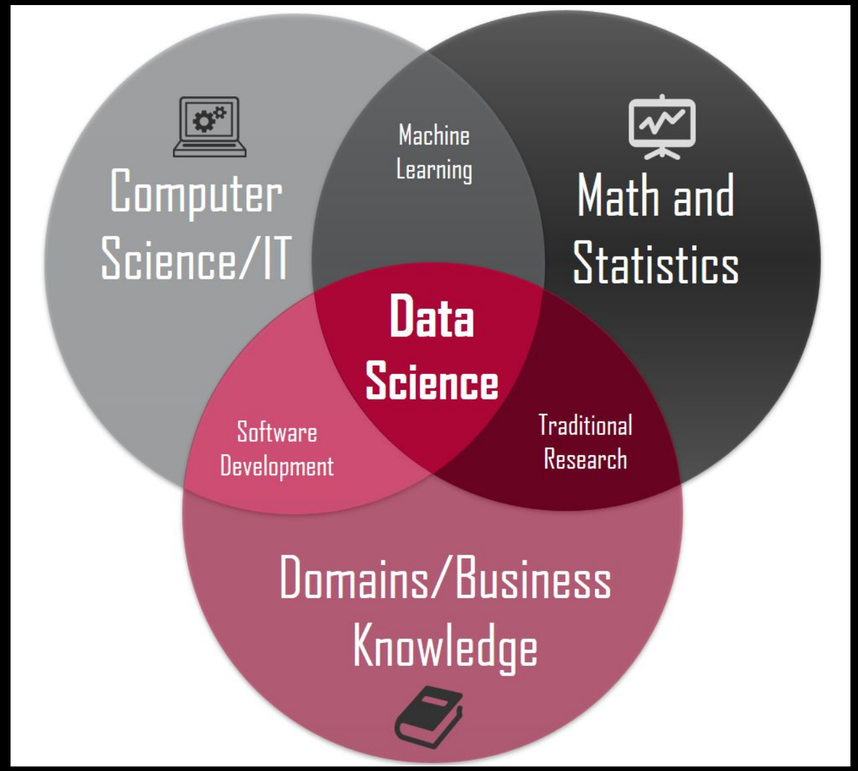
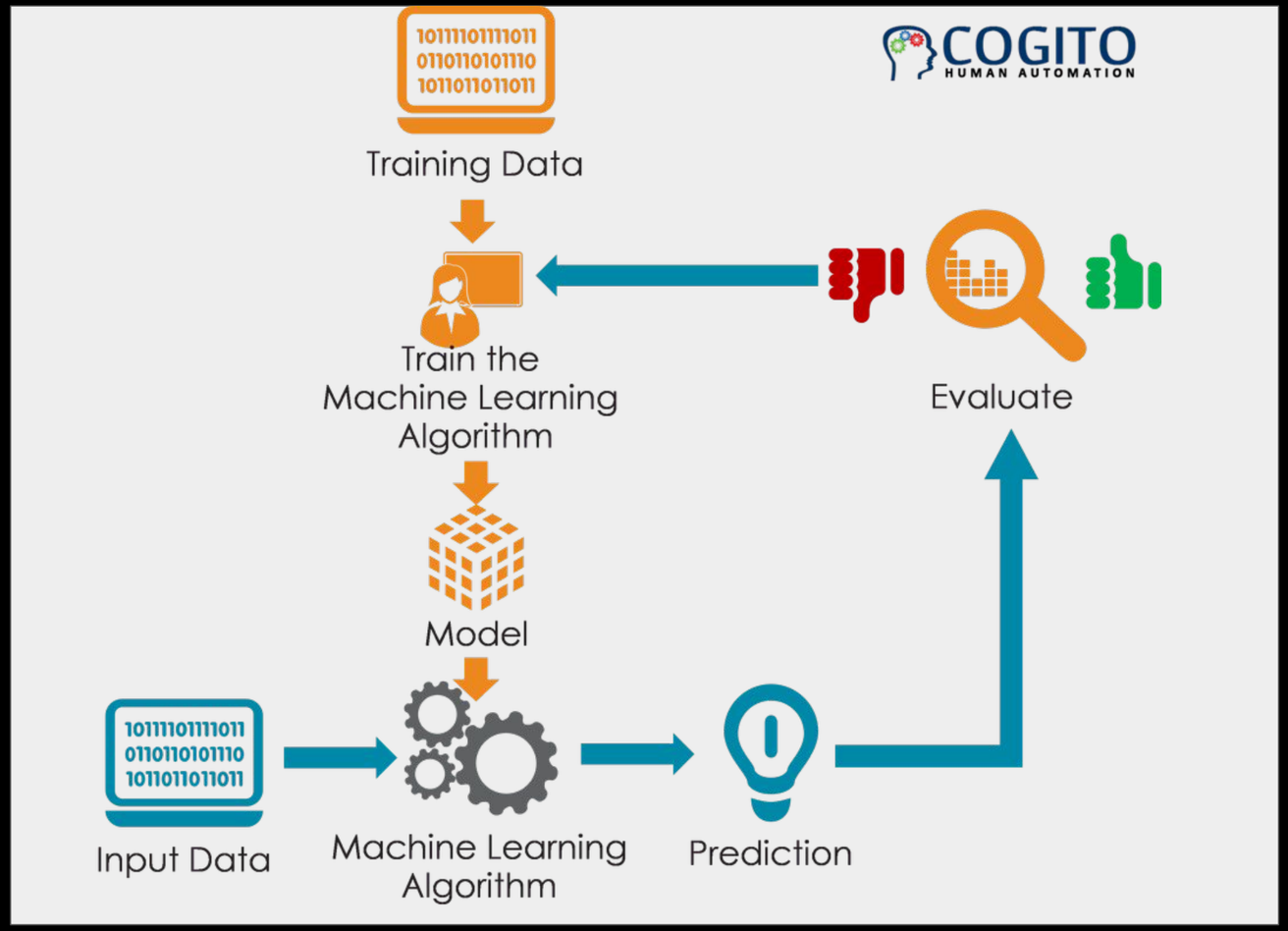
/ Intro to Machine Learning

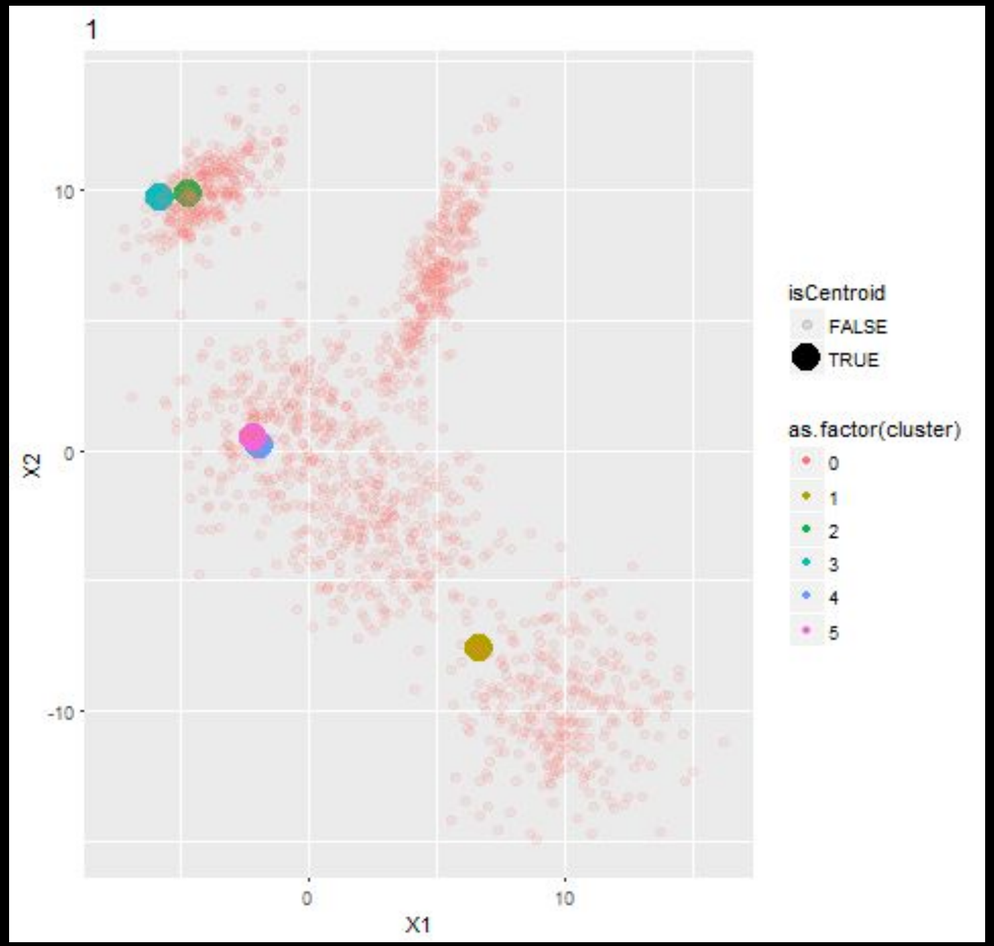
What is Data Science?

Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured.



Machine Learning process





/ Data Science Flow

How is the traditional flow of Data Science?

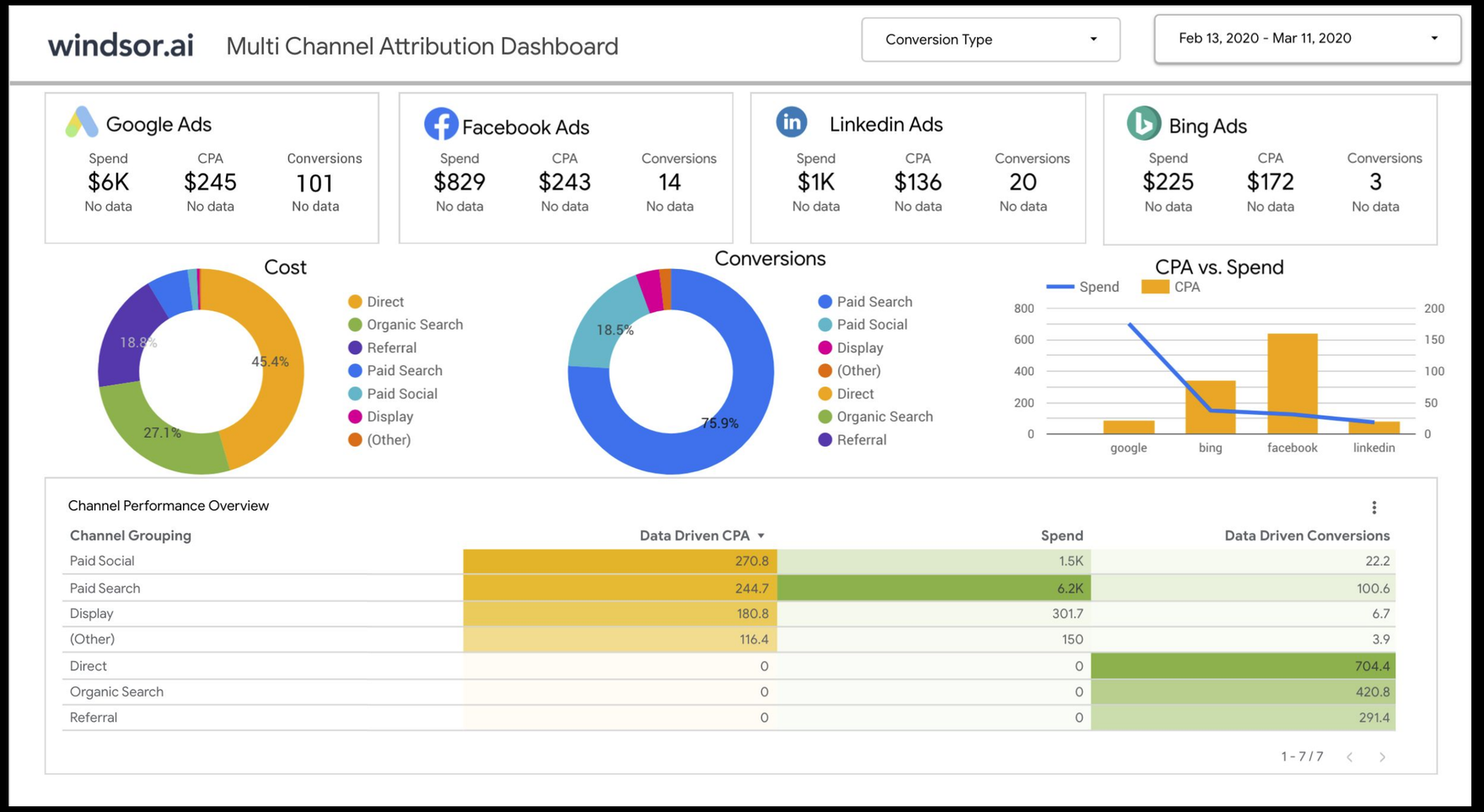
1. Business Problem
2. Data Acquisition
3. Data Preparation
.
.
.
4. Data Analysis
5. Data Modelling
6. Visualization and Communication
7. Deployment and maintenance

How to solve a DS Problem

1. Business Problem
2. Data Acquisition
3. Data Preparation
.
.
.
4. Data Analysis
5. Data Modelling
6. Visualization and Communication
7. Deployment and maintenance

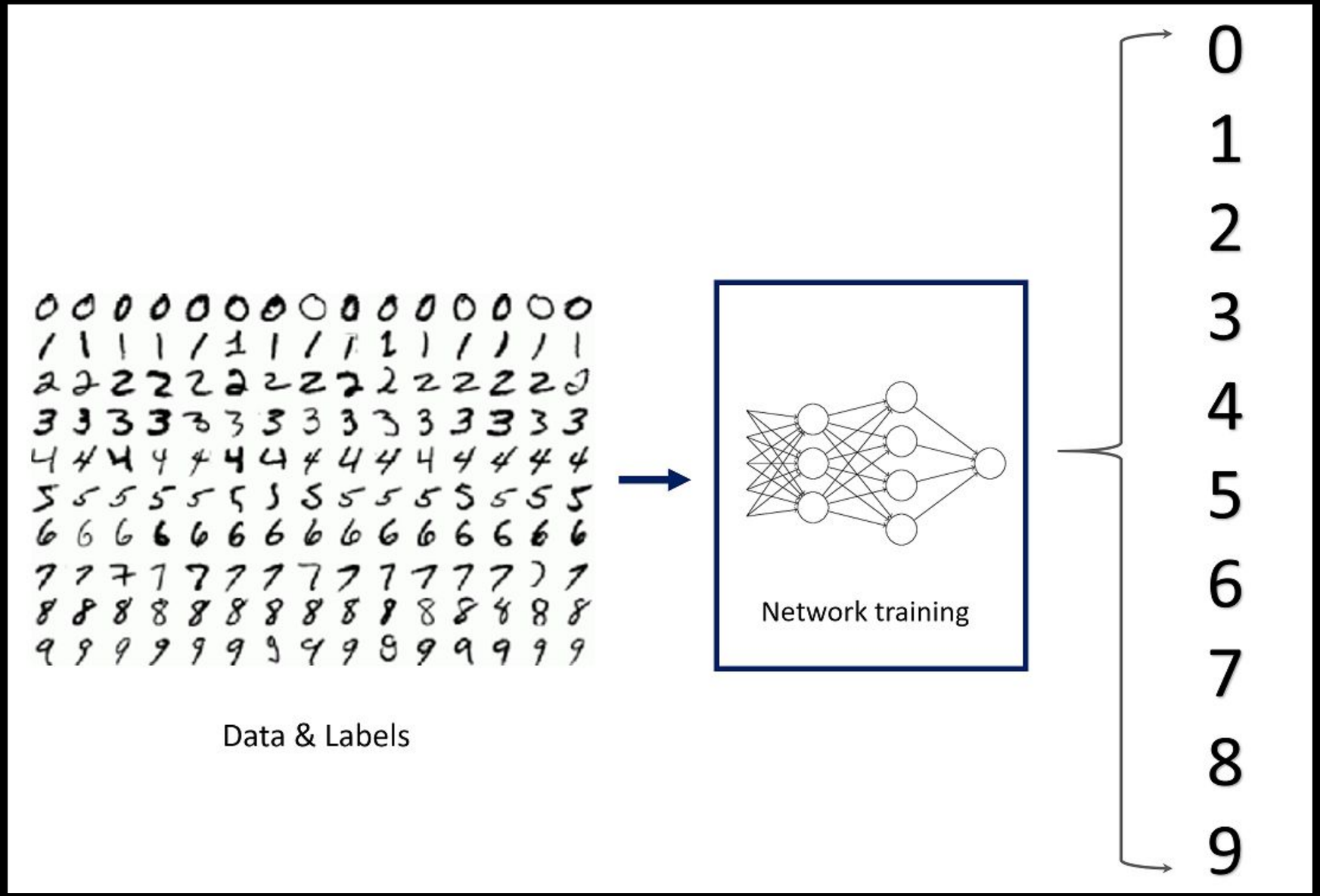
How to solve a DS Problem

1. Business Problem
2. Data Acquisition
3. Data Preparation
.
.
.
4. Data Analysis
5. Data Modelling
6. Visualization and Communication
7. Deployment and maintenance



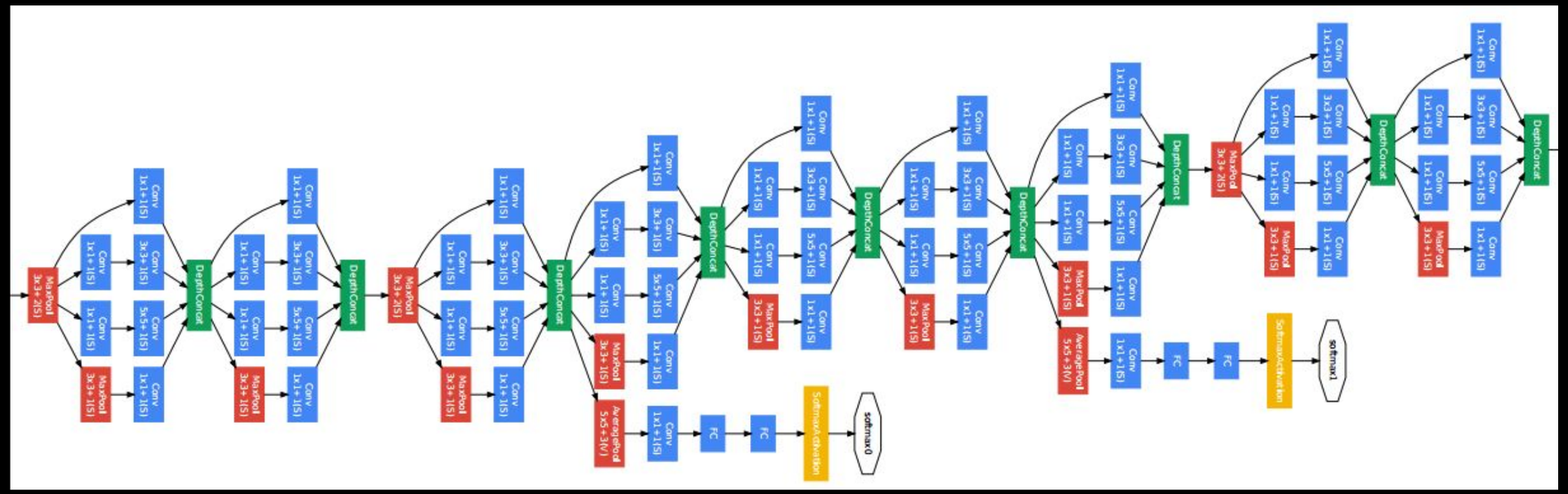
How to solve a DS Problem

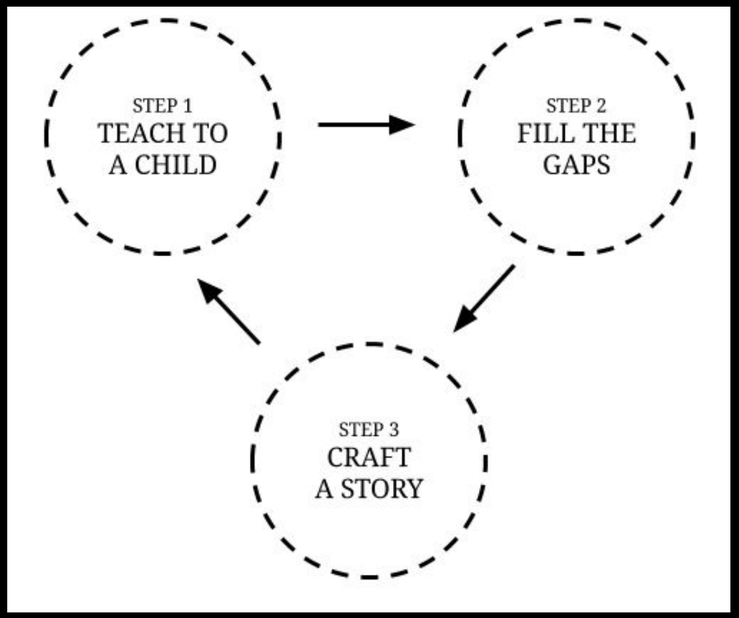
1. Business Problem
2. Data Acquisition
3. Data Preparation
.
.
.
4. Data Analysis
5. Data Modelling
6. Visualization and Communication
7. Deployment and maintenance



How to solve a DS Problem

1. Business Problem
2. Data Acquisition
3. Data Preparation
.
.
.
4. Data Analysis
5. Data Modelling
6. Visualization and Communication
7. Deployment and maintenance

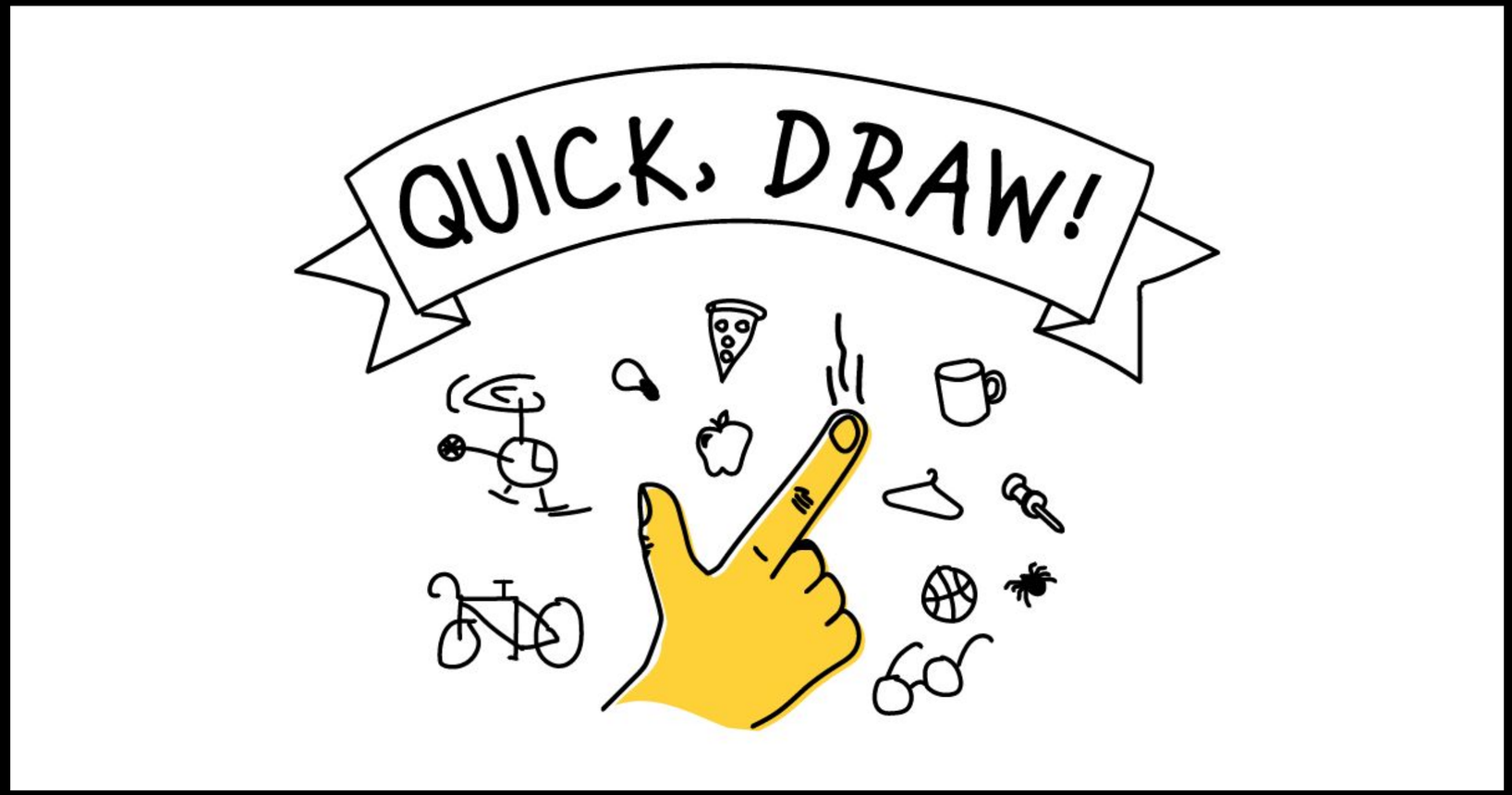




How to solve a DS Problem

Google Quick Draw

1. Business Problem
2. Data Acquisition
3. Data Preparation
.
.
.
4. Data Analysis
5. Data Modelling
6. Visualization and Communication
7. Deployment and maintenance



Machine Learning Definitions

Study of algorithms that given a task T, they improve their performance P based on experience E.

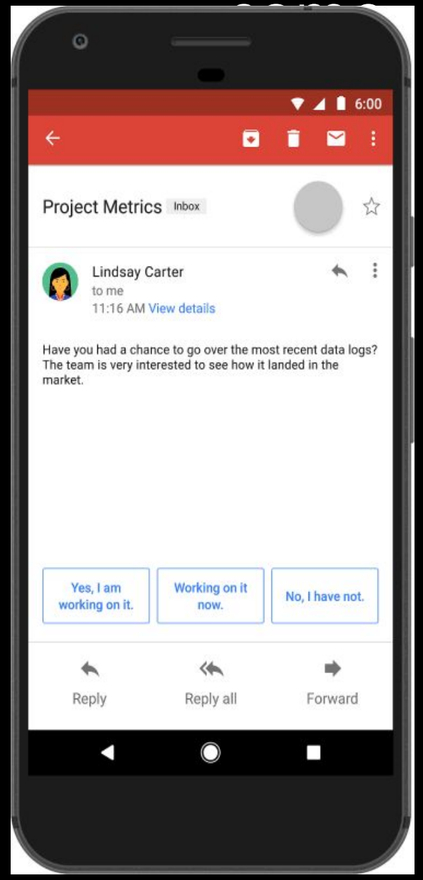
In a way, learning implies : <T,P,E>
Some examples:
T: Handwritten words recognition
P: Percentage of words identified correctly
E: Database of handwritten words

T: Driving autonomously using LIDAR
P: Average distance covered before a mistake
E: Sequence of images and direction commands recorded while a human was driving (millions of km and data actually)

Machine Learning Examples

Study of algorithms that given a **task T**, they improve their **performance P** based on **experience E**.

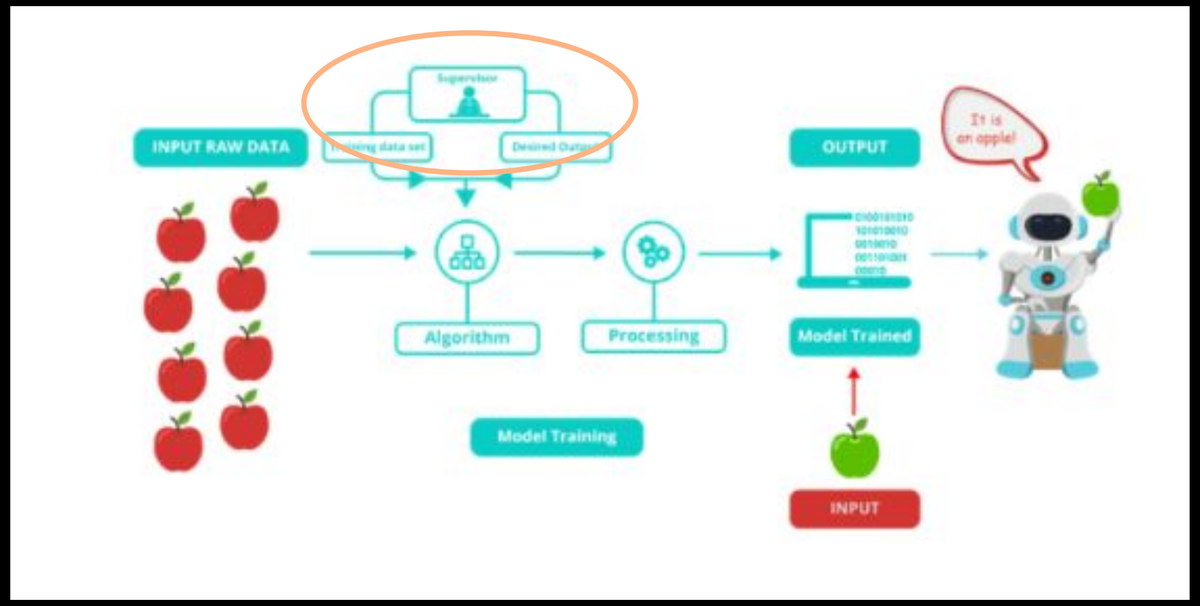
Let's think of some <T, P, E>

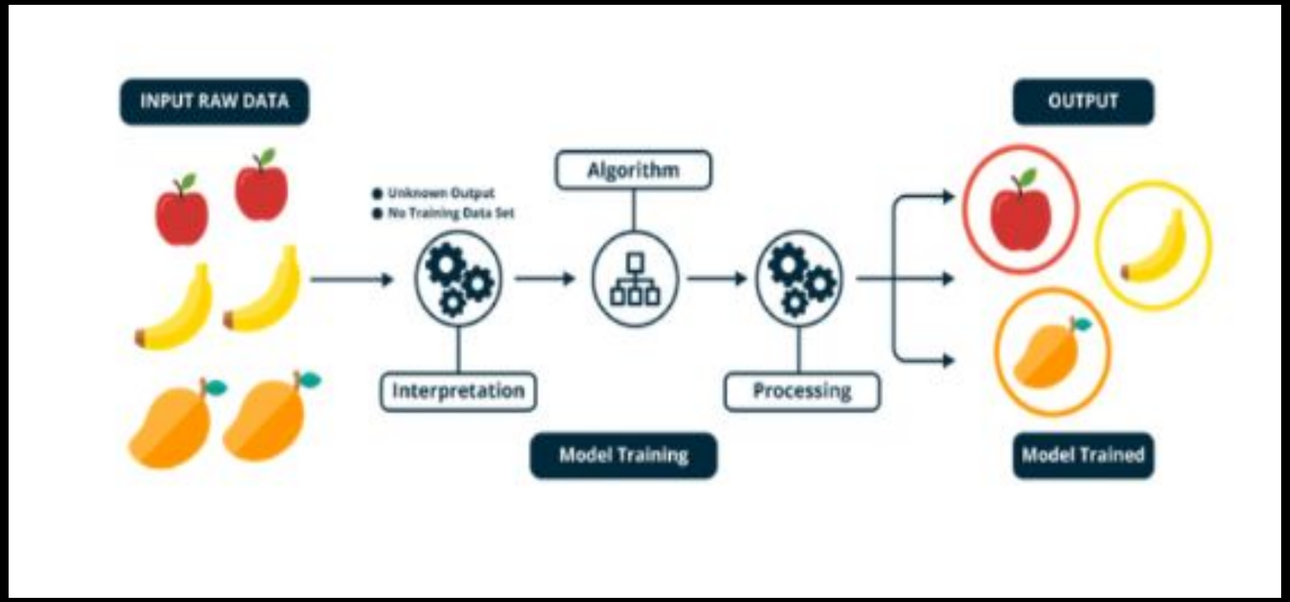


Two types of Machine Learning

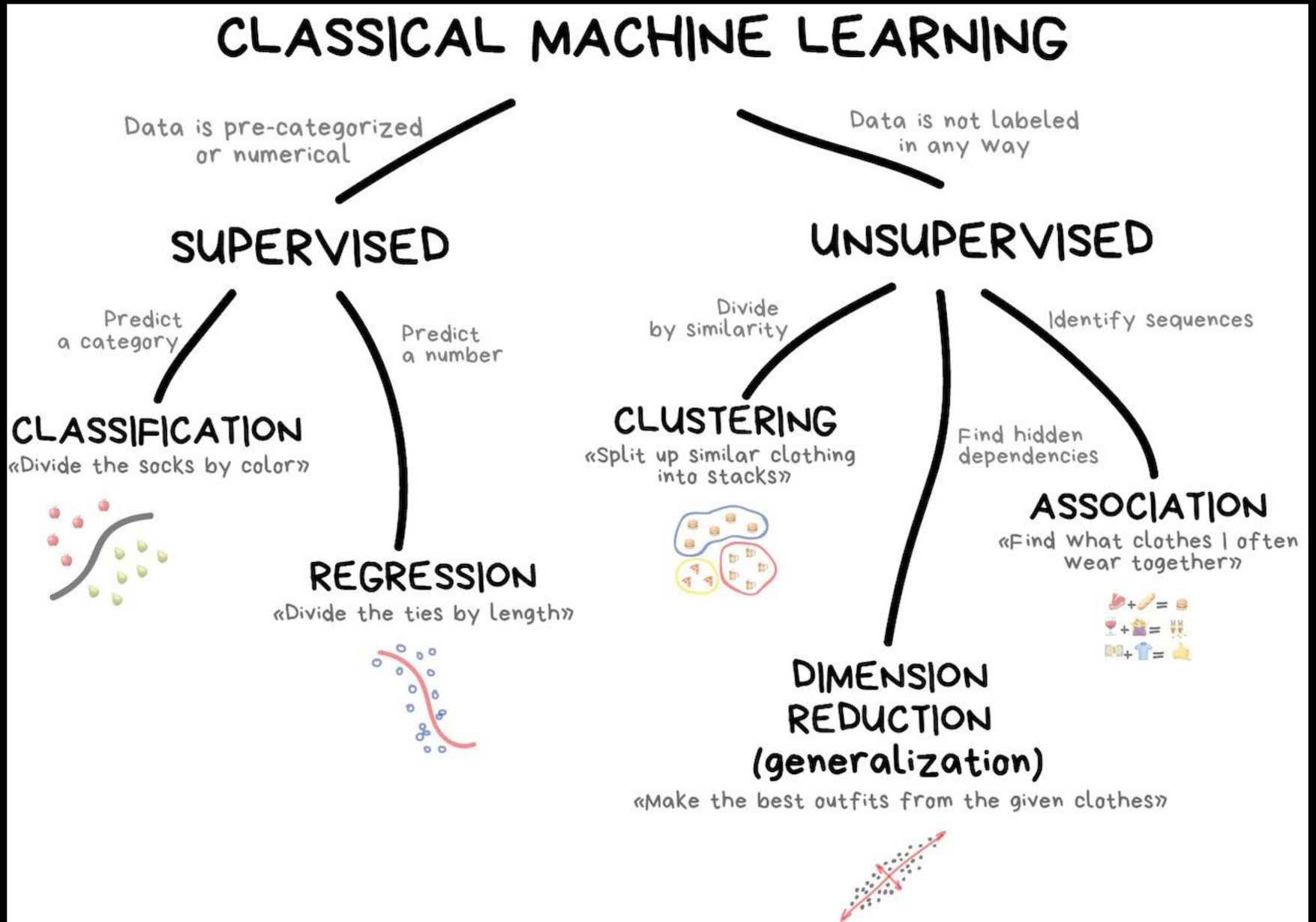
/ Supervised Learning
- > Labelled data
- > Classification
- > Regression

/ Self-supervised Learning
- > Pattern Discovery
- > Clustering
- > Anomaly detection





Fast Machine Learning **overview**



Designing a Machine Learning system

Steps:

1. Picking the data (training experience)
2. Picking what we want to learn (target function)
3. Choosing how to represent the target function
4. Picking a learning algorithm to infer the target function from the training experience.

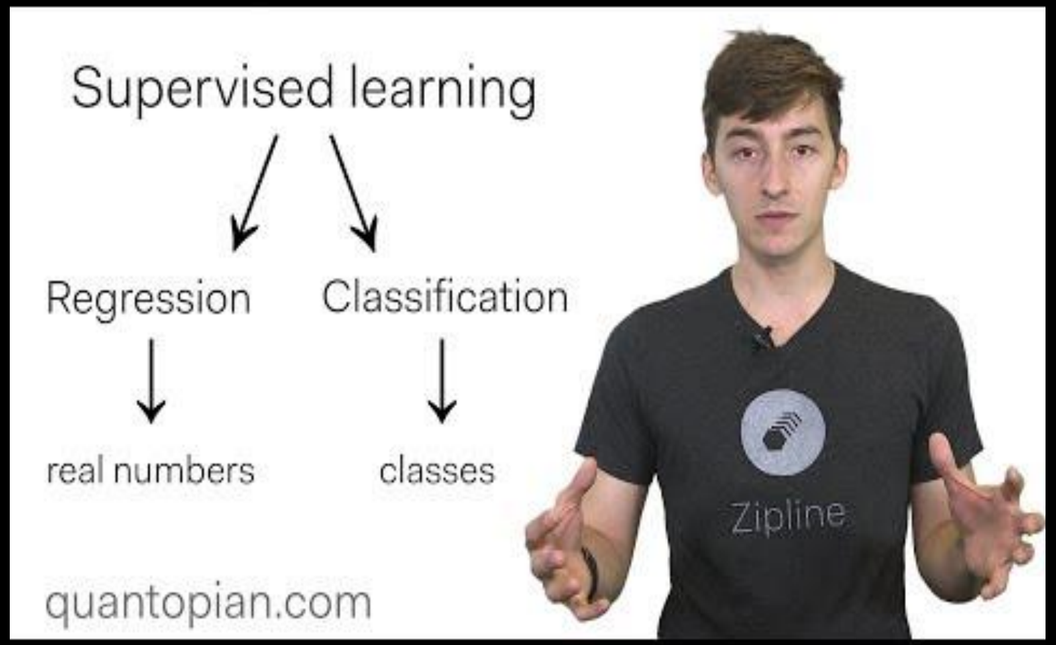
.

Maths for Machine Learning

1. Selecting the right algorithm which includes giving considerations to accuracy, training time, model complexity, number of parameters and number of features.
2. Choosing parameter settings and validation strategies.
3. Identifying underfitting and overfitting by understanding the Bias-Variance tradeoff.
4. Estimating the right confidence interval and uncertainty.

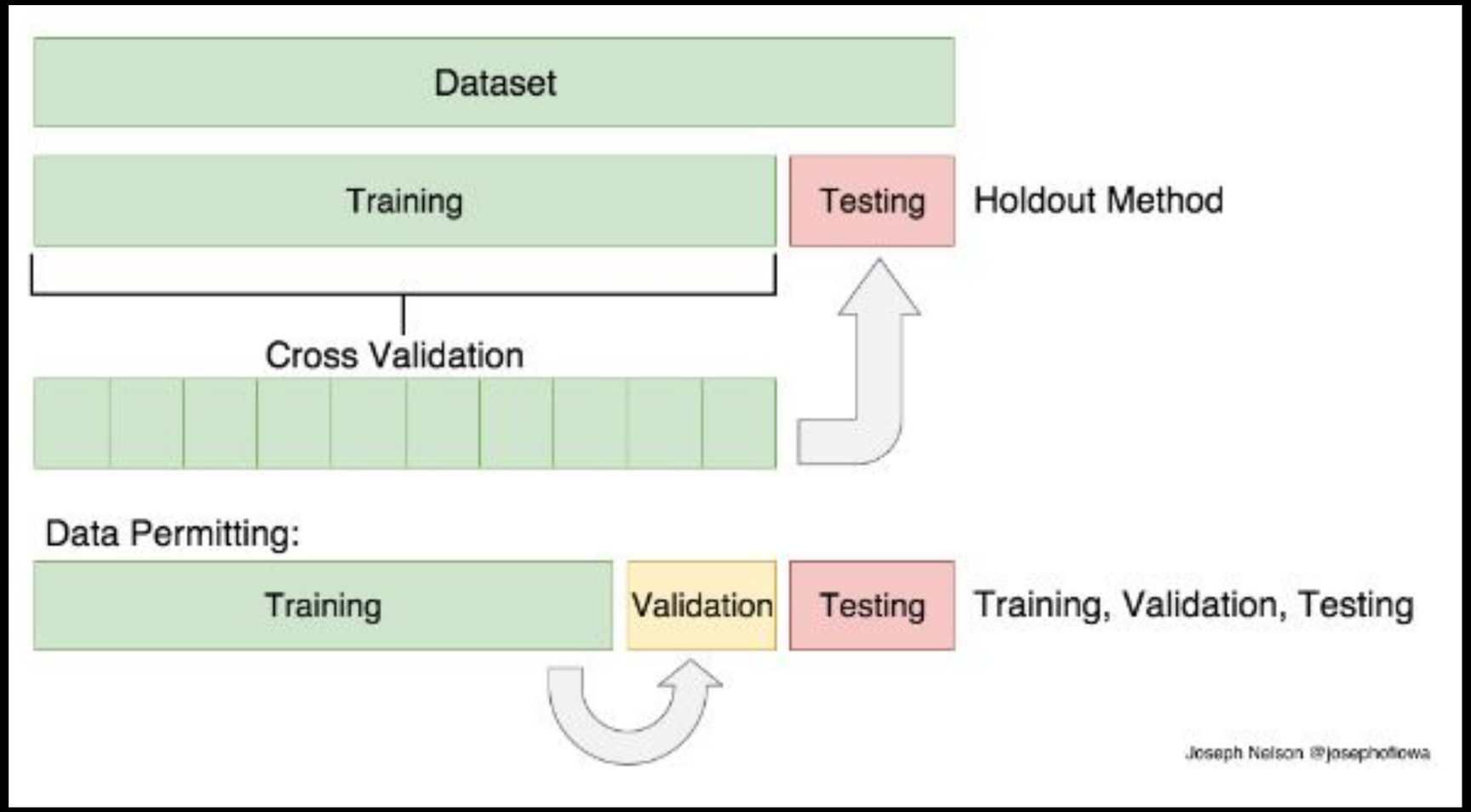
Relevant Concepts

1. What is a **task**?
   a. Classification
   b. Regression
   c. Problem solving / planning / control
2. How to evaluate **performance**?
   a. Classifying the right answers (or error)
   b. The validity of the solution
   c. Quality of the solution
   d. Performance velocity
3. How to represent experience?
   a. Neurons, cases, decision trees, etc



Train/Test Split

Like practicing for an exam with class exercises. Seeing the exam first would defeat the learning challenge

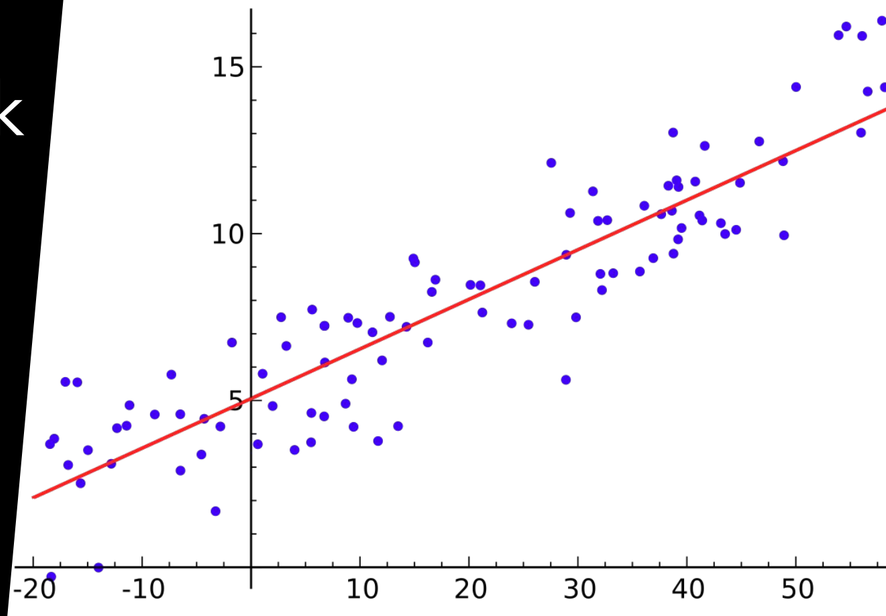


Target function

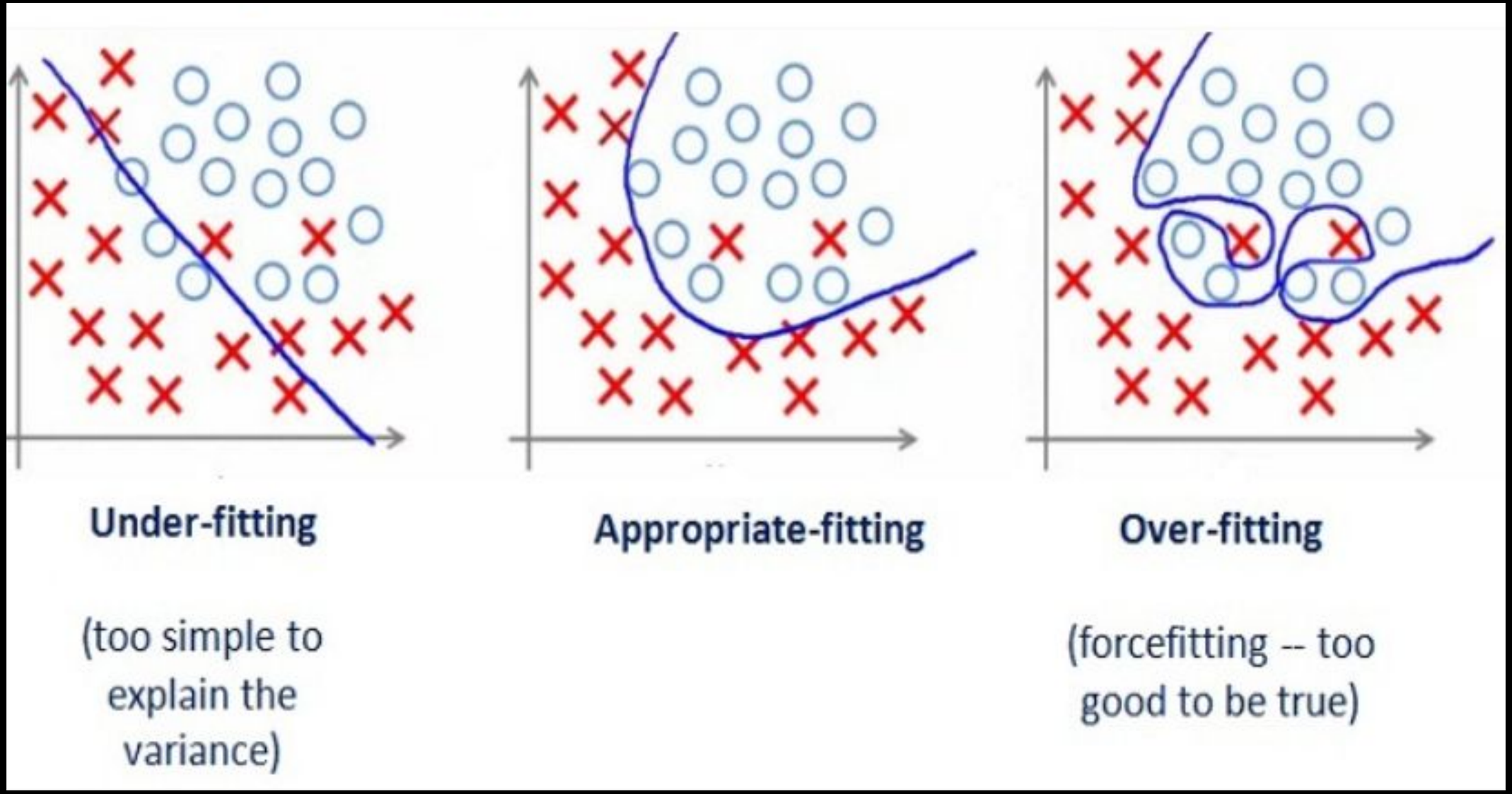
Which function must be learned and how will it be used in the system to incentivize performance?

Target function **can be represented in many ways:** lookup table, symbolic rules, numerical function -like in the picture- or neural network among others.

There is a trade-off between the **expressiveness** of a **representation** and the **ease of learning**.

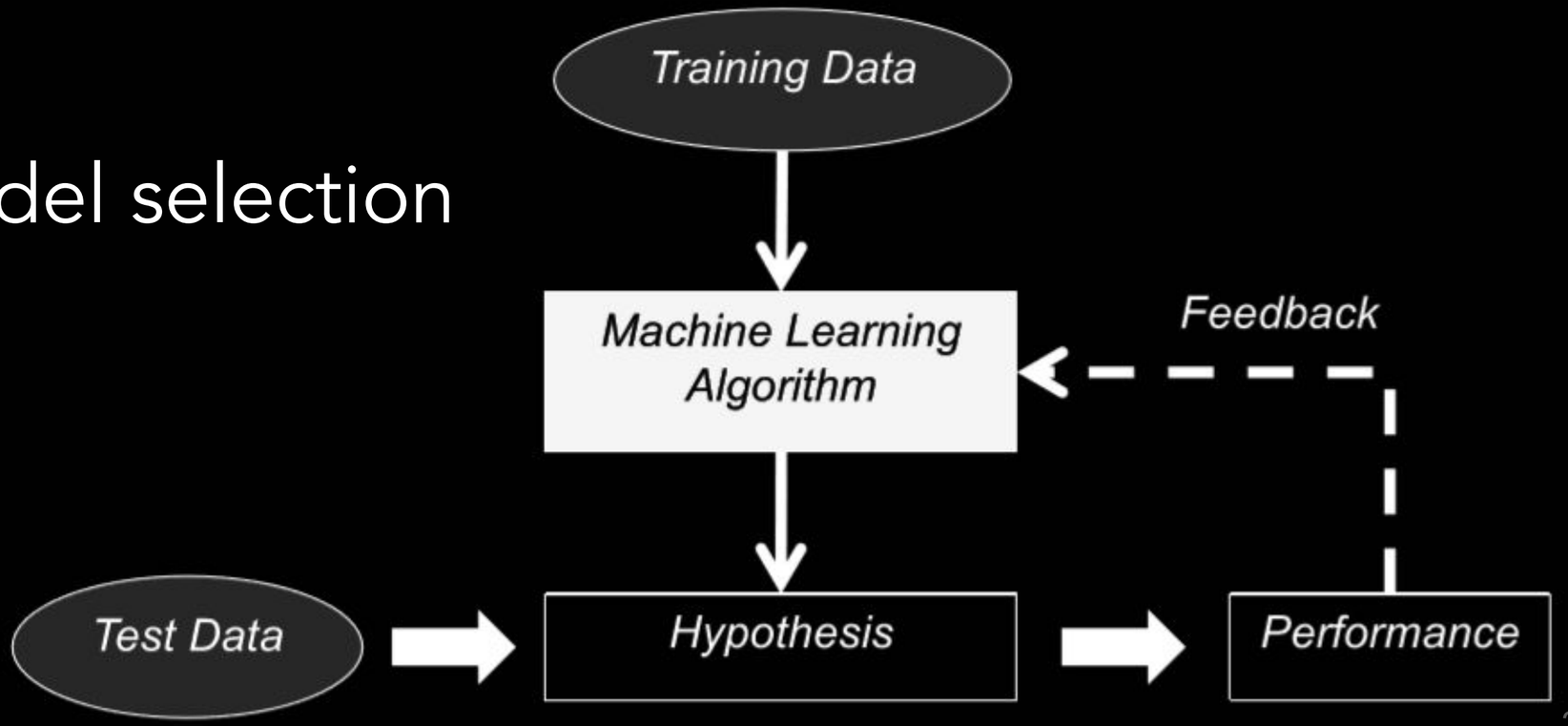


Under-fitting and Overfitting

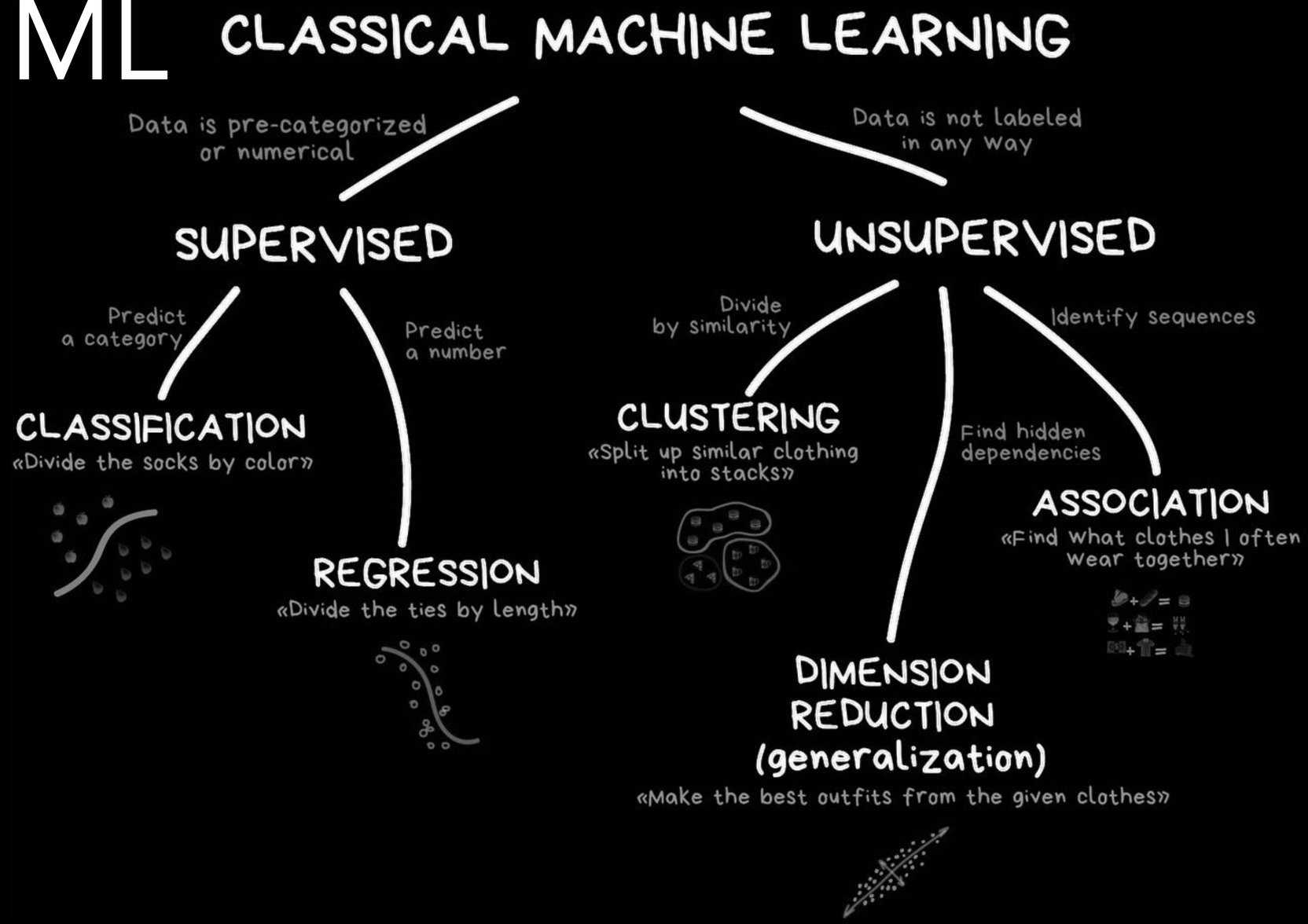


To **review** and **summarize**: ML Process

1. Data collection and Preparation
2. Feature Selection
3. Algorithm Choice
4. Parameter and model selection
5. Training Data
6. Testing Data
7. Evaluation



To **review** and **summarize:** Types of ML



/ Basic ML Dataset and Algorithm

Iris Dataset [1]

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Iris Data Set

Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	3773071

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

What is our Target?

Where do we learn from (features)?

Different types of data:

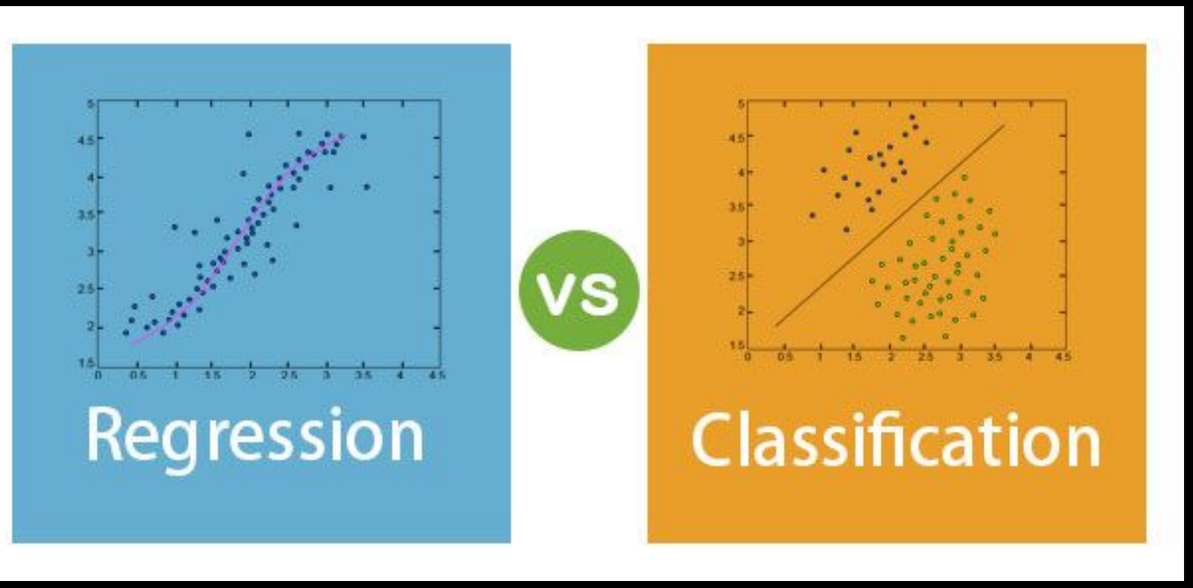
- Numerical
- Categorical (Text)
- Images / Video
- Sound → Text or Numbers

Iris Dataset [1]

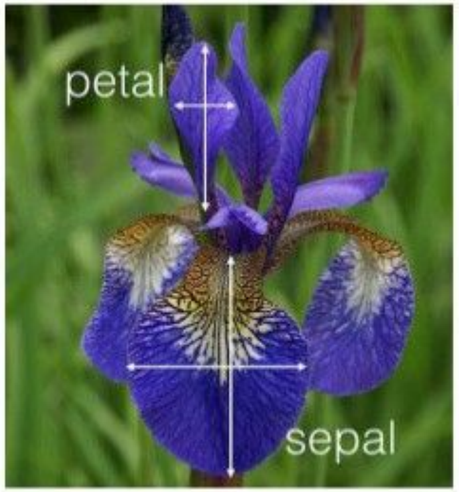
Different types of data:

- Numerical
- Categorical (Text)
- Images / Video
- Sound → Text or Numbers

Regression or Classification?



Supervised learning ??? problem
(using the Iris flower data set)



Training / test data

Features				Labels
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica

Introduction to **Sklearn**

Sklearn is the Swiss knife of machine learning, it comes with dozens of models out of the box and a huge community. It is not the most powerful knife but great to get started. There are also some tutorials to help you get started.

Sklearn comes installed with the conda environment. In other scenario we need to install it by means of pip (which we won't), to install it we just need to run:

```
conda install scikit-learn
```

Sklearn: Types of Models

Models in sklearn are imported separately as for example.

```
from sklearn.ensemble import RandomForestClassifier
```

Inside of sklearn we will find different types of models. I will just introduce the high level API of them:

/ Supervised models to perform predictions.
/ Self-supervised models to group data automatically.
/ Transformation models to perform transformations in the data

Sklearn: Supervised Models

This kind of models are the most intuitive ones. You train them with data and expected outputs and later it will predict outputs for unseen data. To train the algorithm we call the fit method and to predict with it we call the predict function

```
from sklearn.ensemble import RandomForestClassifier

clf = RandomForestClassifier().fit(X, y)
clf.predict(X)
```

Sklearn: Self-supervised Models

Other type of models, in this case it will not predict but find groups of similar elements inside data. To train the algorithm we call the fit method and to get the group of an unseen element we call the predict method

```
from sklearn.cluster import KMeans

clf = KMeans().fit(X)
clf.predict(X)
```

Sklearn: Transformation Models

Other type of models, in this case it will not predict but find groups of similar elements inside data. To train the algorithm we call the fit method and to get the group of an unseen element we call the predict method

```
from sklearn.preprocessing import MinMaxScaler

transformed_data = MinMaxScaler().fit_transform(X)
```

Sklearn: K-Fold CrossValidation

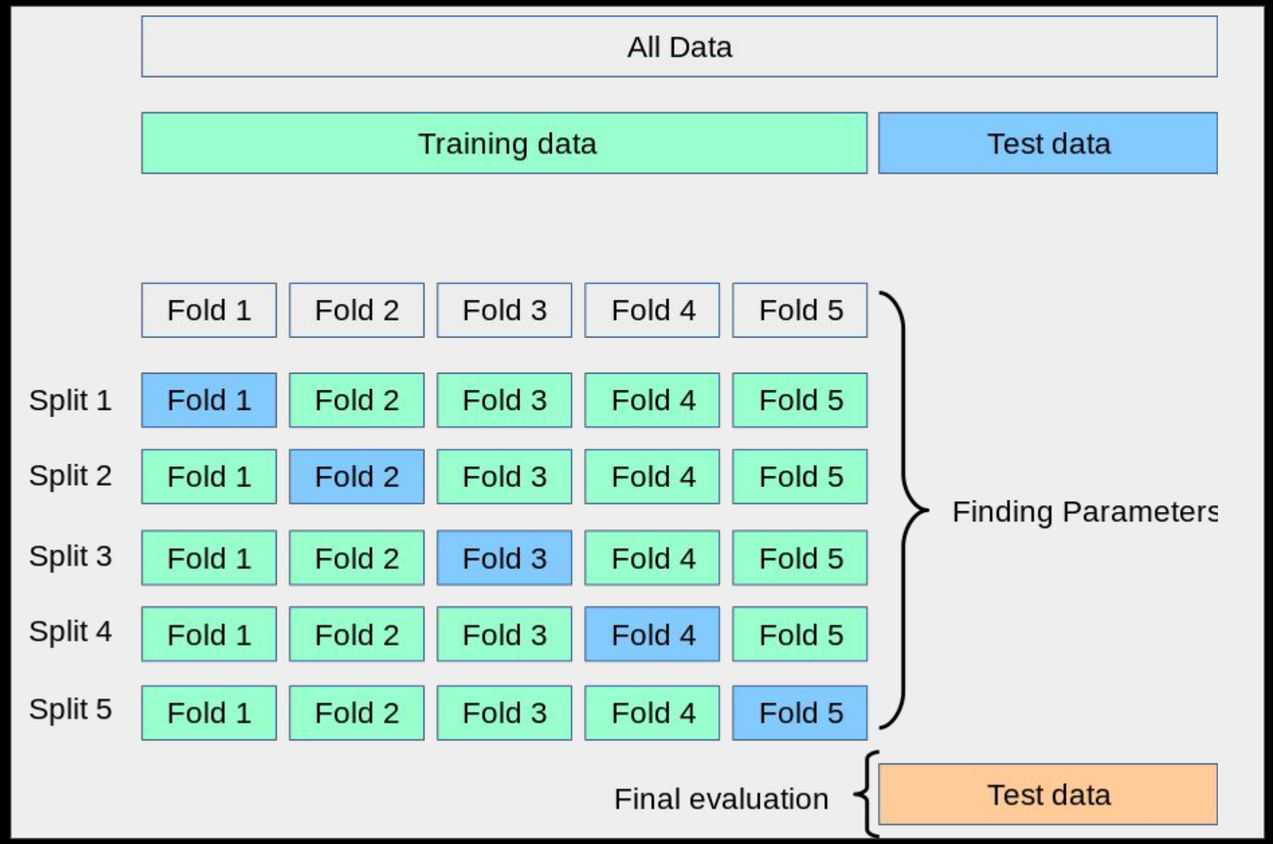
In essence we split data into two: Train and Test. However if there is enough data the ideal split should be:

- Train
- Validation
- Test

But this may not always be the case!

Sklearn: K-Fold CrossValidation

Enter K-fold CrossVal. This is a technique that allows to work on the training set yet also perform validation. The Accuracy of the model is the average of the accuracy of each fold.



Exercise **time**!