Not yet another article on Machine Learning!

Can you engage in a conversation with your boss and simply explain the basics of Machine Learning? Now you can…

Semi Koen

Photo by James Pond on Unsplash

If you have been following the latest trends in technology, you have probably noticed that Machine Learning (ML) is not just a buzzword anymore but is responsible for the most important breakthroughs in Artificial Intelligence (AI). There are lots of examples to validate the claims (from image classification to text generation to language translation), but this post is about a quick overview of ML for people that either start from zero or those that are after a concise summary.

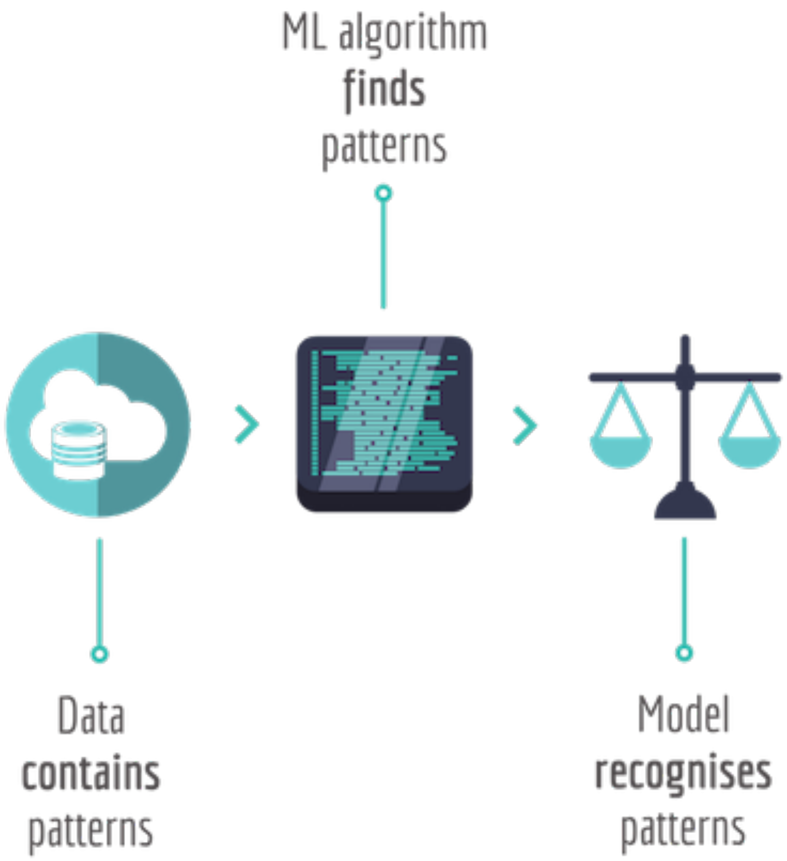
It is not just another introductory article on ML… It is ***THE*** introductory article on ML! 🏆

ML — What is it?

ML enables computers to find **patterns** in data and then use those to make decisions rather than being explicitly programmed to carry out a certain task.

The workflow is pretty simple:

- You have data which ***contains*** patterns.
- You supply it to a ML algorithm which ***finds*** the patterns and generates a model.
- The model ***recognises*** these patterns when presented with new data.



ML Workflow

Every day examples include:

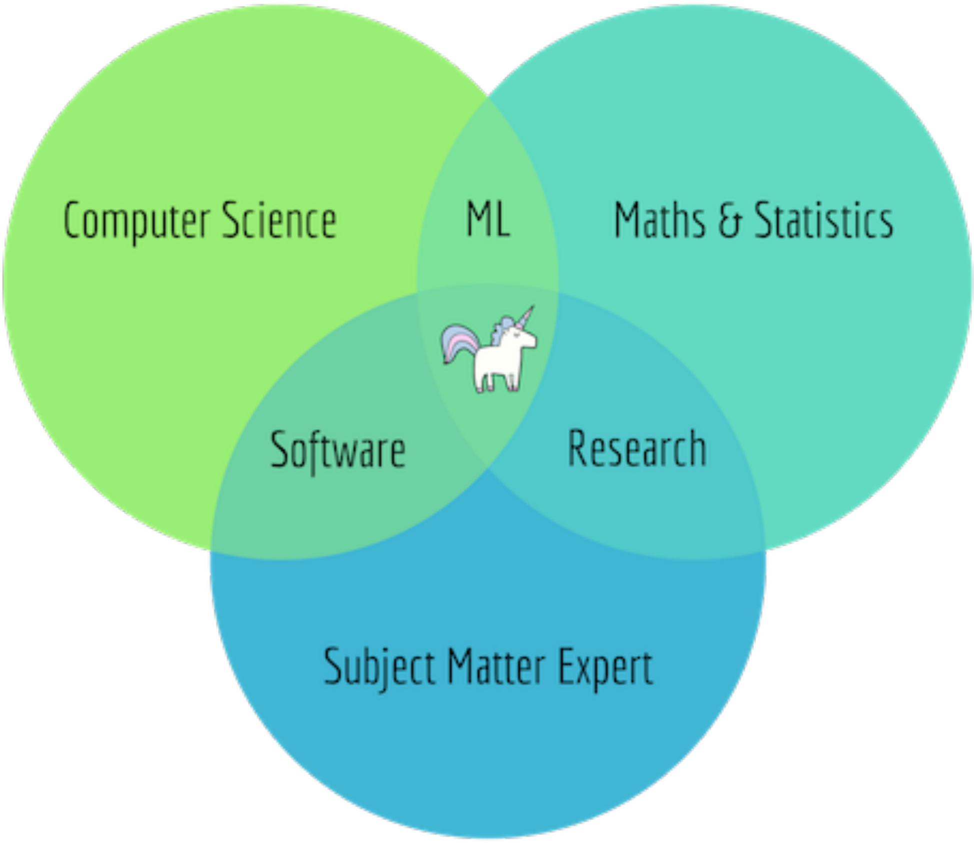
> *Medical diagnosis*
> *Customer's ability to pay back a loan*
> *Market analysis / Stock trading*
> *Credit card fraud detection*
> *Customer segmentation*
> *Spam emails*

Who is a Data Scientist?

The HBR article from 2012 was prophetic…

> *Data Scientist is 'the **sexiest** job of the 21st century'.*

Fast forward to 2019, a Data Scientist is someone with multidisciplinary skills ranging from mathematics, statistics, machine learning, computer science, programming and a business domain expertise. Rightly Steven Geringer calls them ***Unicorns*** as in the 'mythical beast with magical powers who's rumoured to exist but is never actually seen in the wild' 😄 — Oh yes they do!.



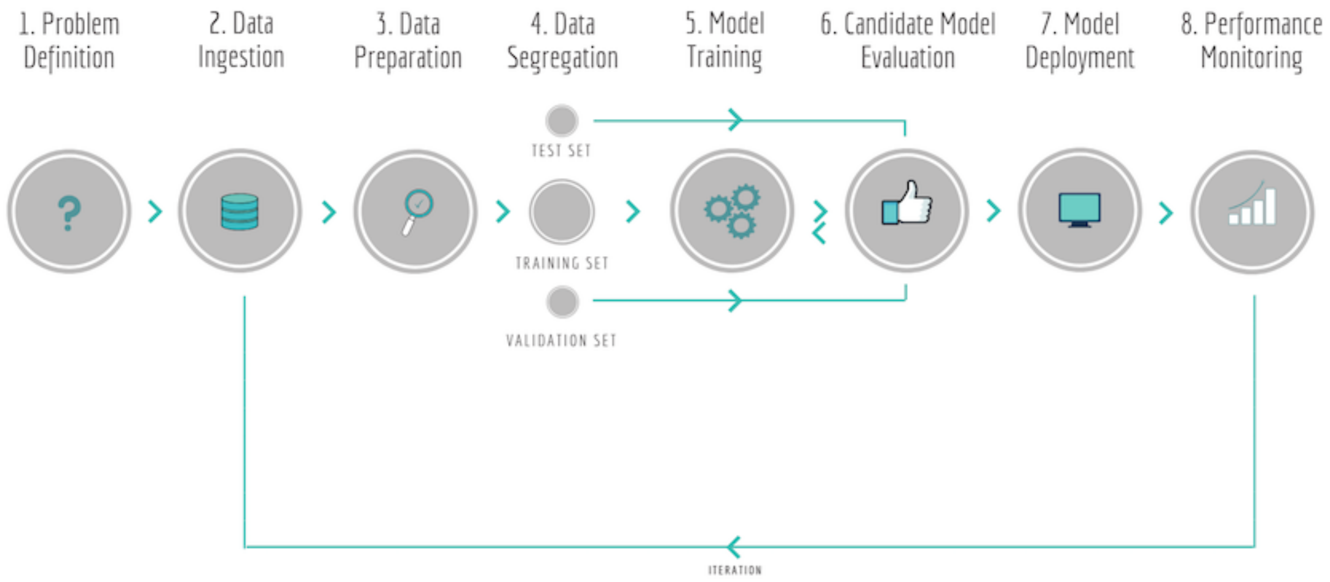
Venn diagram of a data science unicorn [Copyright Steven Geringer*]*

ML Pipeline

Data scientists define a pipeline for data as it flows through their ML solution. Each step of the pipeline is fed data processed from its preceding step. The term 'pipeline' is slightly misleading as it implies a one-way flow of data; instead the ML pipelines are cyclical and iterative as every step is repeated to finally achieve a successful algorithm.
The key stages are described below:

1. **Problem Definition**: Define the business problem you require an answer for.
2. **Data Ingestion**: Identify and gather the data you want to work with.
3. **Data Preparation**: Since the data is raw and unstructured, it is rarely in the correct form to be processed. It usually involves filling missing values or removing duplicate records or normalising and correcting other flaws in data, like different representations of the same values in a column for instance. This is where the feature extraction, construction and selection takes place too.
4. **Data Segregation**: Split subsets of data to *train* the model, *test* it and further *validate* how it performs against new data.
5. **Model Training**: Use the training subset of data to let the ML algorithm recognise the patterns in it.
6. **Candidate Model Evaluation**: Assess the performance of the model using test and validation subsets of data to understand how accurate the prediction is. This is an iterative process and various algorithms might be tested until you have a Model that sufficiently answers your question.
7. **Model Deployment**: Once the chosen model is produced, it is typically exposed via some kind of API and embedded in decision-making frameworks as a part of an analytics solution.
8. **Performance Monitoring**: The model is continuously monitored to observe how it behaved in the real world and calibrated accordingly. New data is collected to incrementally improve it.

And putting it all together in a diagram:



ML Pipeline

Training Algorithms Taxonomy

ML algorithms are divided into two broader categories:

Supervised Learning (SL)

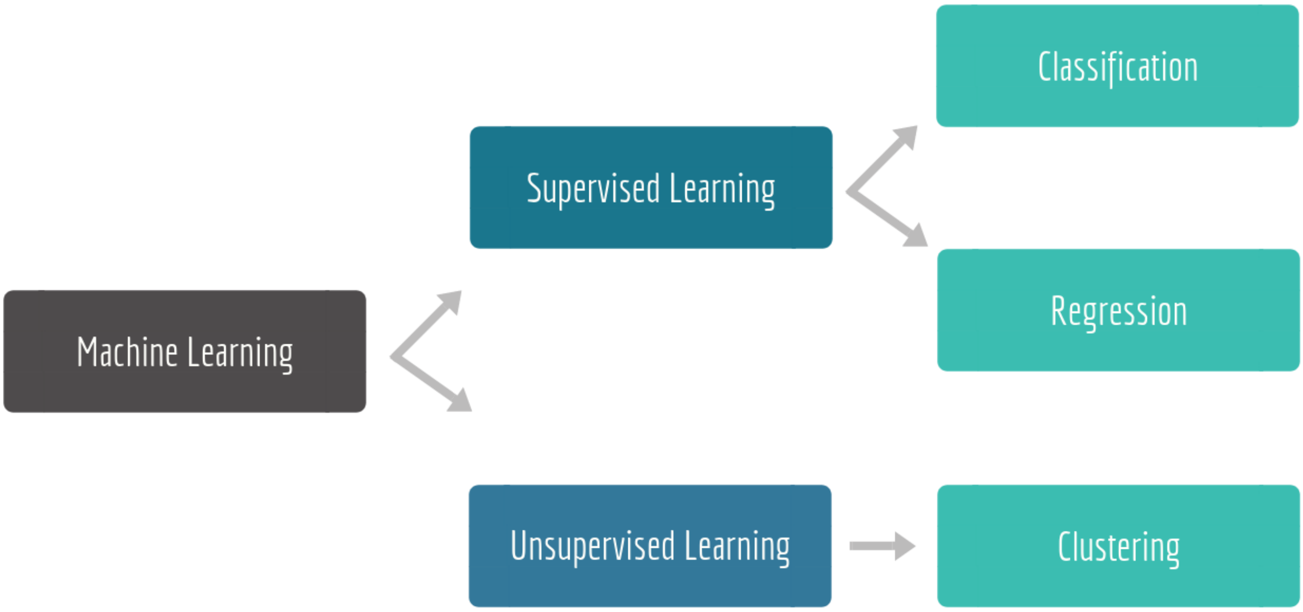
The value you want to predict ***is*** in the training data, so the algorithm can **predict** future outputs in a reasonable manner.

> *Data is **labelled***

Unsupervised Learning (UL)

The value you want to predict ***is not*** in the training data, so the algorithm finds **hidden patterns** (according to similarities or differences) or intrinsic values.

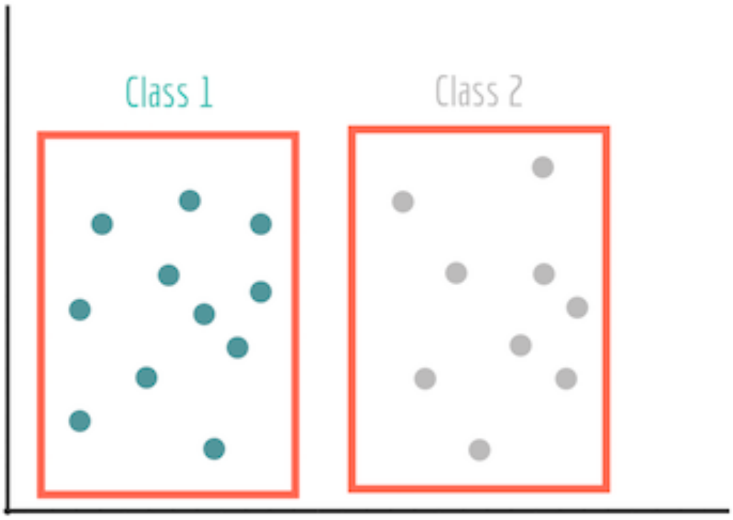
> *Data is **unlabelled***



ML Taxonomy

The main subcategories are:

Classification

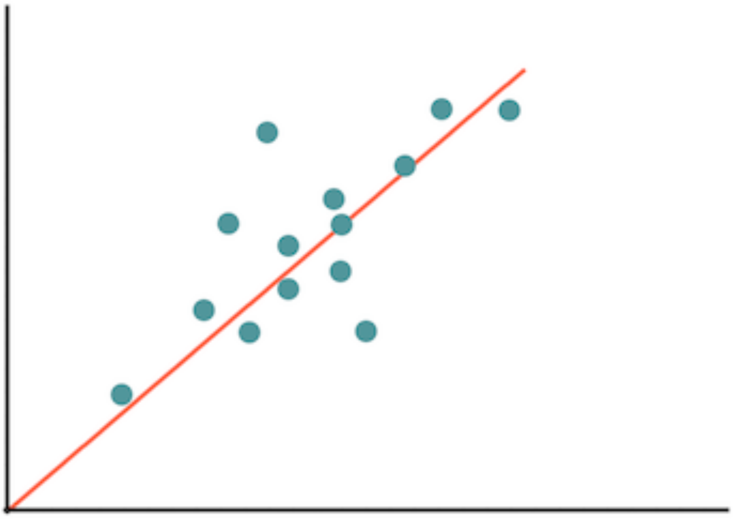


Supervised Learning — Classification

A subcategory of SL, Classification is the process of predicting **categorical/discrete responses** i.e. the input data is classified into categories. Another application is anomaly detection i.e. the identification of outliers/unusual objects that do not appear in a normal distribution. Examples:

> *Yes / No*
> *Genuine / Spam email*
> *Male / Female*
> *Fraudulent / Legit transaction*

Regression

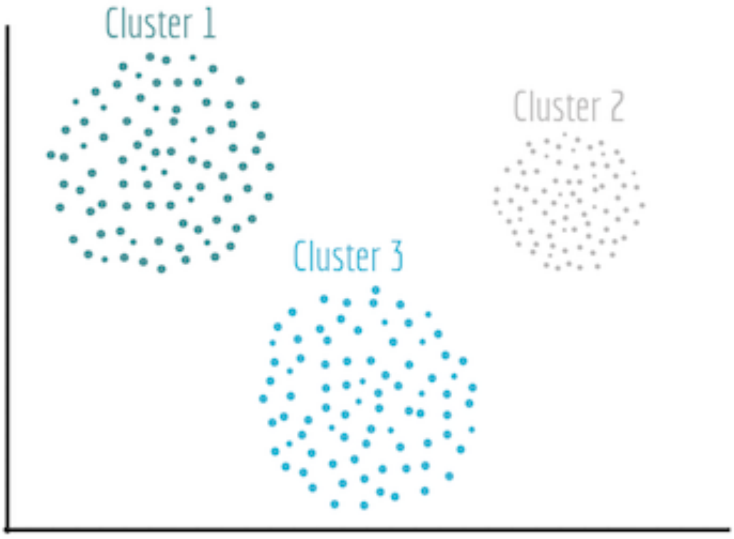


Supervised Learning — Regression

Another subcategory of SL, Regression is the process of predicting **continuous responses** (i.e. numeric values) which normally answer questions like ‘How many’/ ‘How much’.
Examples:

> *Changes in prices*
> *Fluctuations of temperature*
> *Sales figures*
> *Product demand*

Clustering



Unsupervised Learning — Clustering

A subcategory of UL, Clustering is the process used for exploratory data analysis to find hidden patterns or **groupings/partitions** of data. Examples:

> *Customer segmentation*
> *Market research*
> *Recommendation engine*

Finale

Machine Learning is an exciting subject; it is art and it is science! In this article we have just explored the basics — my aim was to make ML '*as simple as possible, but not one bit simpler*' — as Einstein used to say!

Thanks for reading!

I regularly write about Technology & Data on Medium — if you would like to read my future posts then simply 'Follow' me!