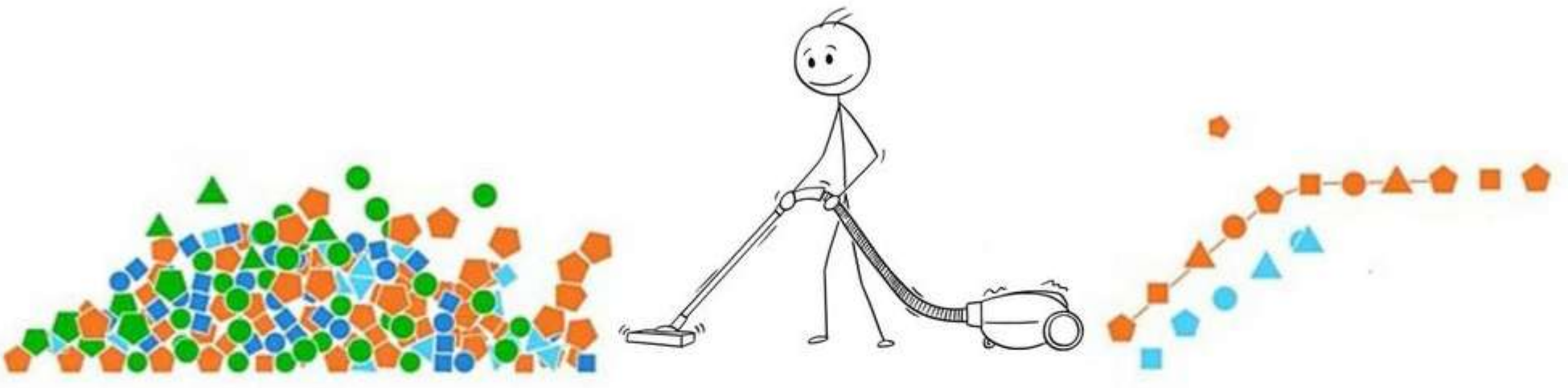
Data wrangling



Data wrangling is the process of gathering, selecting, cleaning, structuring and enriching raw data into the desired format for better decision making in less time.

If you want to create an efficient ETL pipeline (extract, transform, and load) or create beautiful data visualizations, you should be prepared to do a lot of data wrangling- Springboard.

Data Imputation

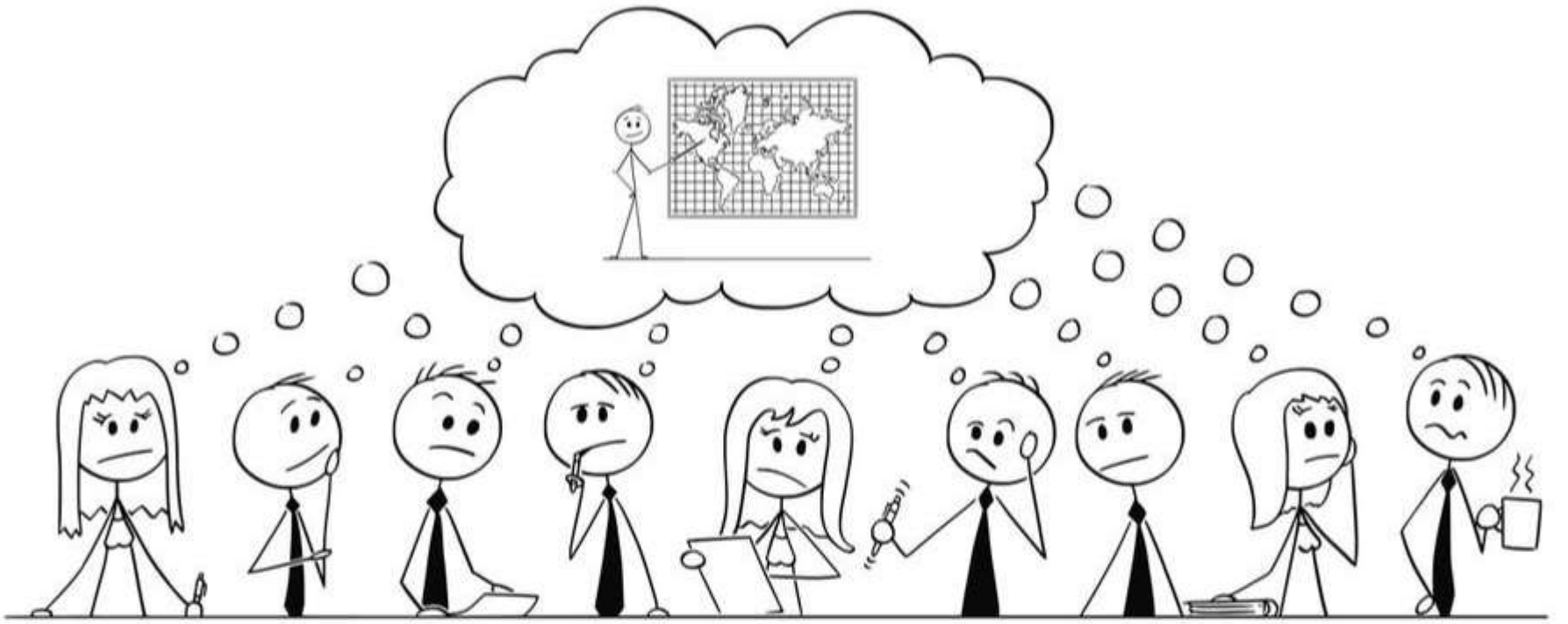
0	2	5.0	3.0	6.0	NaN
1	9	NaN	9.0	0.0	7.0
2	19	17.0	NaN	9.0	NaN
3	7	10.0	3.0	6.0	4.0
4	2	8.0	10.0	NaN	3.0

0	2.0	5.0	3.00	6.00	4.666667
1	9.0	10.0	9.00	0.00	7.000000
2	19.0	17.0	6.25	9.00	4.666667
3	7.0	10.0	3.00	6.00	4.000000
4	2.0	8.0	10.00	5.25	3.000000

Data imputation is the substitution of estimated values for missing or inconsistent data items (fields). The substituted values are intended to create a data record that does not fail edits.

The most common technique is mean imputation, where you take the mean of the existing data in the field and fill in the blanks with this.

Supervised learning



Supervised learning is an approach to creating artificial intelligence (AI), where the program is given labelled input data and the expected output results.

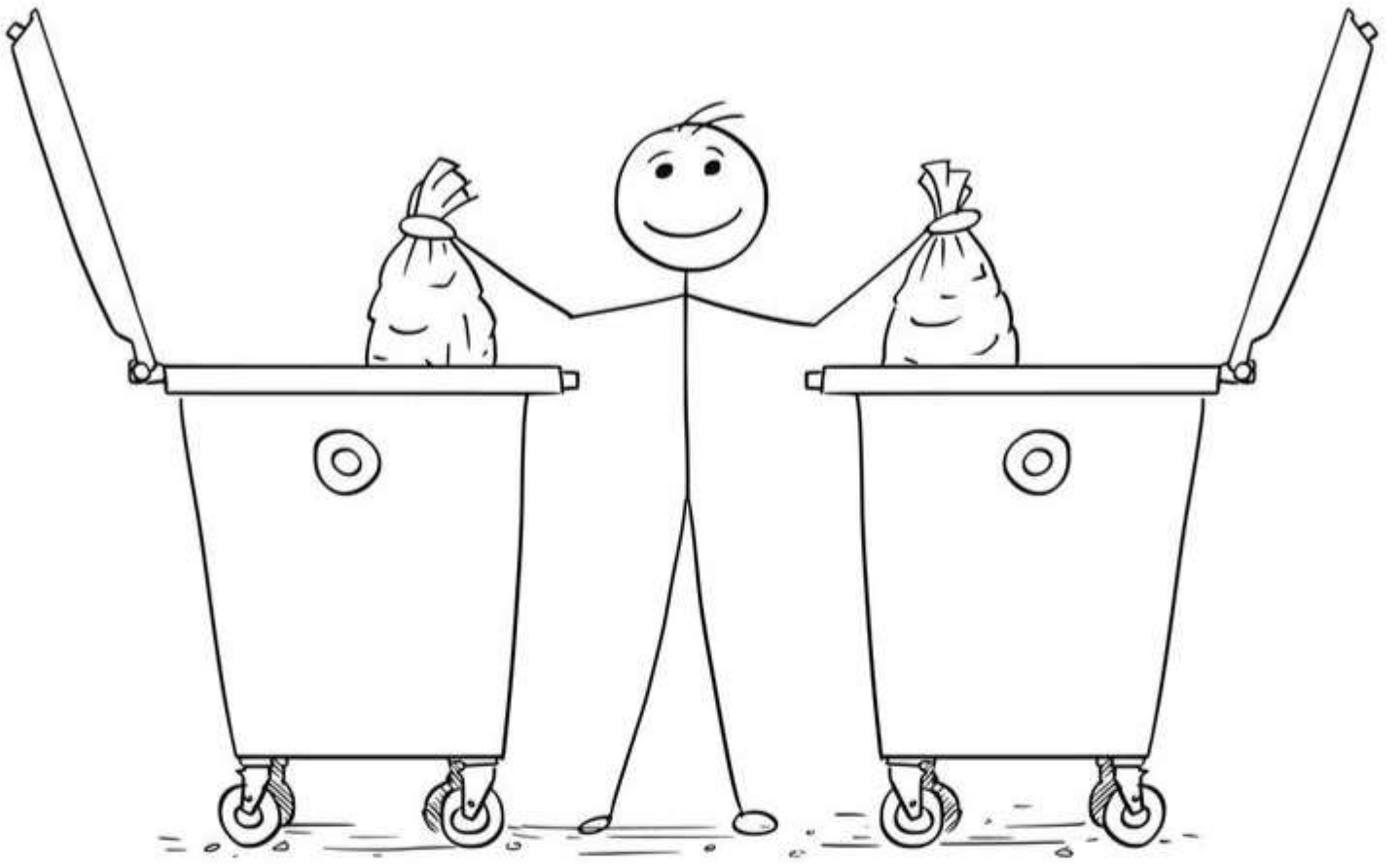
The AI system is specifically told what to look for, thus the model is trained until it can detect the underlying patterns and relationships, enabling it to yield good results when presented with never-before-seen data.

UnSupervised learning

In unsupervised learning, a dataset is provided without labels, and a model learns useful properties of the structure of the dataset. We do not tell the model what it must learn, but allow it to find patterns and draw conclusions from the unlabeled data.

The algorithms in unsupervised learning are more difficult than in supervised learning since we have little or no information about the data. Unsupervised learning tasks typically involve grouping similar examples together, dimensionality reduction, and density estimation.

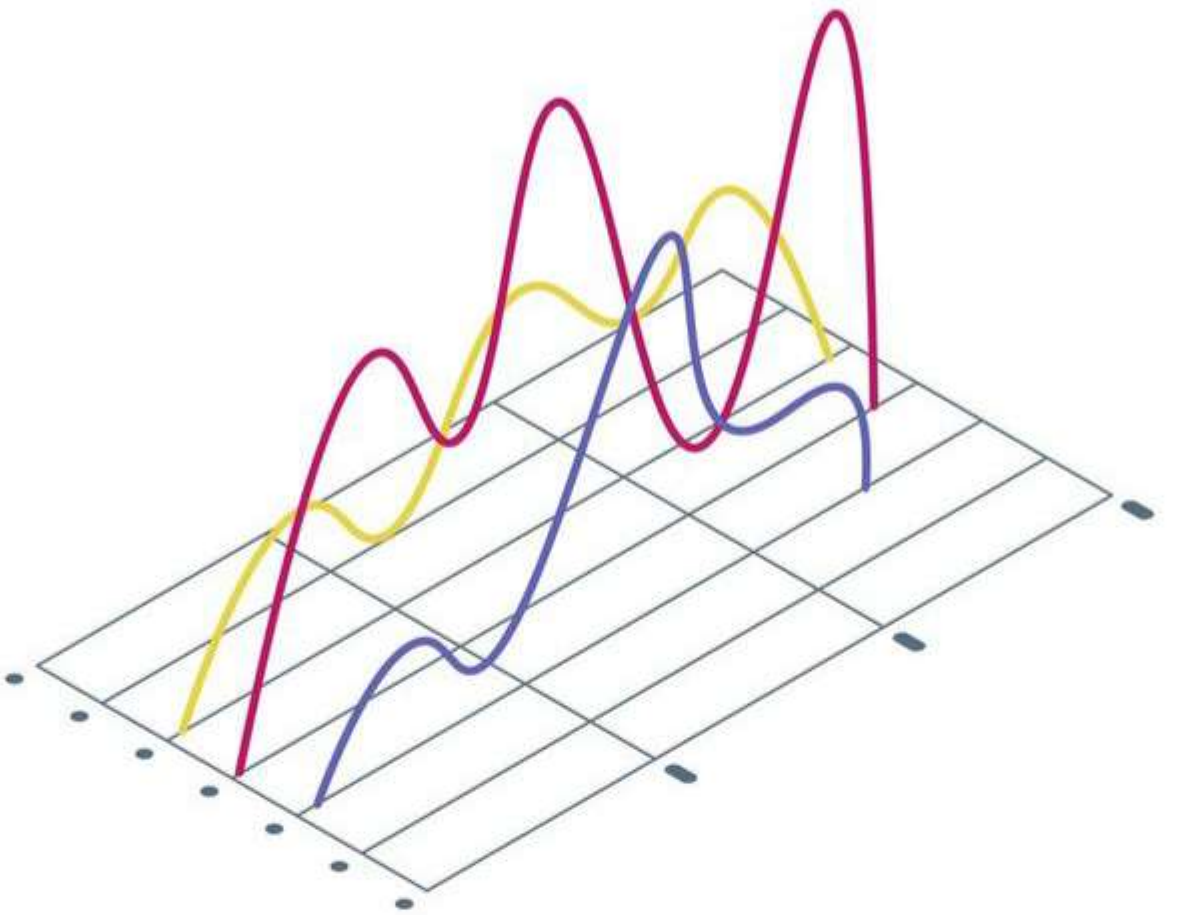
Classification



A classification algorithm tries to determine the class or the category of the data it is presented with.

Many times, an object might belong to several categories, and the AI needs to determine what those categories are and how much confidence the algorithm has in its predictions.

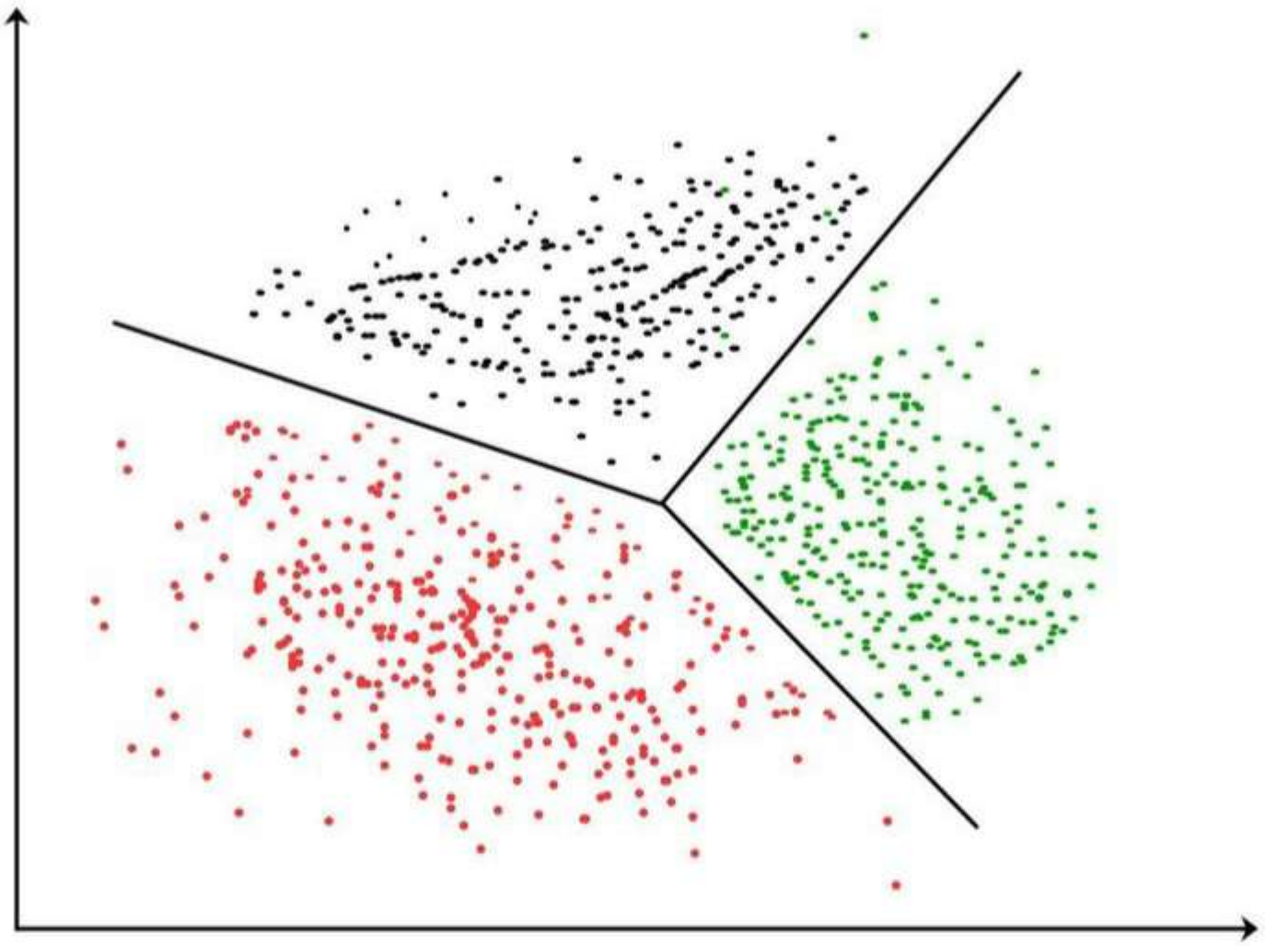
Regression



Regression is the type of Supervised Learning in which labelled data is used, and this data is used to make predictions in a continuous form.

Regression problems include types where the output variables are set as a real number. The format for this problem often follows a linear format.

Clustering



Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Evaluation metrics

Evaluation metrics are used to measure the quality of the statistical or machine learning model.

There are many different types of evaluation metrics available to test a model. These include classification accuracy, logarithmic loss, confusion matrix, and others.