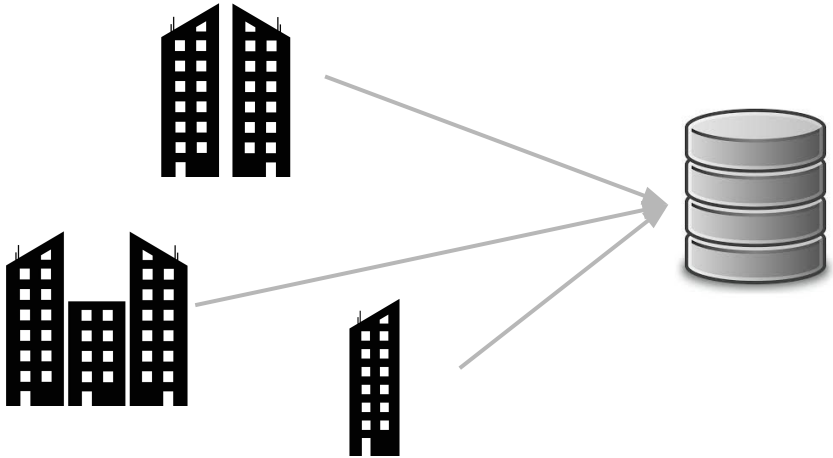

EHH. . . NO

Data Collection

We need **samples**, a lot of **samples**, usually coming from different sources.

Common errors:

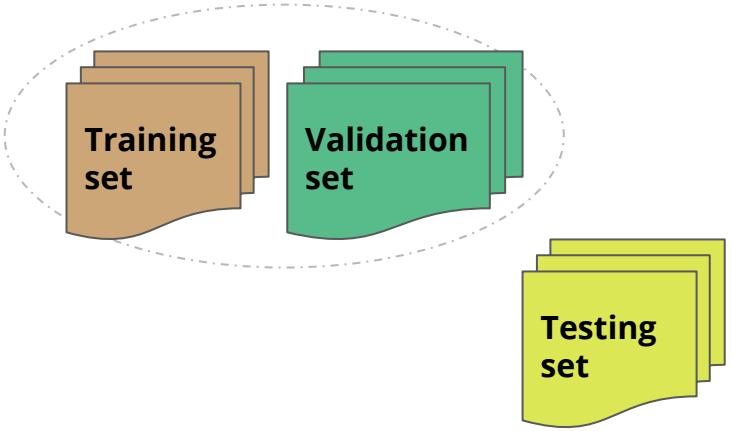
- Not check the independence of the sources → May introduce a Bias
  - Different operators/drivers
  - Different devices/instruments
  - Different acquisition software.
  - . . .
- Not cleaning the dataset
  - Outliers
  - Unit measures (*meter/mile,ºC/ºF, ...*)
  - . . .



Data Preparation: Balancing data

The set will be represented with the **different classes** to study/classify. We have to consider the **proportion of each class**.

- Balanced / Unbalanced data
  - Data augmentation
  - Collect more data
  - Changing performance metric
  - Resampling
  - Synthetic data
  - . . .



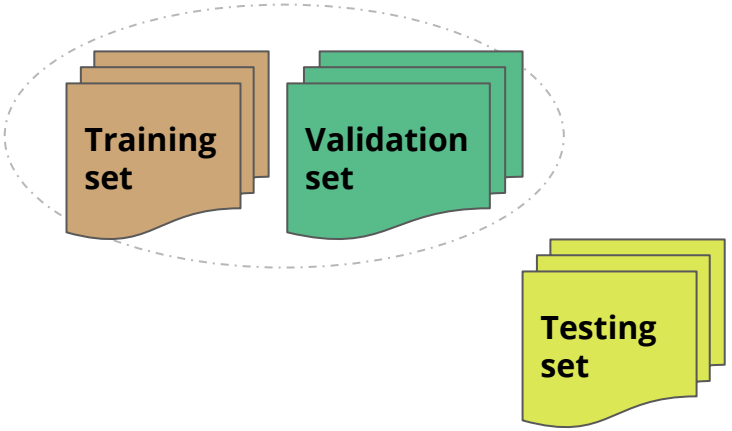
Data Preparation: Dimensionality

The larger the number of **inputs** → The larger/complex the model.

- ↑ number of hidden layers in a neural network
- ↑ bigger the size of a decision tree
- . . .

Possible solutions:

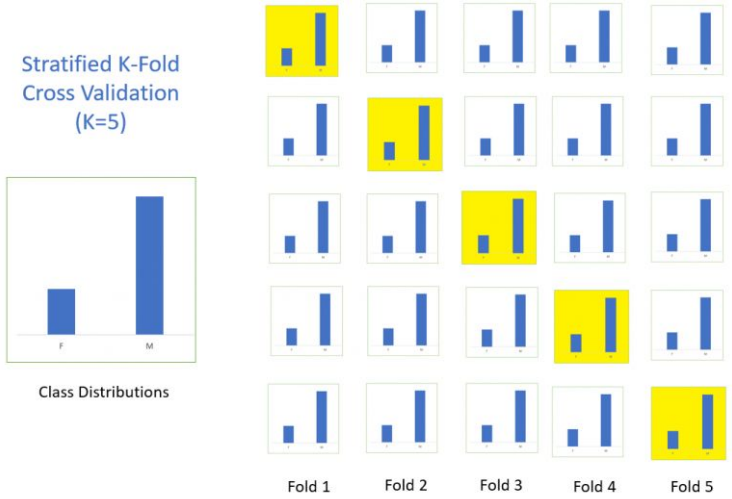
- Look for Data Correlation
- Reduce the dimensionality
- . . .



Data Preparation: Representativity

All classes have more or less the same representation in each set (*training, validation, testing set*).

- Stratified K-Fold
  - This is like cross-validation, that returns stratified folds. The folds are made by preserving the percentage of samples for each class.

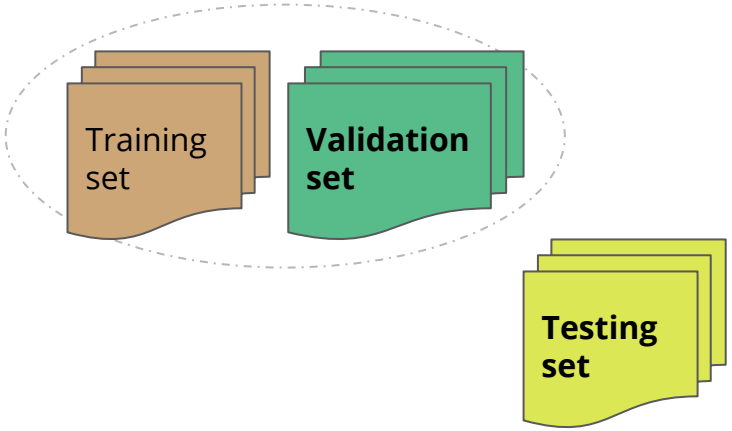


Data Preparation: Data Leakage

Data Leakage is when information from outside the training dataset is used to create the model. This affects the composition of **validation** and **testing** **sets**.

- Training set (**X**) ∩ Validation set (**X**) = ∅
- Validation set (**X**) ∩ Testing set (**X**) = ∅

  X subset of our dataset.



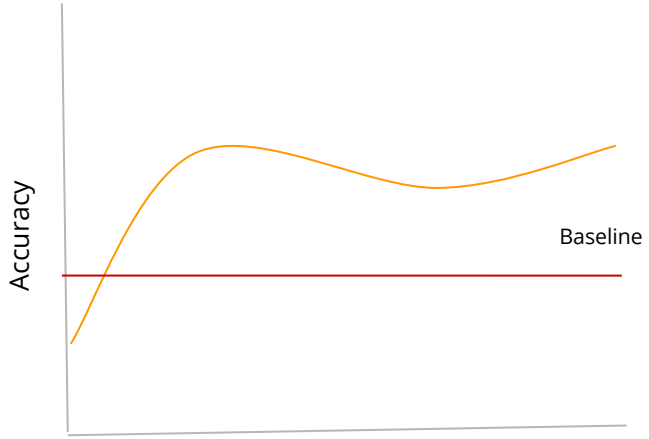
Testing: Baseline

How do we know the model is good?

We can use a **baseline** to compare our model.

Test set...

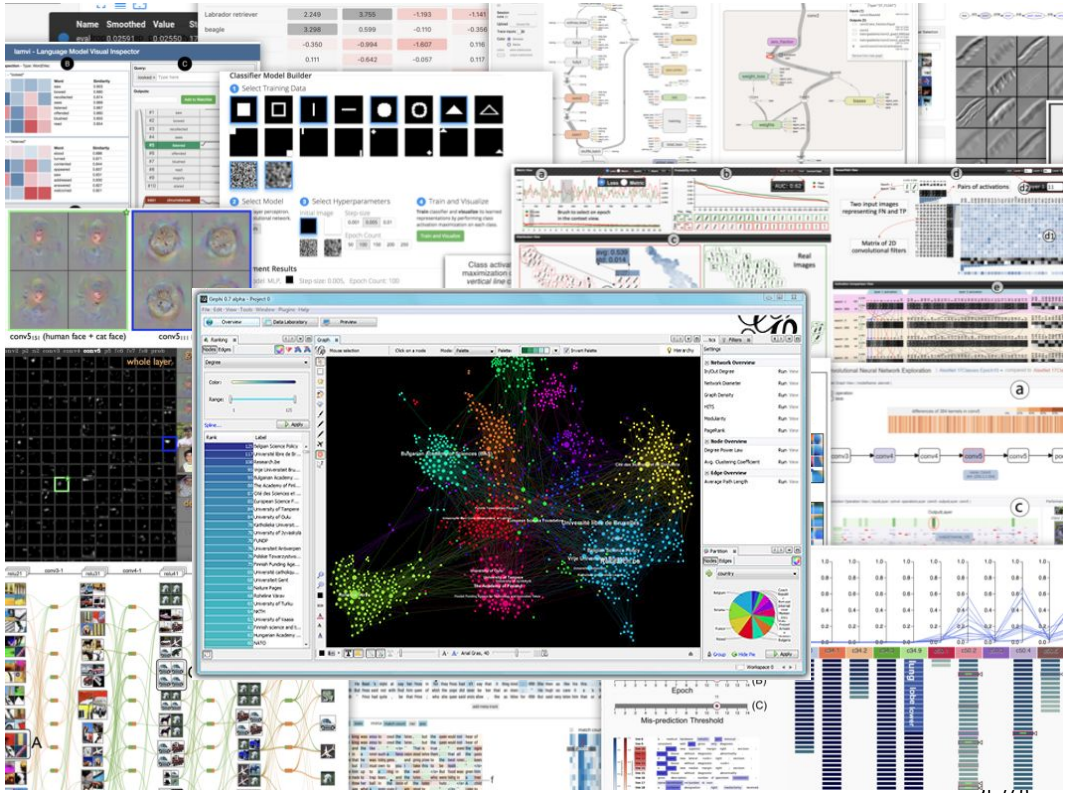
- ...using the same input
  (more abundant class)
- ...using Random input



Visualization and Interpretation

A correct visualization of the data and interpretation helps to generate better models

- Helps to understand (complex) data
- Look for correlation in the data



Source: https://gephi.org/

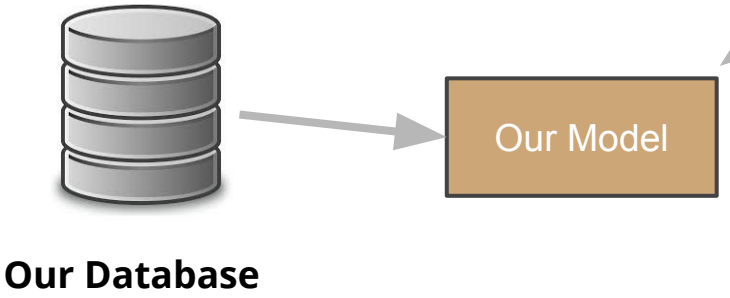
Other ways to make better models

- Apply different classifiers and takes the best
- Tuning the trained model
  - Suppress different part of the model to check their contribution
  - . . .

Restrictions of our model:

- The model needs to explain the taken decision → No Deep Learning

We are ready! Let's rock!



- Restrictions of our data
- Data correlation
- Balanced
- Check baseline
- Data Leakage
- Dimensionality

Now we are ready to do a good model!

Thanks for your attention!

Luis González Jaime

lgonzalezjaime@gmail.com

@luisgj

linkedin.com/in/luisgjaime