Machine Learning

WHAT IS DATA SCIENCE

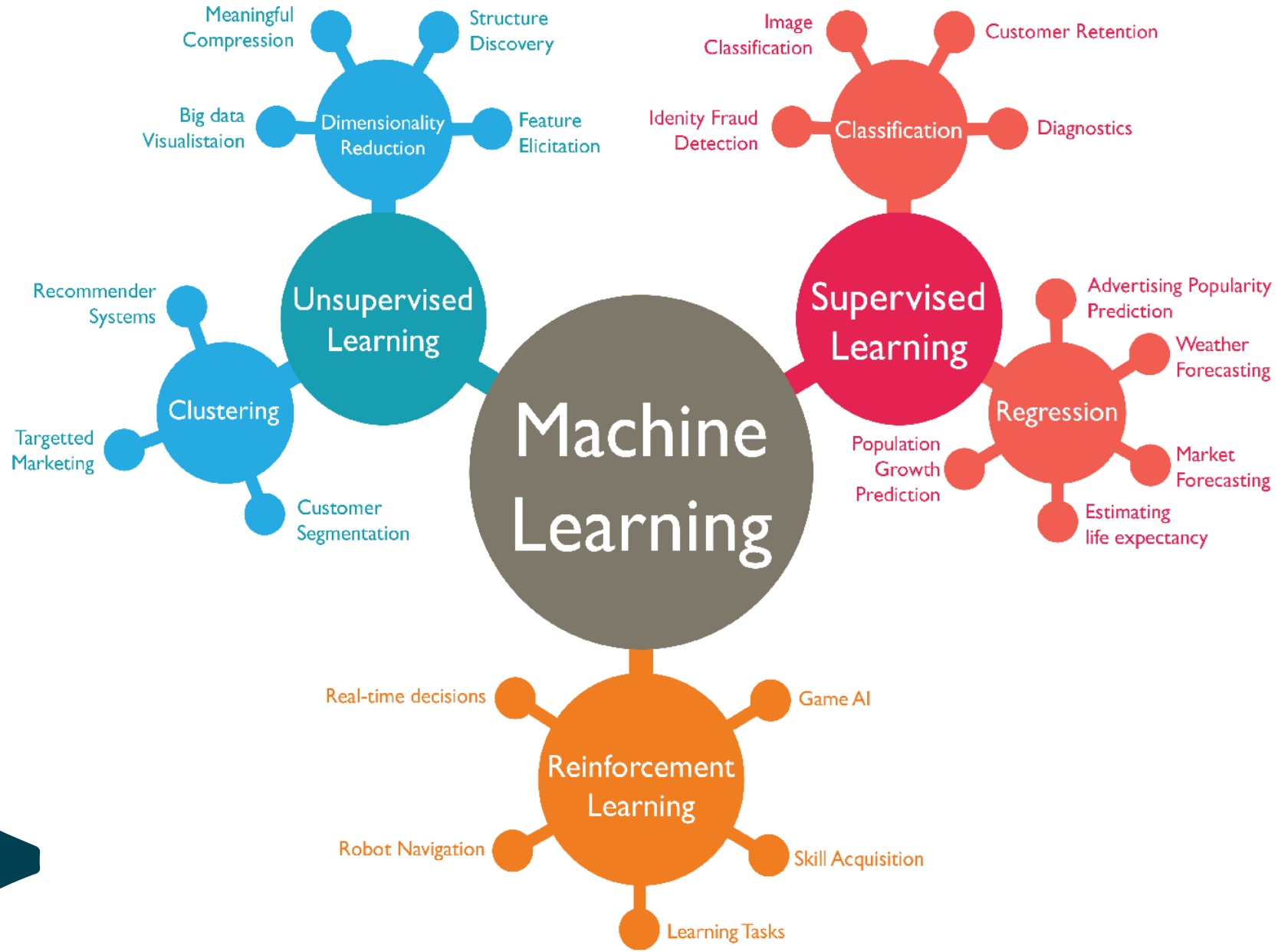
“At its core, data science involves using automated methods to analyse massive amounts of data and to extract knowledge from them.”

What is Machine Learning?

WHAT IS MACHINE LEARNING

Machine learning uses sophisticated algorithms to “learn” from massive volumes of Big Data. The more data the algorithms can access, the more they can learn.

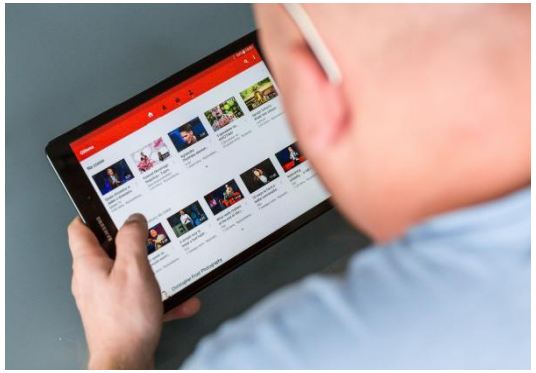
Types of Machine Learning



Applications of Machine Learning

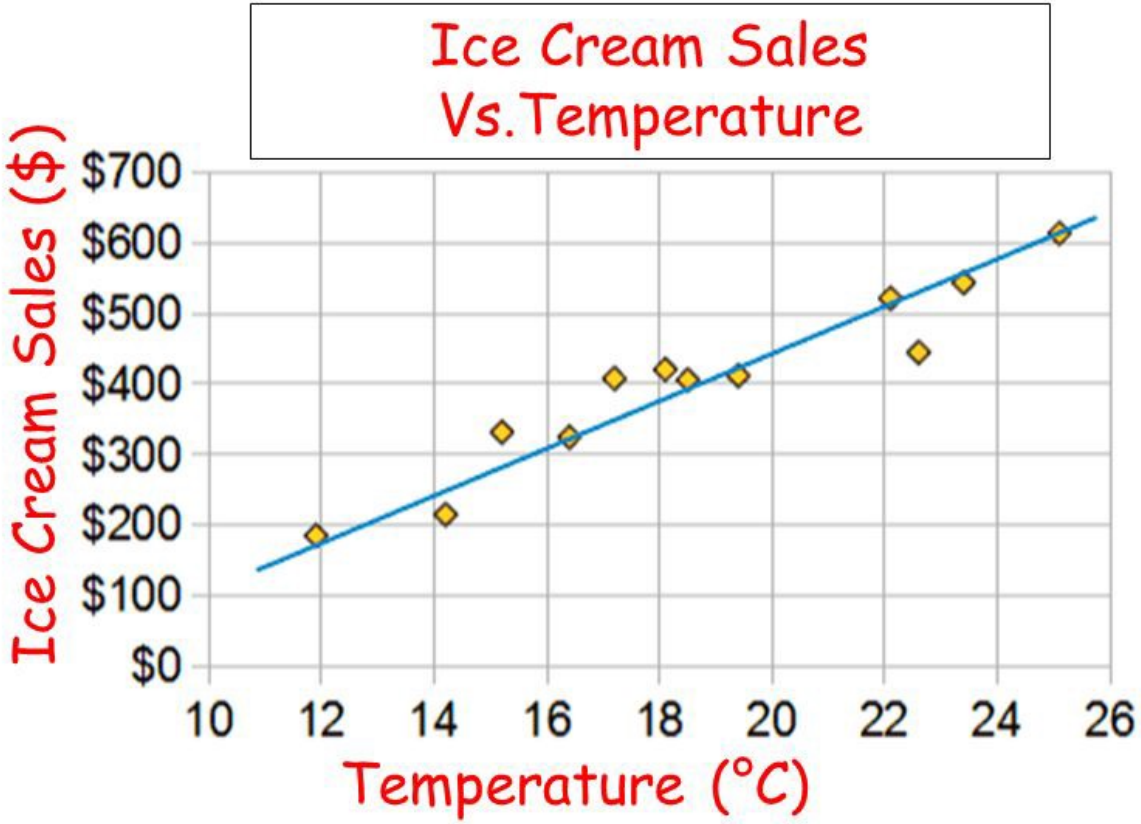
Autonomous Driving

AlphaGo



Youtube recommendations

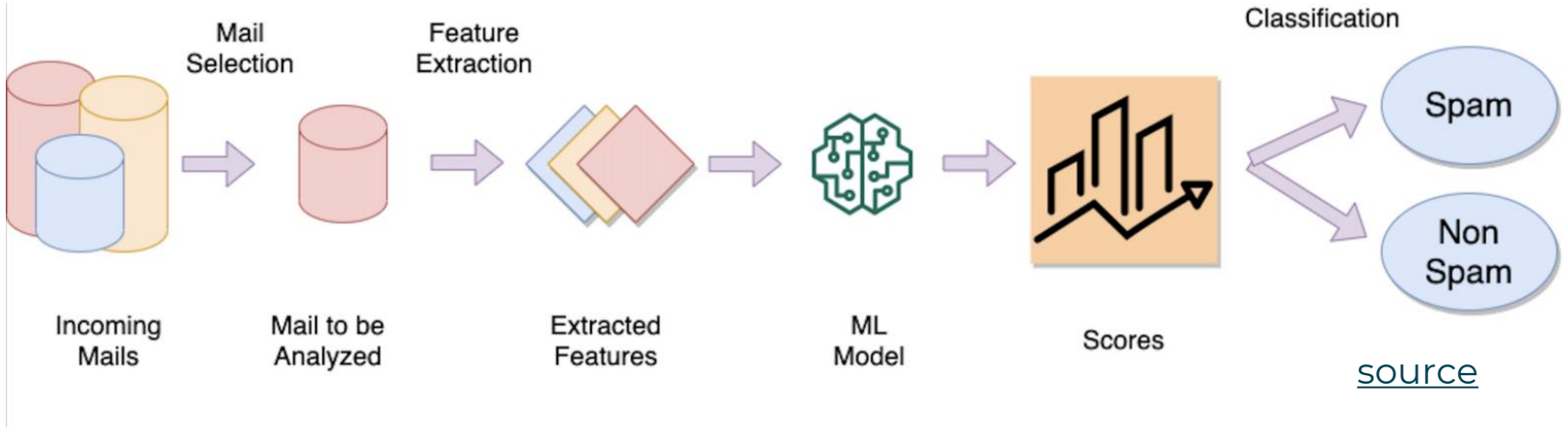
REGRESSION



Can you think of any other information that may be helpful for this model?

Here we only have one input, but if we had more, what ideas would you have for them?

CLASSIFICATION



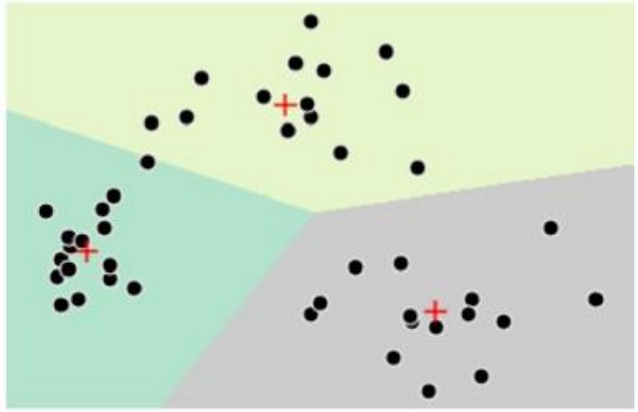
source

Classic example of classifying if emails are spam or not.

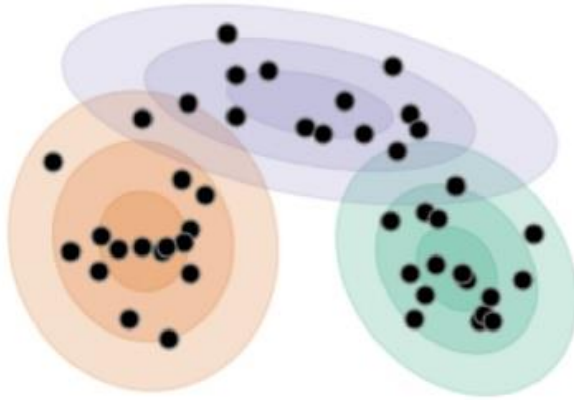
This is important functionality to have as cyber attacks are common place.

As we know, no system is perfect, especially when there are people deliberately trying to bypass them

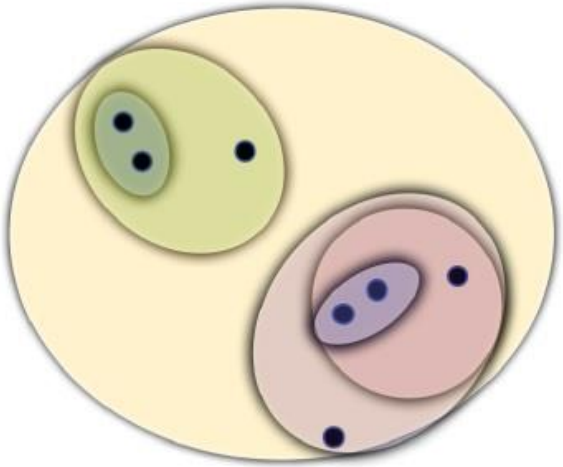
CLUSTERING



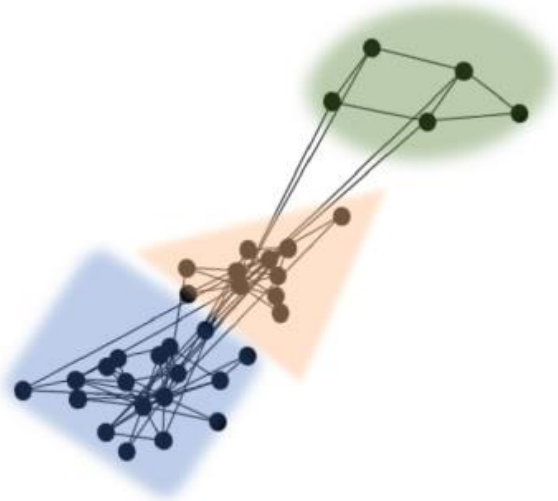
K-means clustering



Mixture model (Gaussian)



Hierarchical clustering



Graph based clustering

Bias and variance

BIAS AND VARIANCE

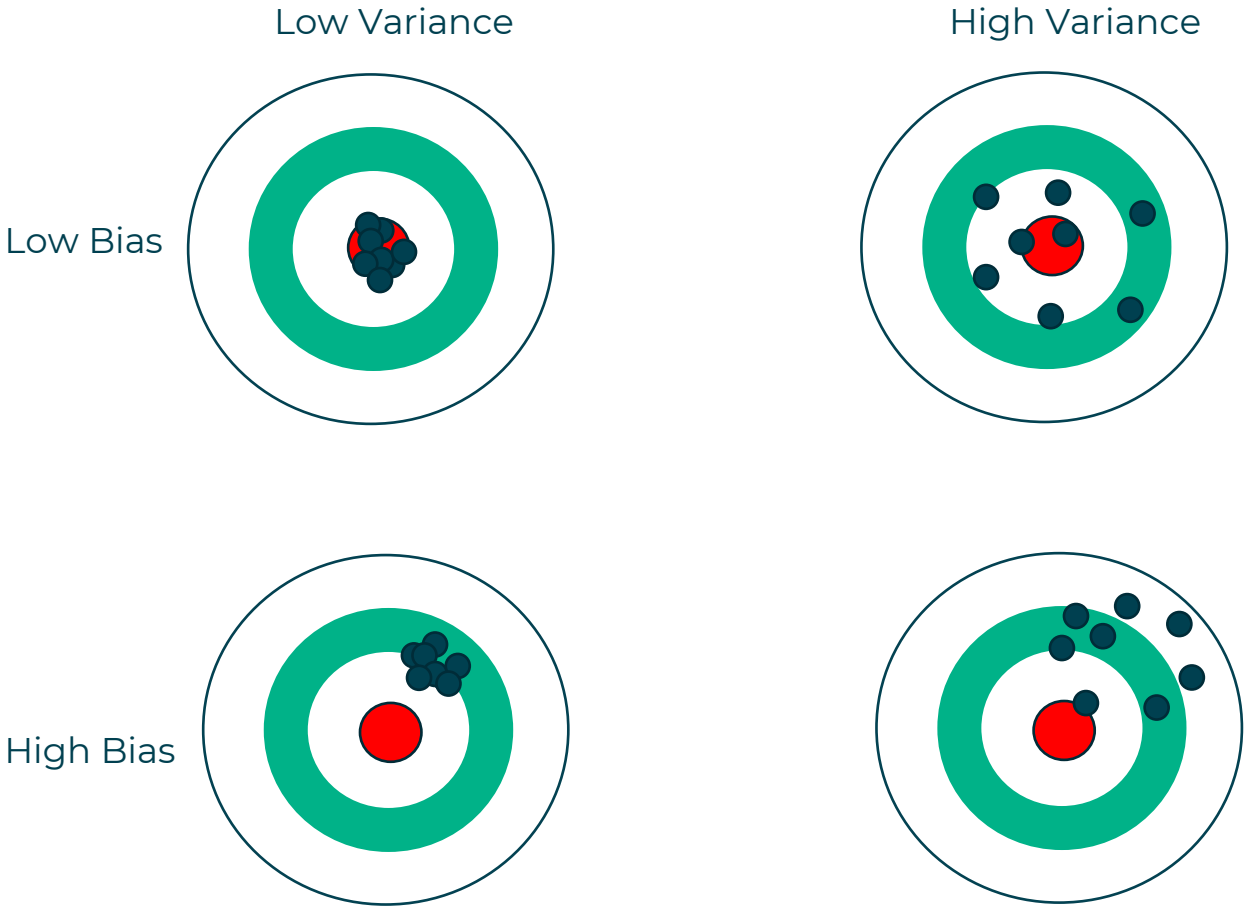
When we build a model, we are trying to create a model that is a perfect fit to the population

Typically there is error in the model and that error can be due to two broad categories:

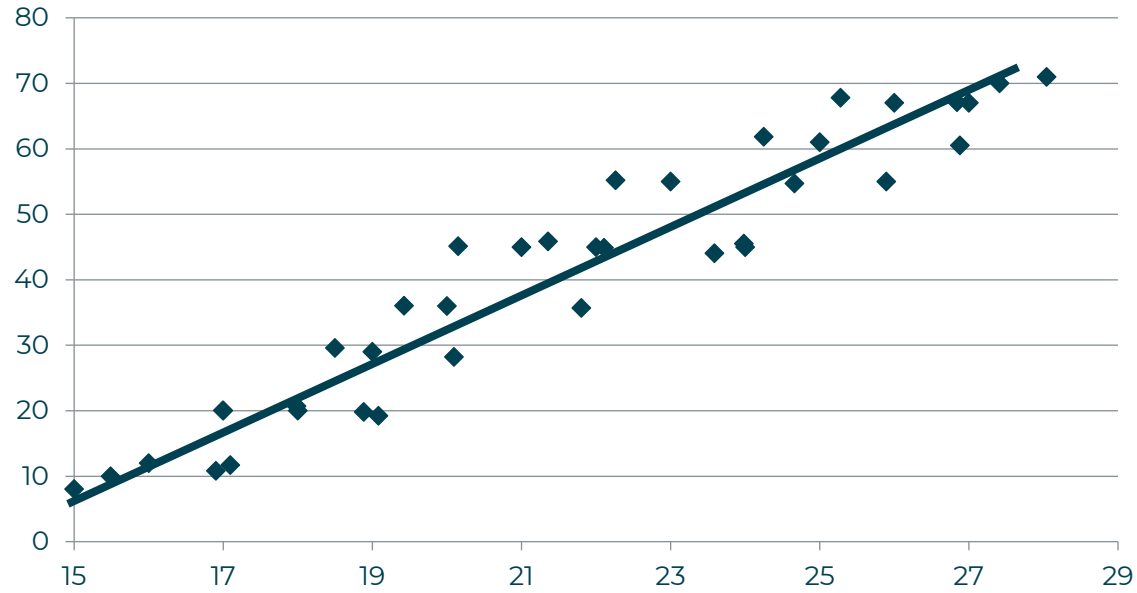
Irreducible error

Reducible error

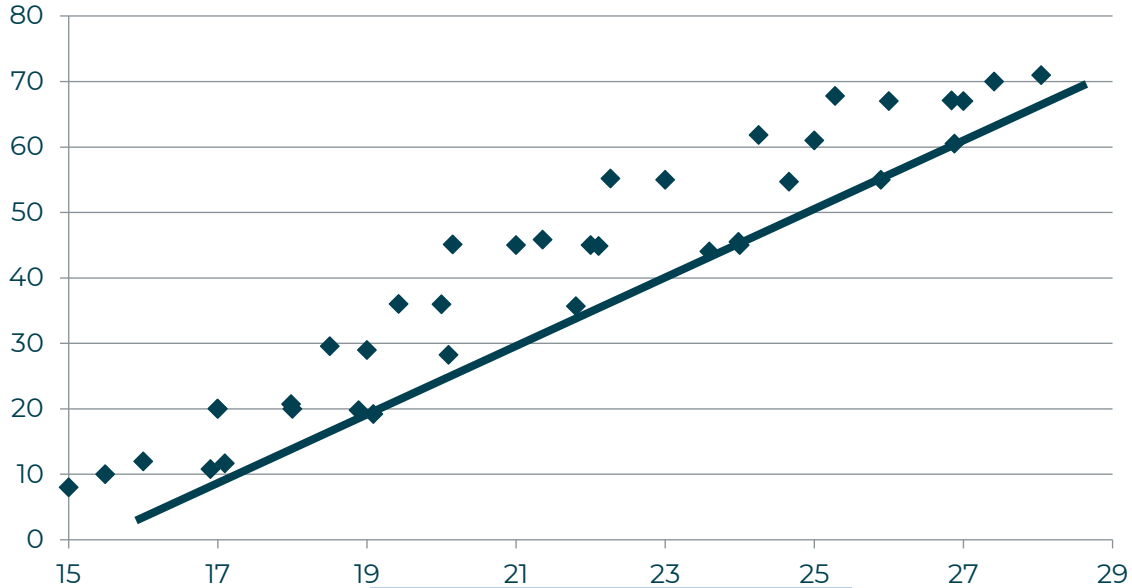
BIAS AND VARIANCE



BIAS

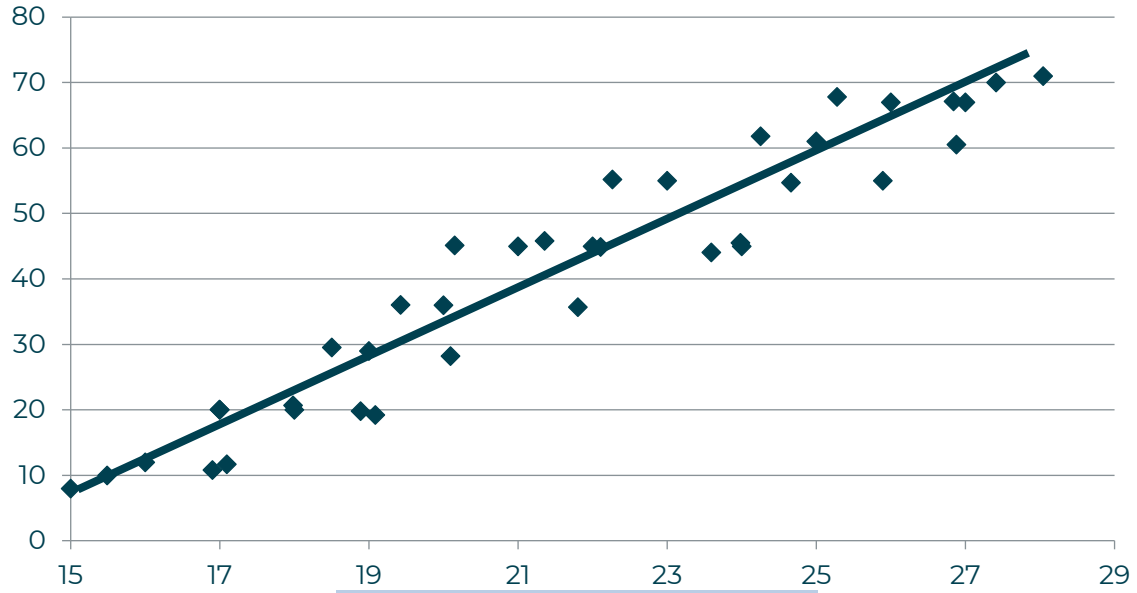


Low Bias

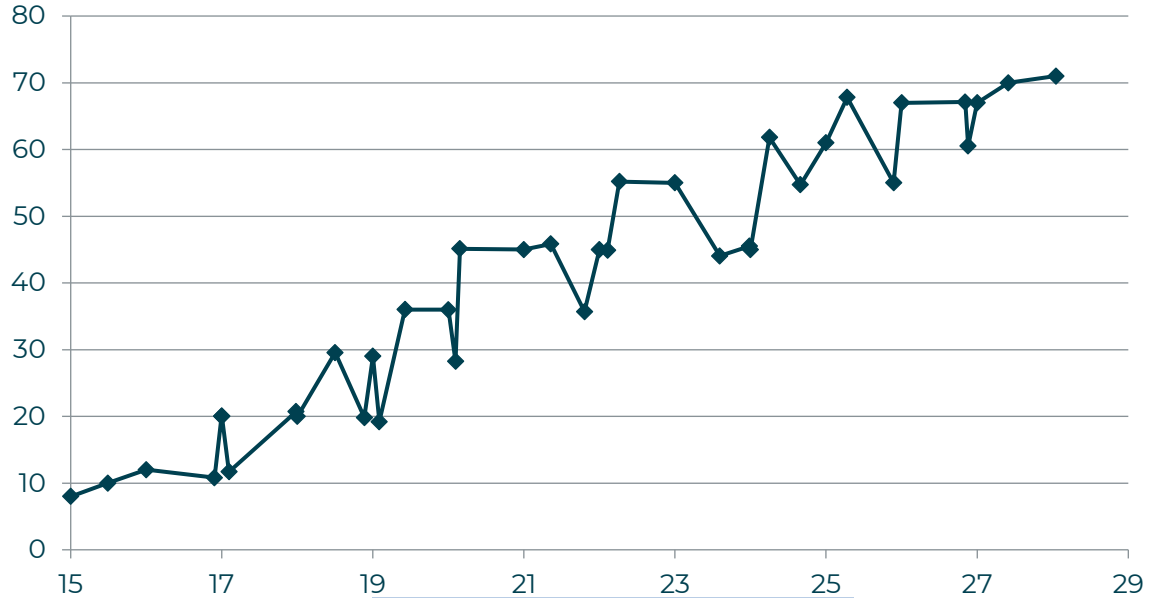


High Bias

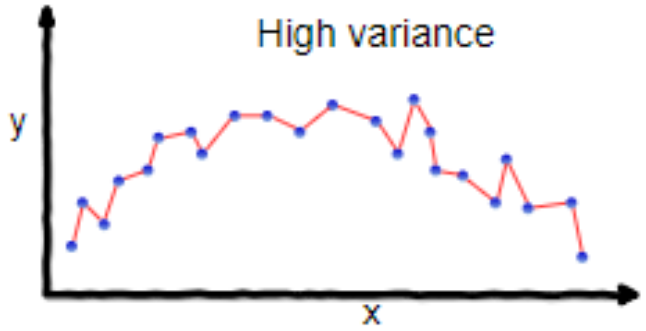
VARIANCE



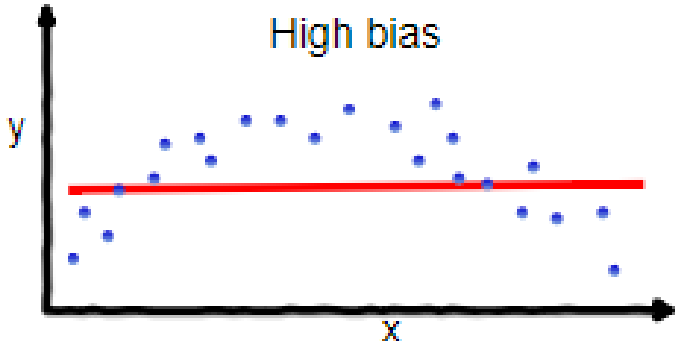
Low Variance



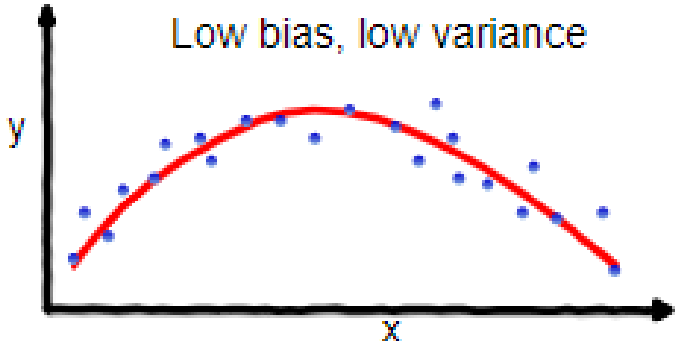
High Variance



overfitting

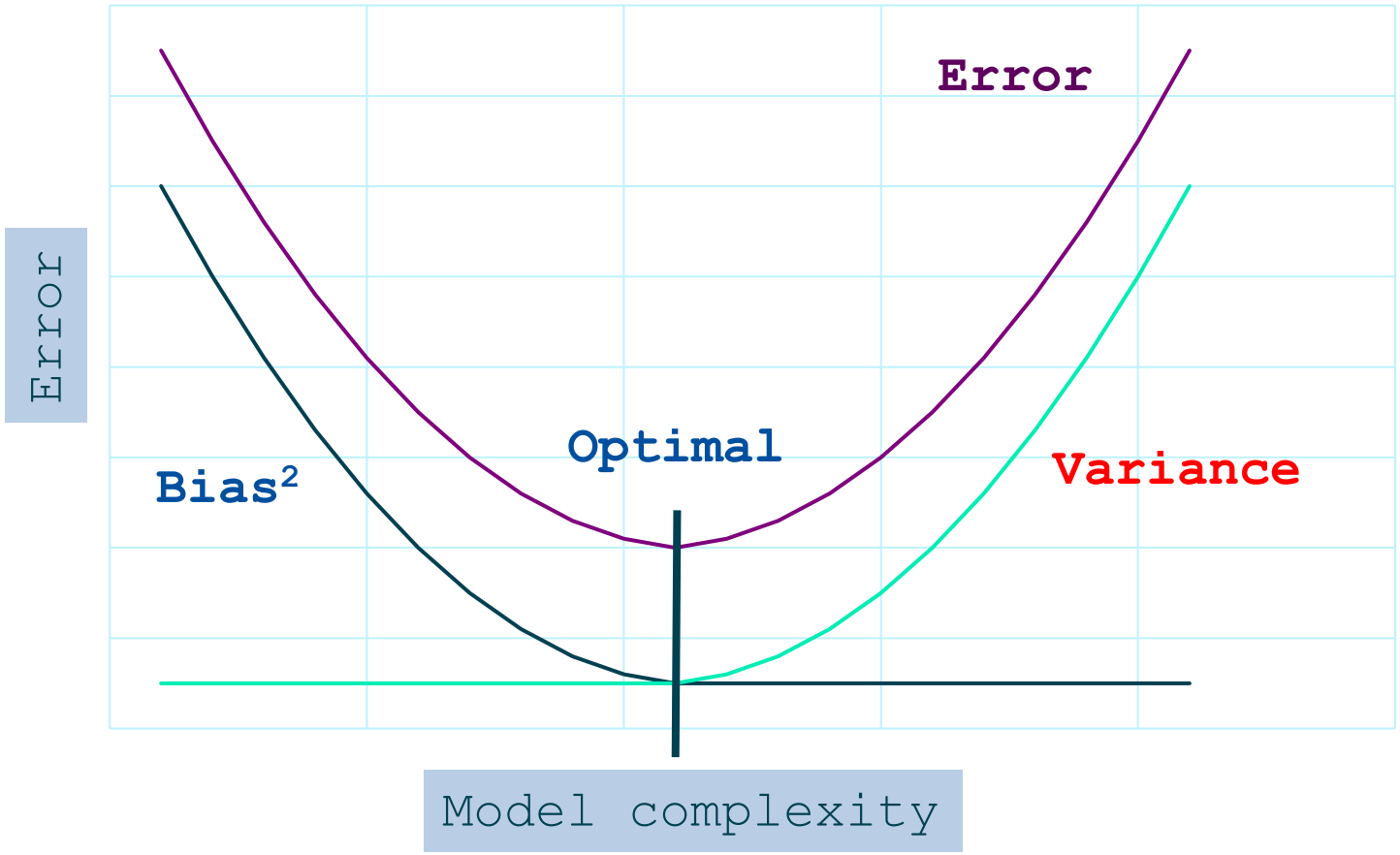


underfitting

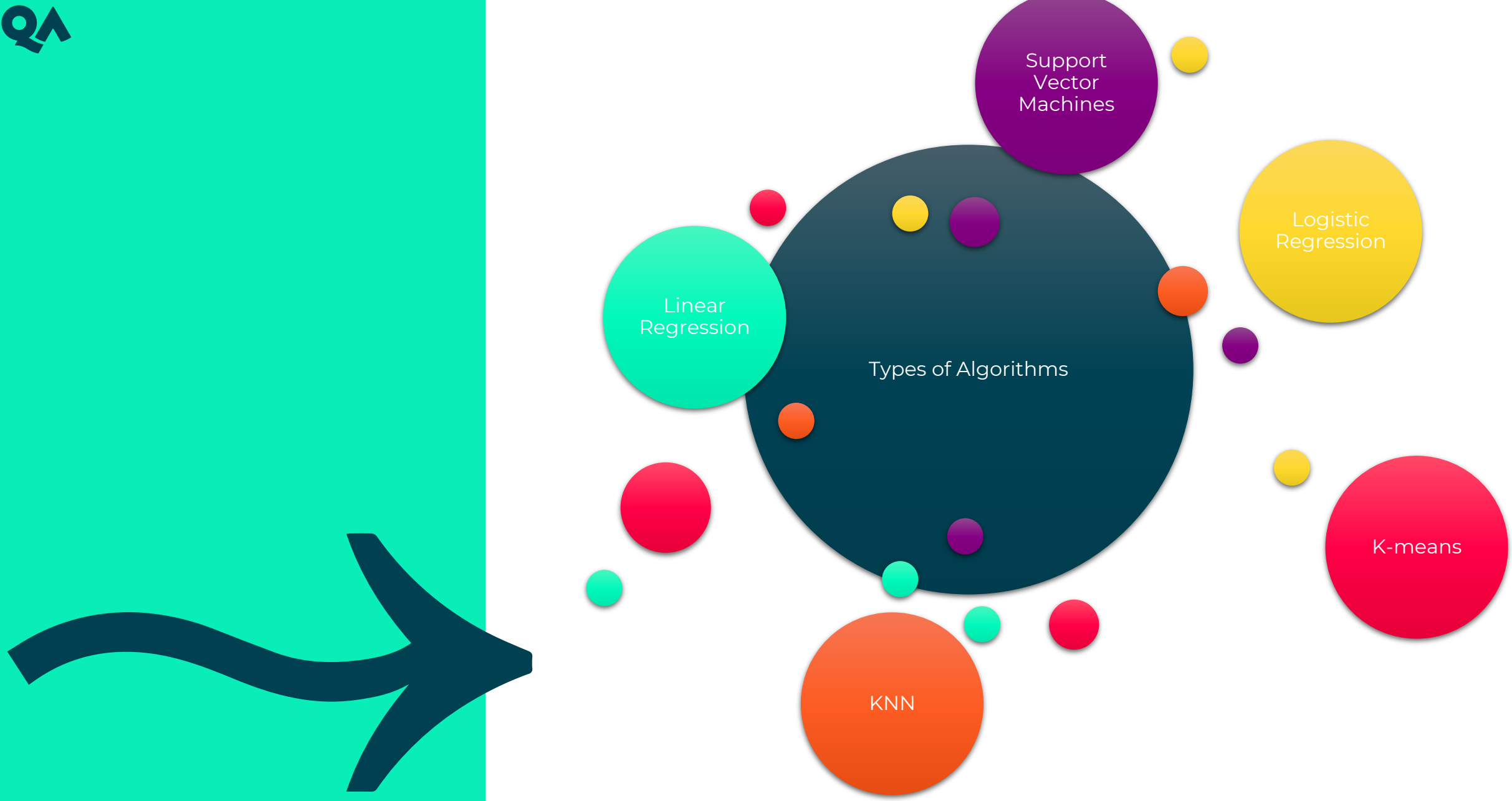


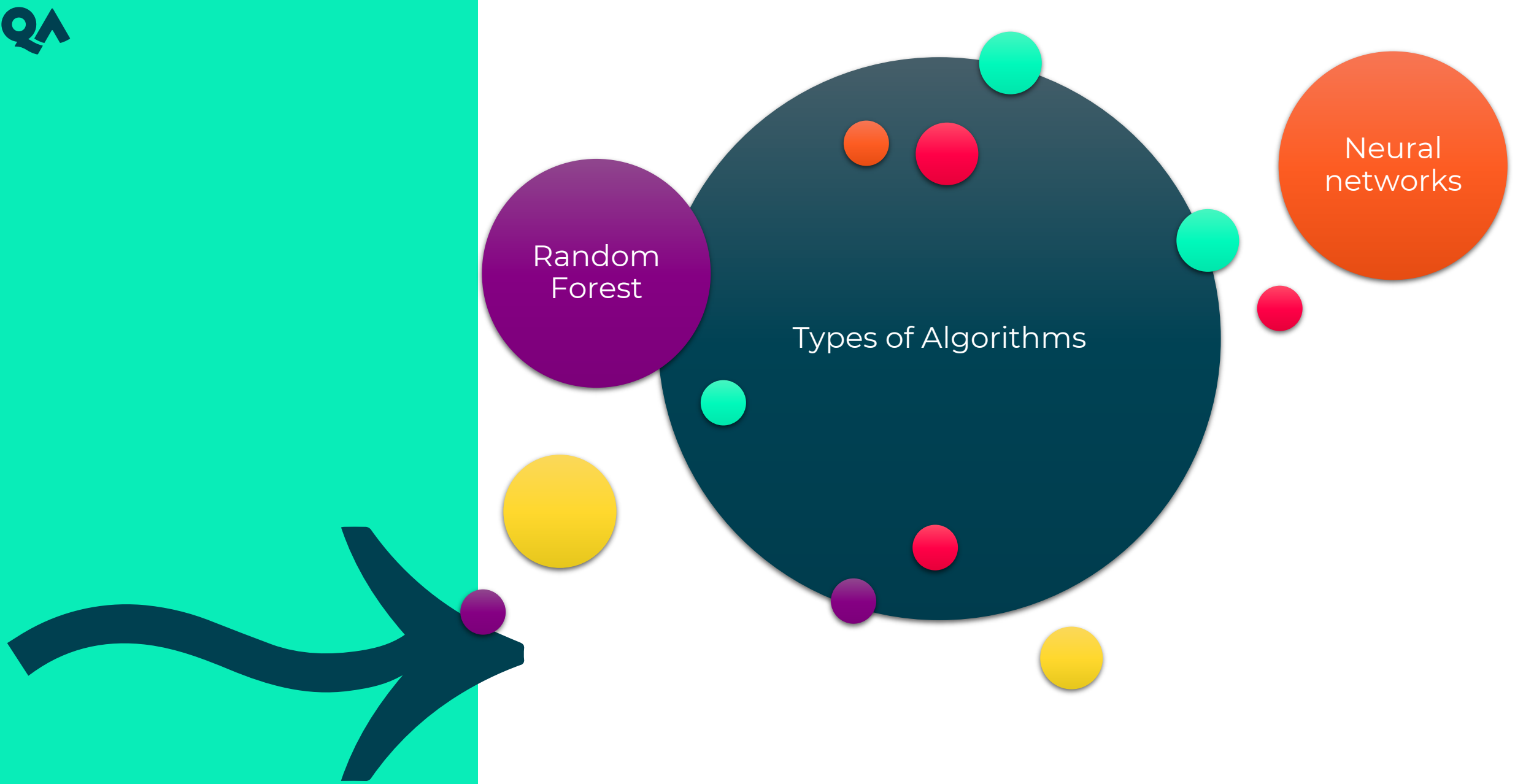
Good balance

BIAS AND VARIANCE




Algorithm Selection


Support Vector Machines
Logistic Regression
Linear Regression
Types of Algorithms
K-means
KNN


Random Forest
Types of Algorithms
Neural networks

CRITERIA

Considerations when choosing a model

- Is it a regression or classification problem?
- How well does this model usually perform with this task?
  - Research and
  - Experiment if possible
- What available resources do you have? – GPU's for NNs

The biggest question of all

- What is the data like?
- Is it good enough?
- is there enough of it?


End of unit