MTU

Ollscoil Teicneolaíochta na Mumhan
Munster Technological University

Machine Learning

Lecture 2: Introduction to machine learning

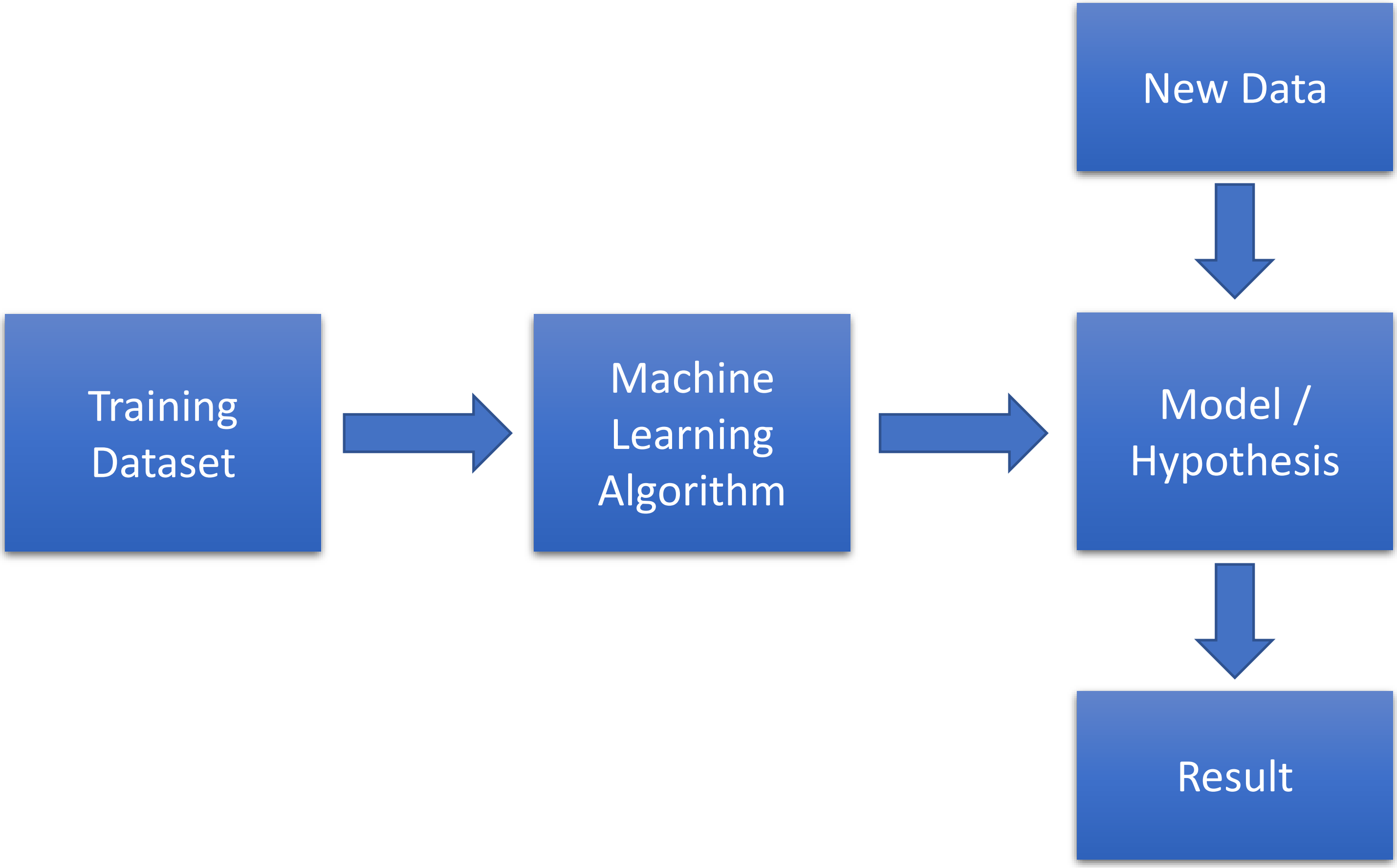
Reminder: What is Machine Learning?

- Machine learning is the use of statistical analysis of existing data to infer a relation between input and output with the goal of computing this relationship for new data

- Depending on the output data we distinguish
  - Classification
  - Regression
  - Clustering

- Depending on the input data we distinguish
  - Supervised learning algorithms
  - Unsupervised learning algorithms
  - Reinforcement learning algorithms

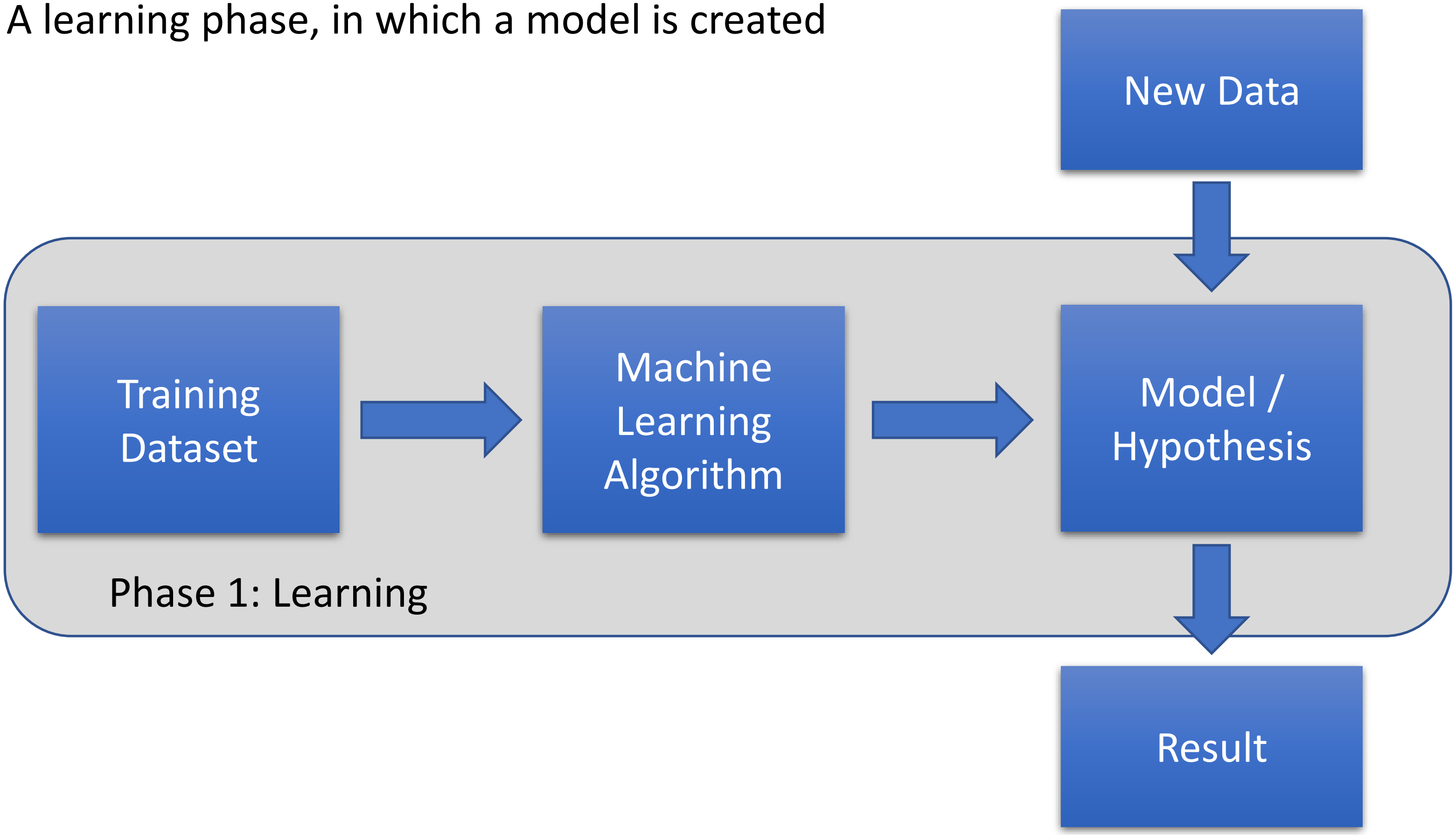
High-level view on Machine Learning



High-level view on Machine Learning

A typical ML problem can be divided into two phases:

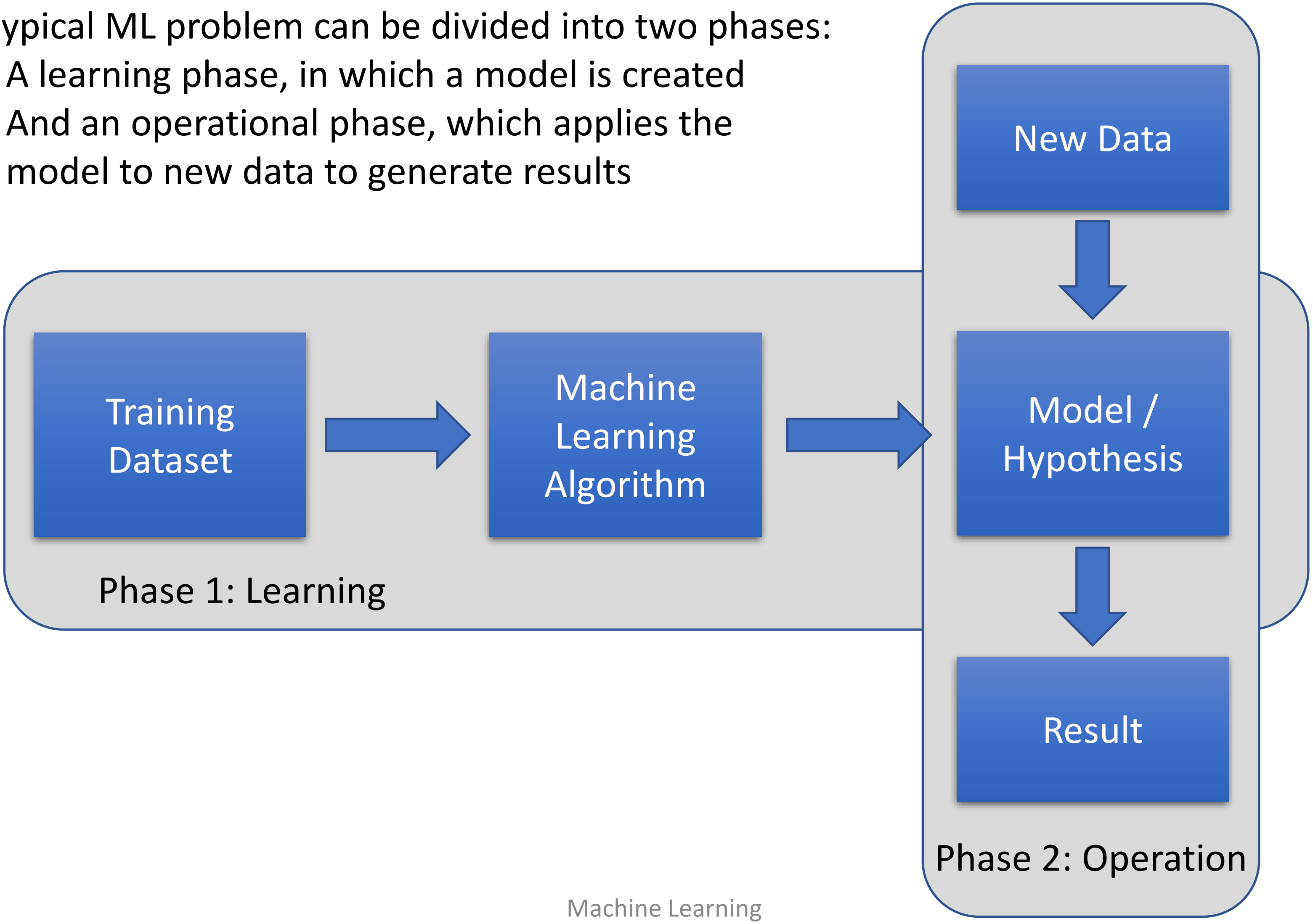
- A learning phase, in which a model is created



High-level view on Machine Learning

A typical ML problem can be divided into two phases:

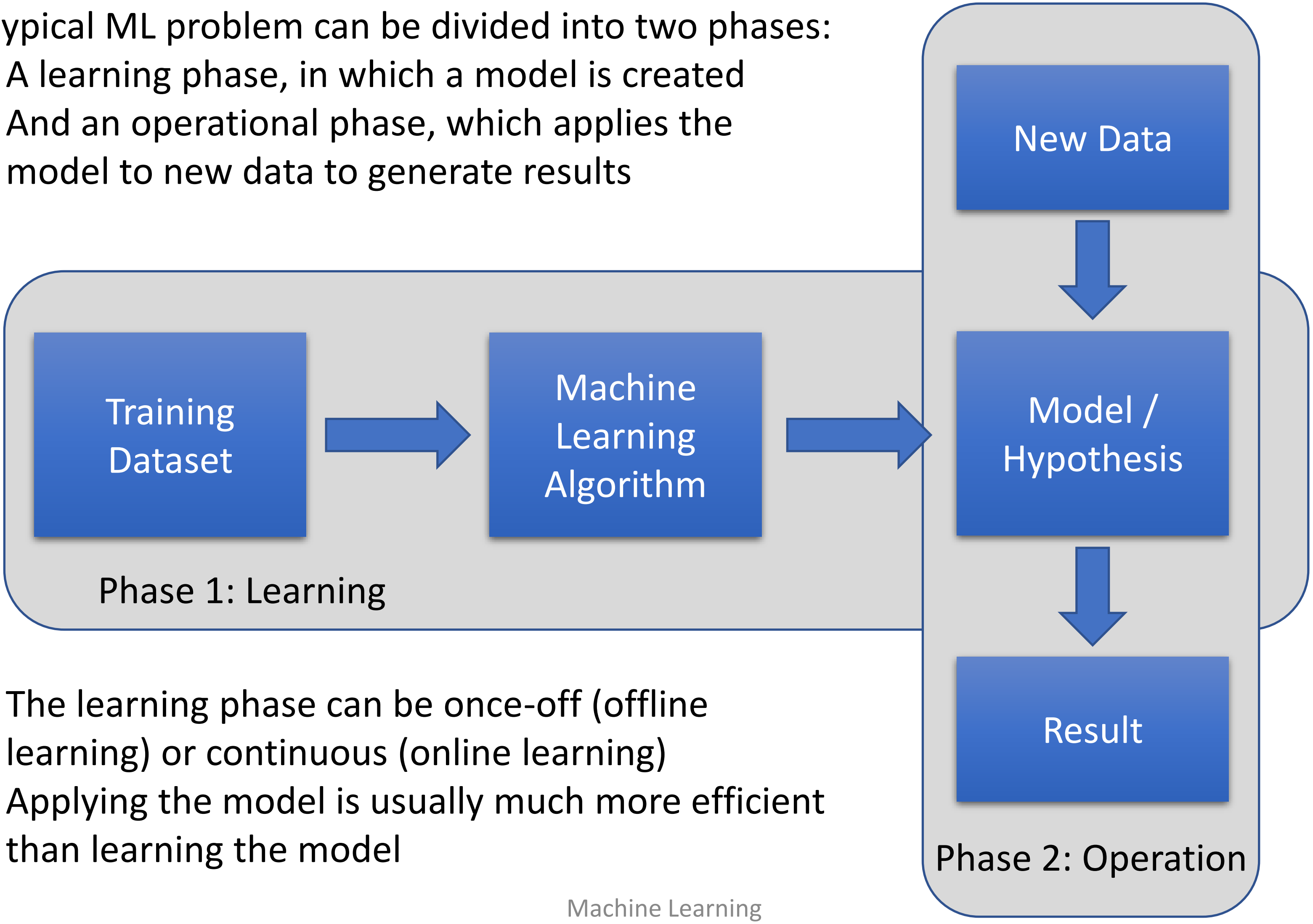
- A learning phase, in which a model is created
- And an operational phase, which applies the model to new data to generate results



High-level view on Machine Learning

A typical ML problem can be divided into two phases:

- A learning phase, in which a model is created
- And an operational phase, which applies the model to new data to generate results

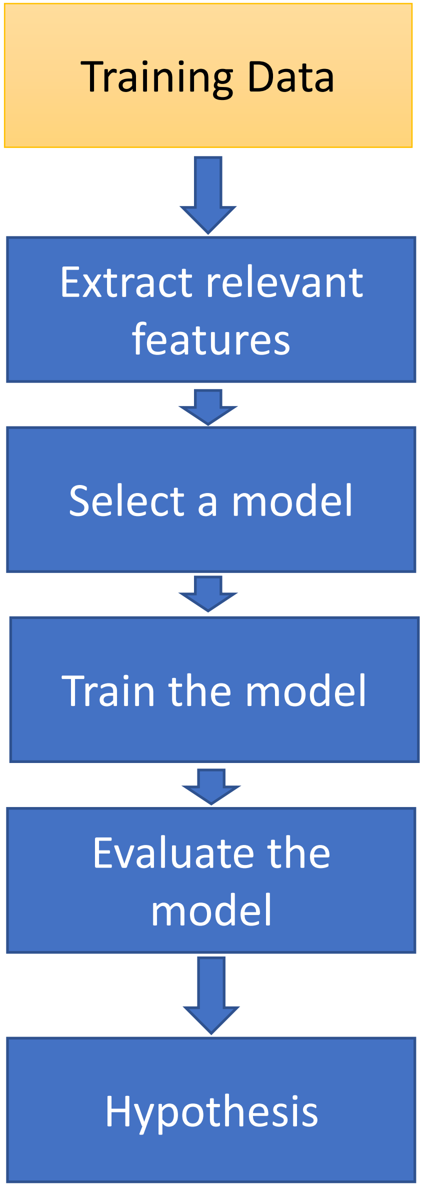


- The learning phase can be once-off (offline learning) or continuous (online learning)
- Applying the model is usually much more efficient than learning the model

The Learning Phase

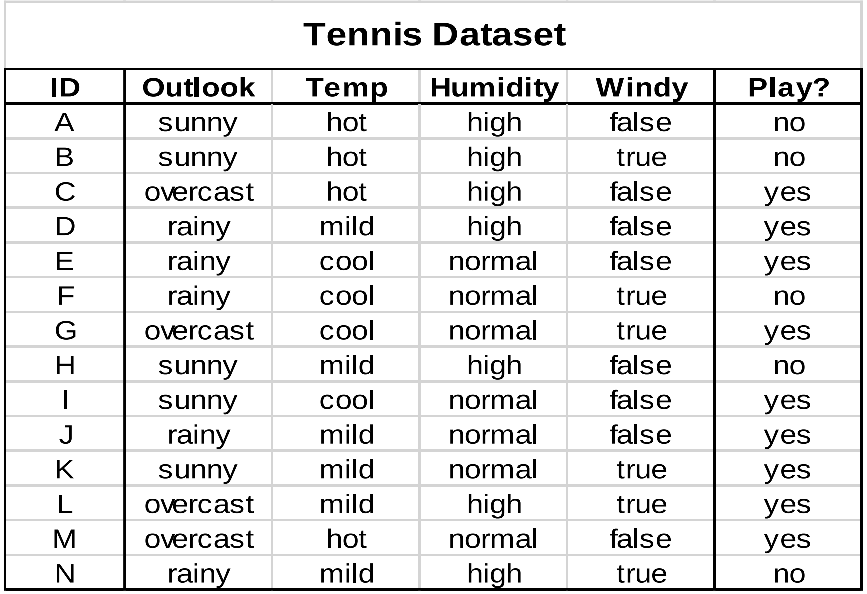
- Training data collection is often the most time consuming, most difficult and most expensive step of the process
- ML algorithms rely solely on statistical patterns in this dataset, so it needs to encode everything there is to know about the problem domain

- The ultimate goal is a balanced, bias-free, representative, outlier-free and complete set of samples



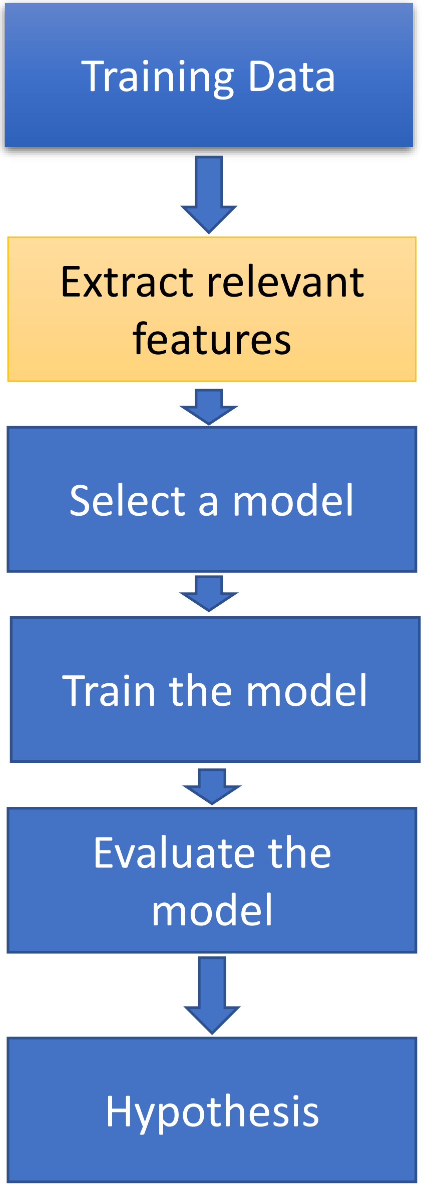
The Learning Phase

- A set of features needs to be extracted from the training data
- This can involve aggregation, categorisation or indeed any form of processing of the raw training data
- For example a temperature measurement can be categorised as “hot”, when > 25deg



Tennis Dataset

ID	Outlook	Temp	Humidity	Windy	Play?
A	sunny	hot	high	false	no
B	sunny	hot	high	true	no
C	overcast	hot	high	false	yes
D	rainy	mild	high	false	yes
E	rainy	cool	normal	false	yes
F	rainy	cool	normal	true	no
G	overcast	cool	normal	true	yes
H	sunny	mild	high	false	no
I	sunny	cool	normal	false	yes
J	rainy	mild	normal	false	yes
K	sunny	mild	normal	true	yes
L	overcast	mild	high	true	yes
M	overcast	hot	normal	false	yes
N	rainy	mild	high	true	no



Training Data

Extract relevant features

Select a model

Train the model

Evaluate the model

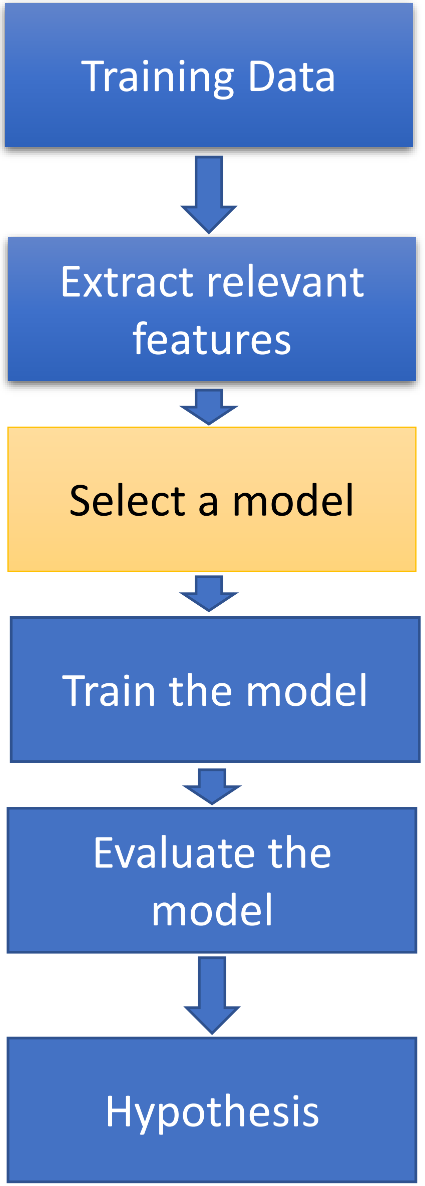
Hypothesis

The Learning Phase

- A model needs to be selected, which determines how the final hypothesis can look like
- This is either a class of functional relationships between input and output, e.g. some class of f so that f[O,T,H,W] → Play, for supervised learning
- Or some measure of distance between features for unsupervised learning (more later, when we discuss unsupervised learning)

Tennis Dataset

ID	Outlook	Temp	Humidity	Windy	Play?
A	sunny	hot	high	false	no
B	sunny	hot	high	true	no
C	overcast	hot	high	false	yes
D	rainy	mild	high	false	yes
E	rainy	cool	normal	false	yes
F	rainy	cool	normal	true	no
G	overcast	cool	normal	true	yes
H	sunny	mild	high	false	no
I	sunny	cool	normal	false	yes
J	rainy	mild	normal	false	yes
K	sunny	mild	normal	true	yes
L	overcast	mild	high	true	yes
M	overcast	hot	normal	false	yes
N	rainy	mild	high	true	no



Training Data

Extract relevant features

Select a model

Train the model

Evaluate the model

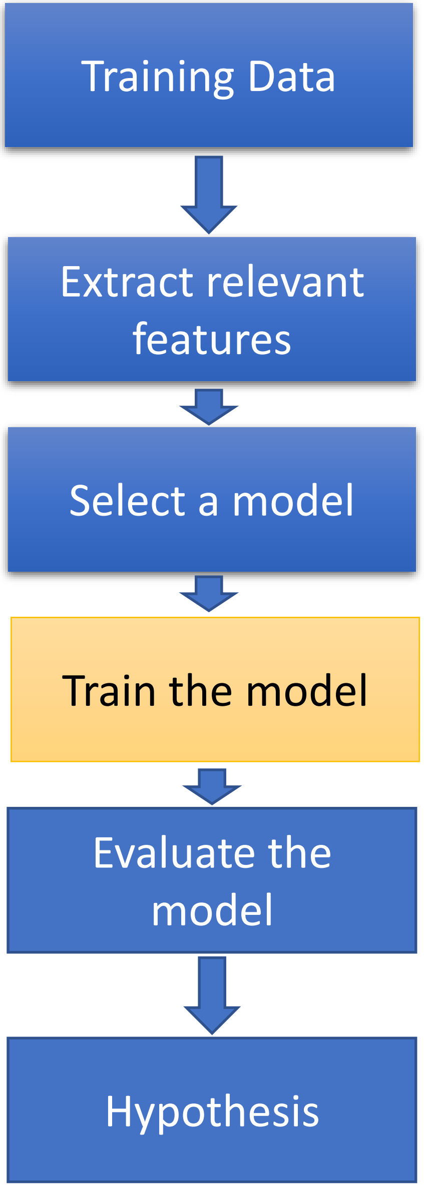
Hypothesis

The Learning Phase

- The model is trained, i.e. the functional relationship between input and output features is estimated based on the examples in the training data
- Out of the class of functions from the previous step a particular functional relationship is generated
- We now have a tentative hypothesis, allowing to predict values for new data, e.g. predict if someone plays tennis given the current weather prediction

Tennis Dataset

ID	Outlook	Temp	Humidity	Windy	Play?
A	sunny	hot	high	false	no
B	sunny	hot	high	true	no
C	overcast	hot	high	false	yes
D	rainy	mild	high	false	yes
E	rainy	cool	normal	false	yes
F	rainy	cool	normal	true	no
G	overcast	cool	normal	true	yes
H	sunny	mild	high	false	no
I	sunny	cool	normal	false	yes
J	rainy	mild	normal	false	yes
K	sunny	mild	normal	true	yes
L	overcast	mild	high	true	yes
M	overcast	hot	normal	false	yes
N	rainy	mild	high	true	no



Training Data

Extract relevant features

Select a model

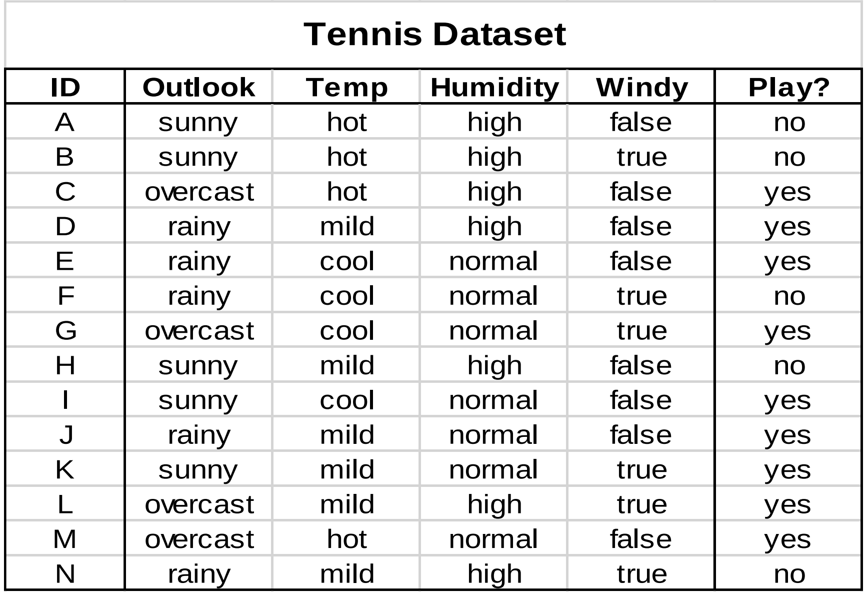
Train the model

Evaluate the model

Hypothesis

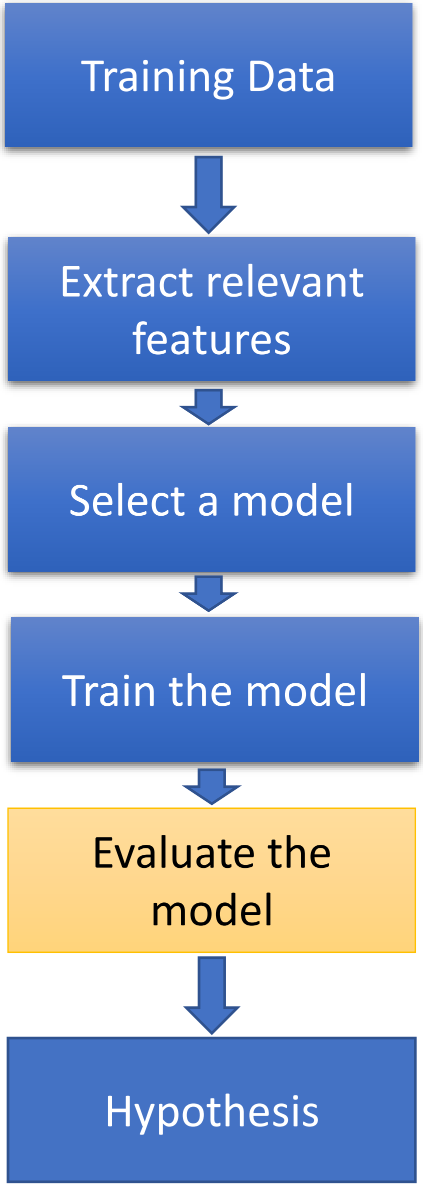
The Learning Phase

- A lot of decisions went into the process up to here, the question is how valid are our assumptions
- We can run our tentative hypothesis on the training data and see how it performs there
- Because we fitted our hypothesis to the training data the result should be good
- Unfortunately this only tells us that we made no mistake up until now
- To evaluate the performance of our hypothesis, we need to test against new data or split the training set into two subsets



Tennis Dataset

ID	Outlook	Temp	Humidity	Windy	Play?
A	sunny	hot	high	false	no
B	sunny	hot	high	true	no
C	overcast	hot	high	false	yes
D	rainy	mild	high	false	yes
E	rainy	cool	normal	false	yes
F	rainy	cool	normal	true	no
G	overcast	cool	normal	true	yes
H	sunny	mild	high	false	no
I	sunny	cool	normal	false	yes
J	rainy	mild	normal	false	yes
K	sunny	mild	normal	true	yes
L	overcast	mild	high	true	yes
M	overcast	hot	normal	false	yes
N	rainy	mild	high	true	no



Training Data

Extract relevant features

Select a model

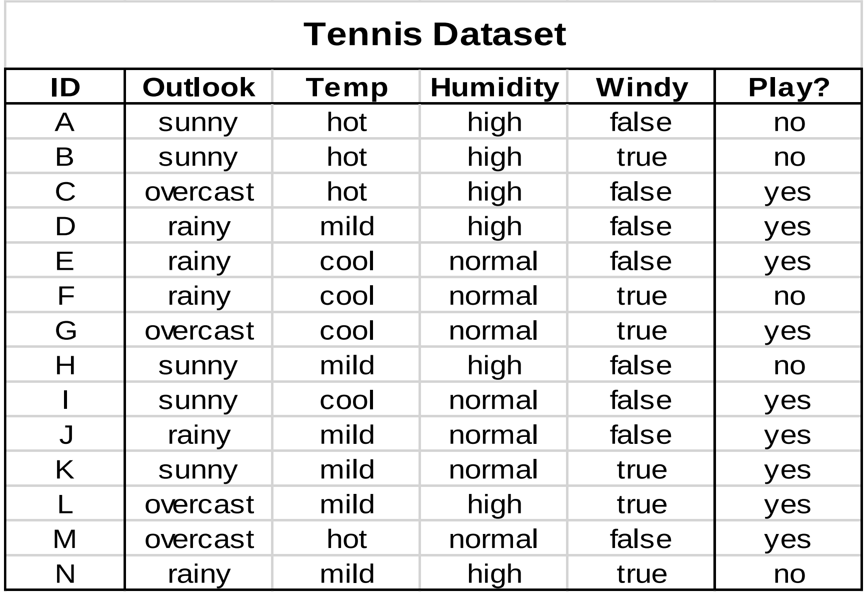
Train the model

Evaluate the model

Hypothesis

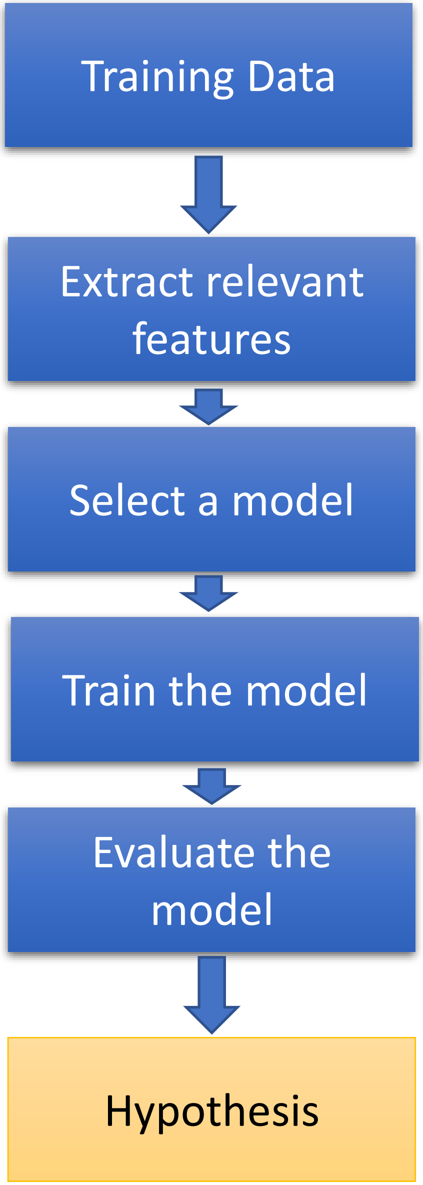
The Learning Phase

- The result of this learning phase is am algorithm, which represents the training data
- It should allow us to predict future outcomes from previously unseen data, e.g. tell us if the tennis court is likely to be used by somebody given the predicted or measured weather conditions



Tennis Dataset

ID	Outlook	Temp	Humidity	Windy	Play?
A	sunny	hot	high	false	no
B	sunny	hot	high	true	no
C	overcast	hot	high	false	yes
D	rainy	mild	high	false	yes
E	rainy	cool	normal	false	yes
F	rainy	cool	normal	true	no
G	overcast	cool	normal	true	yes
H	sunny	mild	high	false	no
I	sunny	cool	normal	false	yes
J	rainy	mild	normal	false	yes
K	sunny	mild	normal	true	yes
L	overcast	mild	high	true	yes
M	overcast	hot	normal	false	yes
N	rainy	mild	high	true	no



Training Data

Extract relevant features

Select a model

Train the model

Evaluate the model

Hypothesis

So what is the big deal, why is machine learning so difficult?

Why isn’t there a black box machine learning algorithm, that takes all the data available on the internet and is able to predict everything?

So what is the big deal, why is machine learning so difficult?

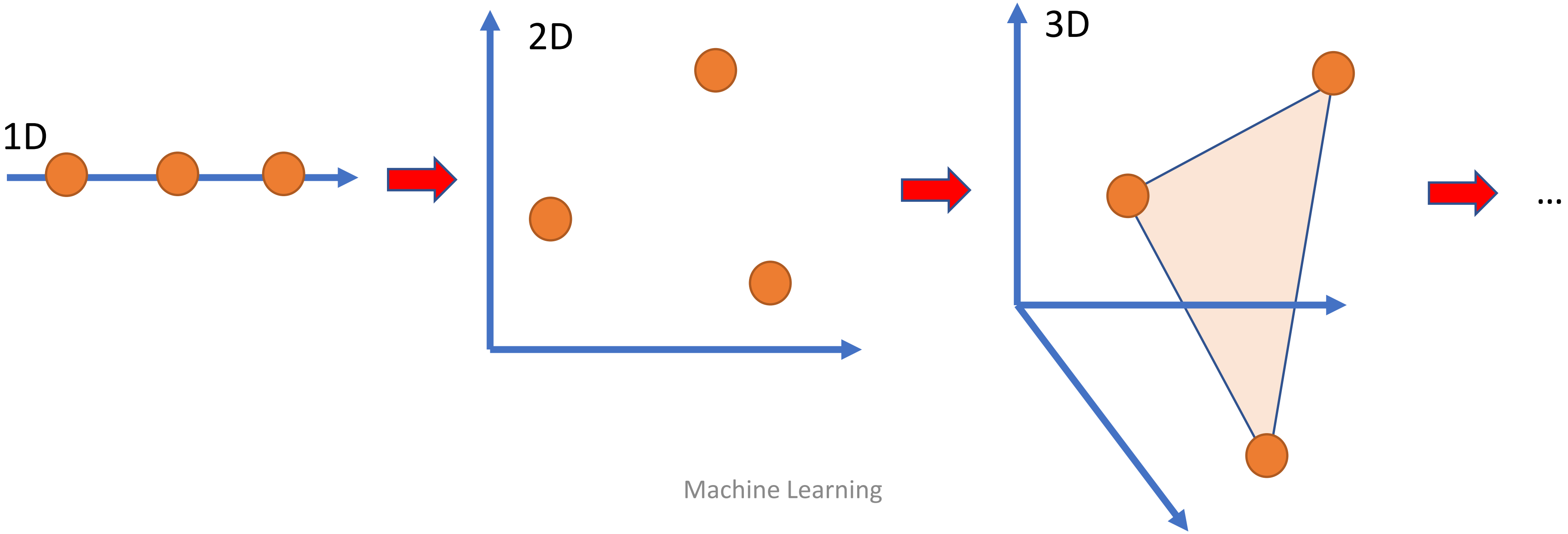
Why isn’t there a black box machine learning algorithm, that takes all the data available on the internet and is able to predict everything?

There are two major concerns that influence what can be achieved in any given application space by machine learning:

- High dimensional input data and feature spaces that lead to what is known as the “curse of dimensionality”
- Model complexity and generalisation performance

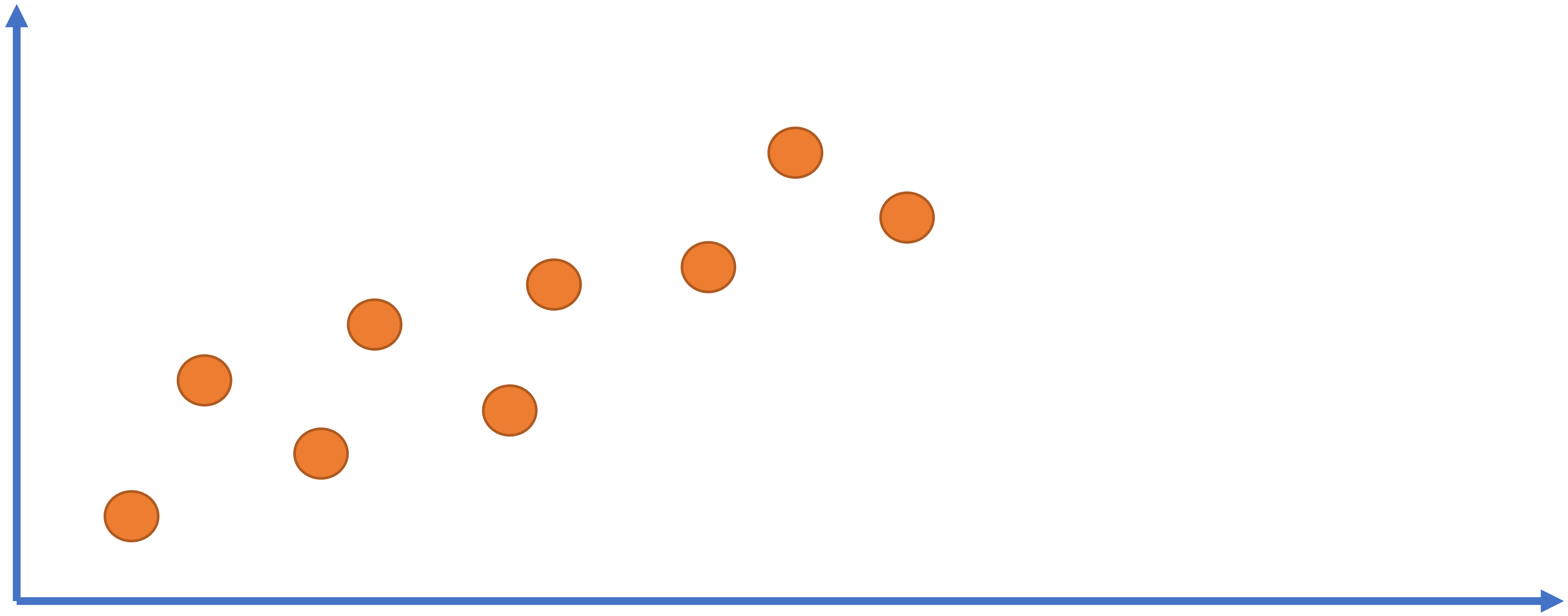
“Curse of dimensionality”

- Each additional feature add another dimension to the input space
- High dimensional spaces are empty, so the training data is representing less and less of the total space volume
- n Points span an n-1 dimensional subspace (2 points -> 1d line, 3 points -> 2d plane, 4 points -> 3d volume, etc.), so any additional dimension is redundant
- Euclidean distances in high dimensional spaces are problematic, almost all distances are almost equal making objects indistinguishable from each other



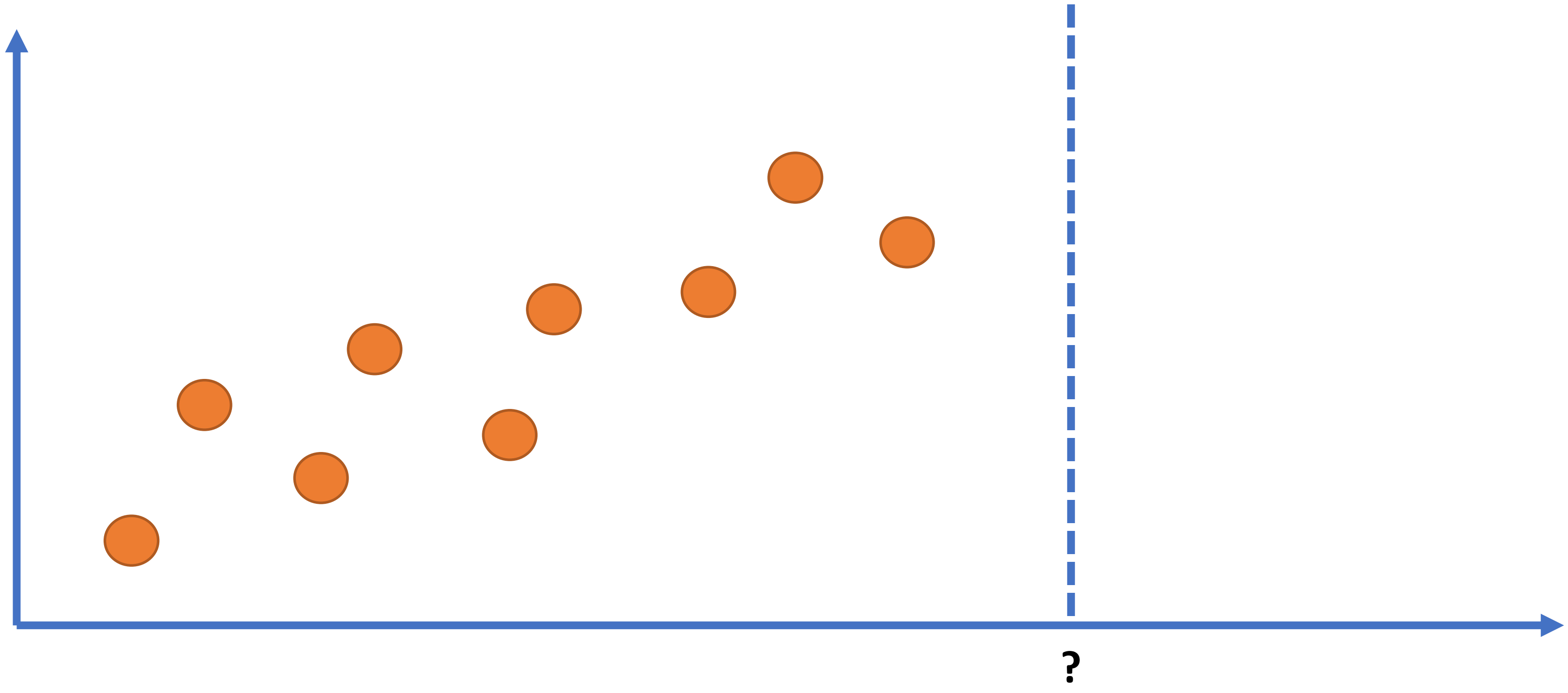
Model complexity and overfitting

- The choice of the “right” model complexity for a problem is crucial



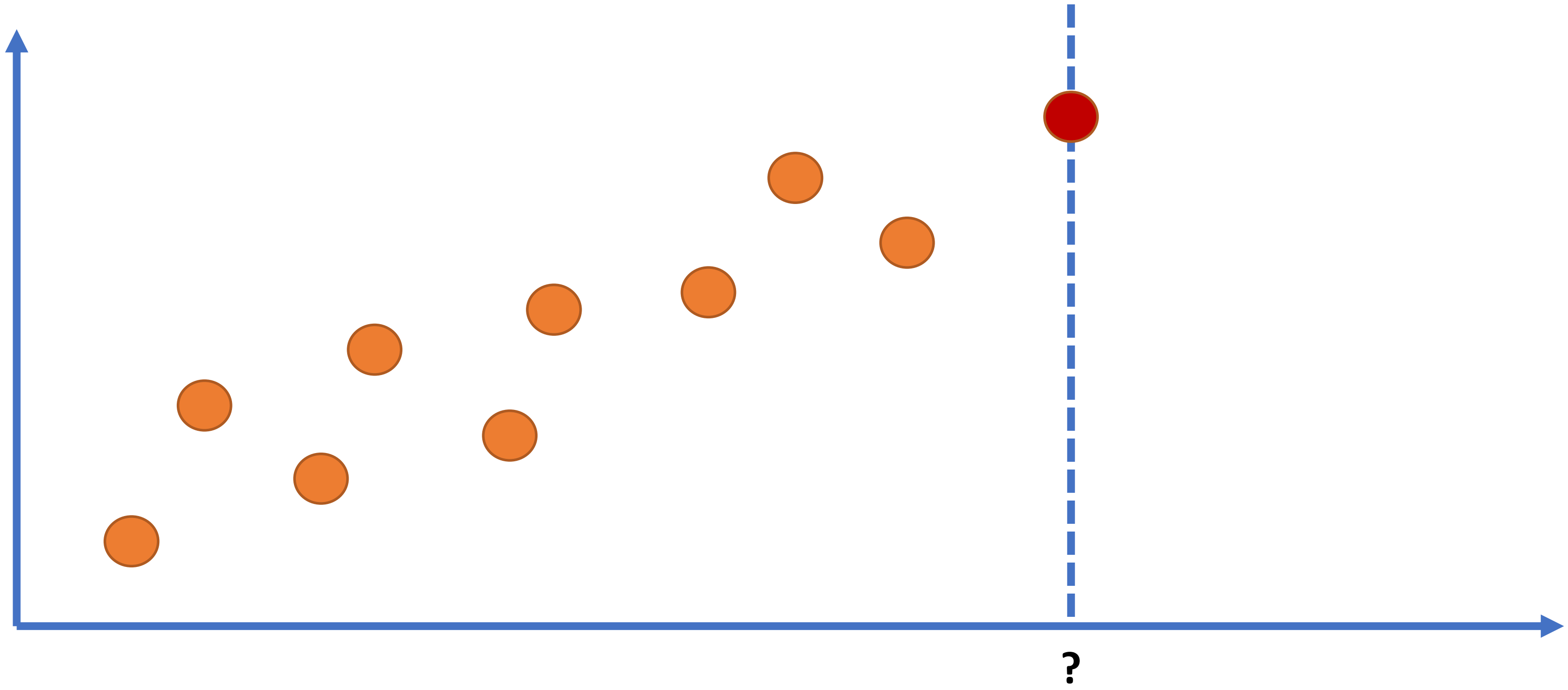
Model complexity and overfitting

- The choice of the “right” model complexity for a problem is crucial
- Where do we predict this new point?



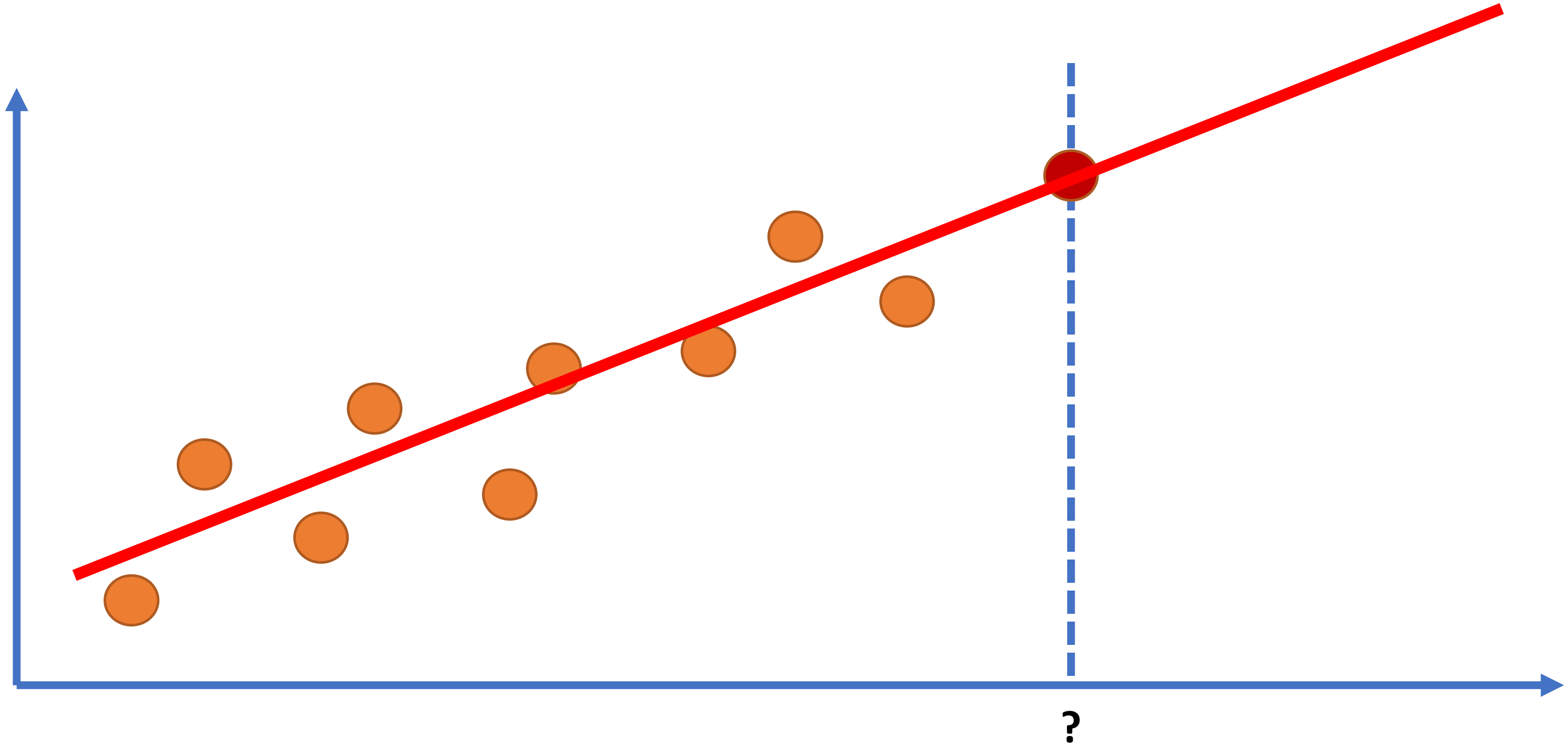
Model complexity and overfitting

- The choice of the “right” model complexity for a problem is crucial
- Where do we predict this new point? It is probably there, isn’t it?



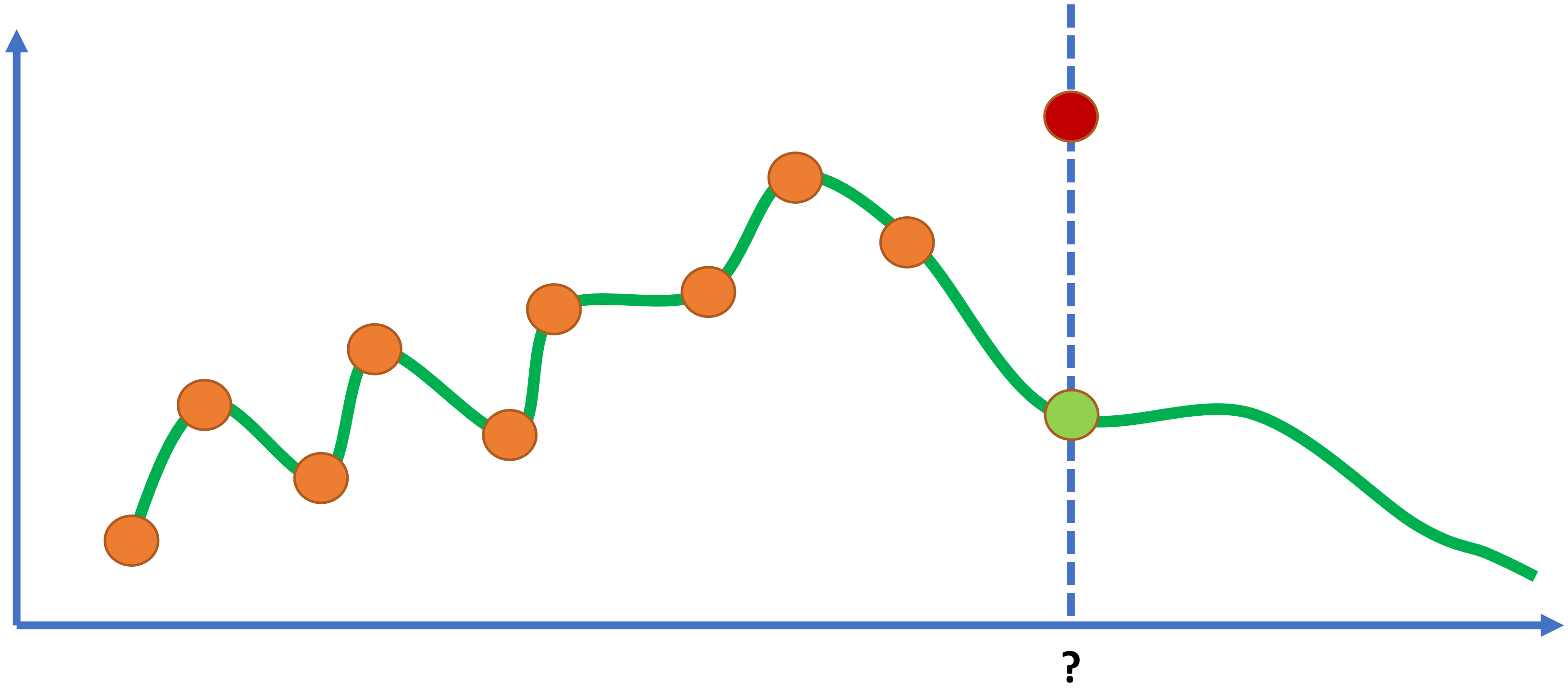
Model complexity and overfitting

- The choice of the “right” model complexity for a problem is crucial
- Where do we predict this new point? It is probably there, isn’t it?



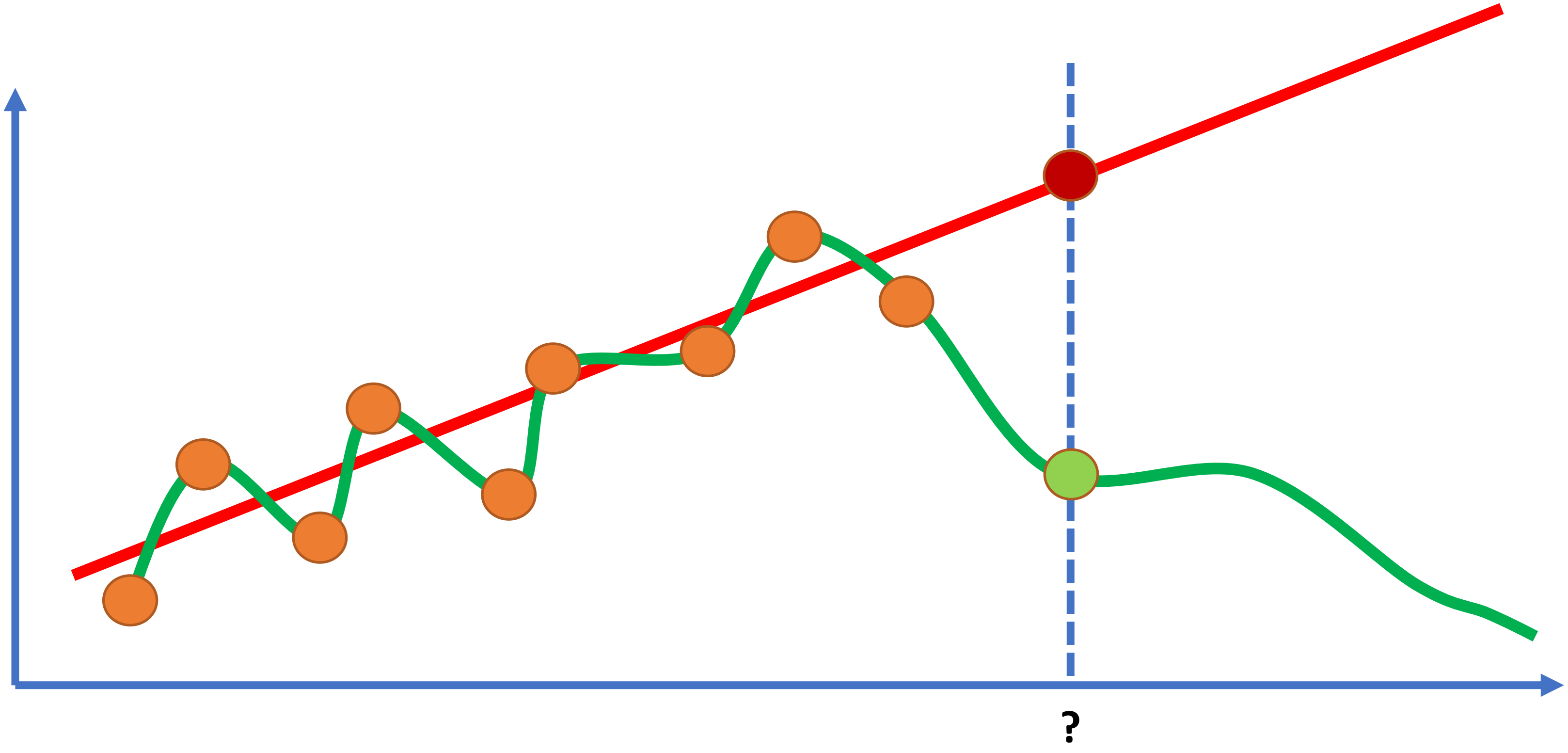
Model complexity and overfitting

- The choice of the “right” model complexity for a problem is crucial
- Where do we predict this new point? It is probably there, isn’t it?
- But why isn’t it there? The green curve fits all points of the training set perfectly.



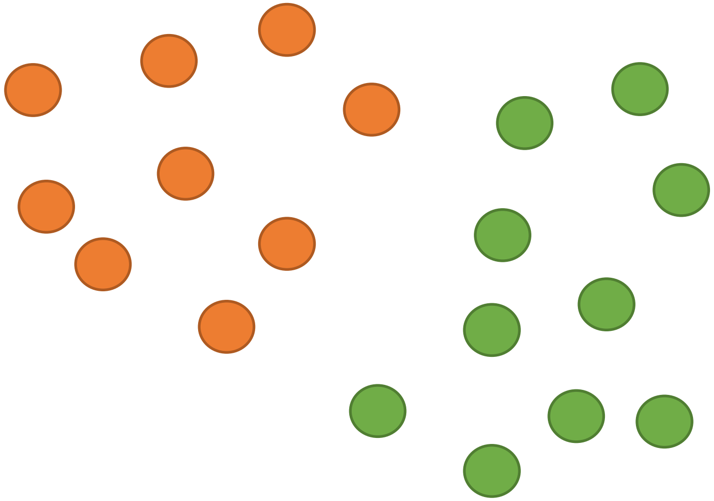
Model complexity and overfitting

- The choice of the “right” model complexity for a problem is crucial
- Where do we predict this new point? It is probably there, isn’t it?
- But why isn’t it there? The green curve fits all points of the training set perfectly. The red line doesn’t fit a single one, and yet it seems better!



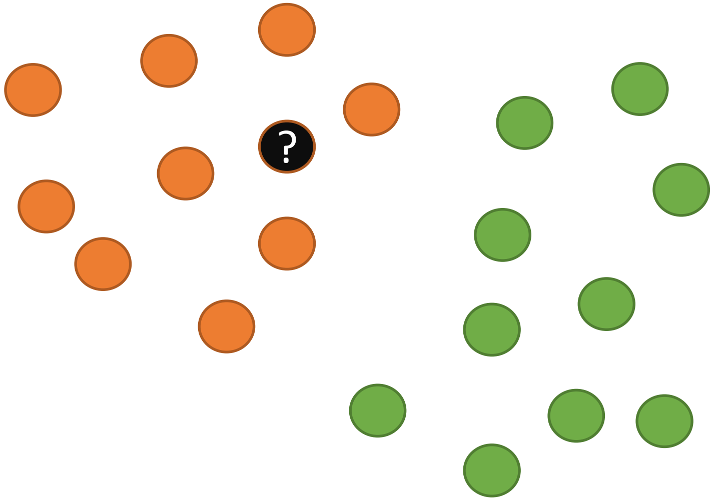
Model complexity and overfitting

- The choice of the "right" model complexity for a problem is crucial
- Overfitting is also an issue for classification problems



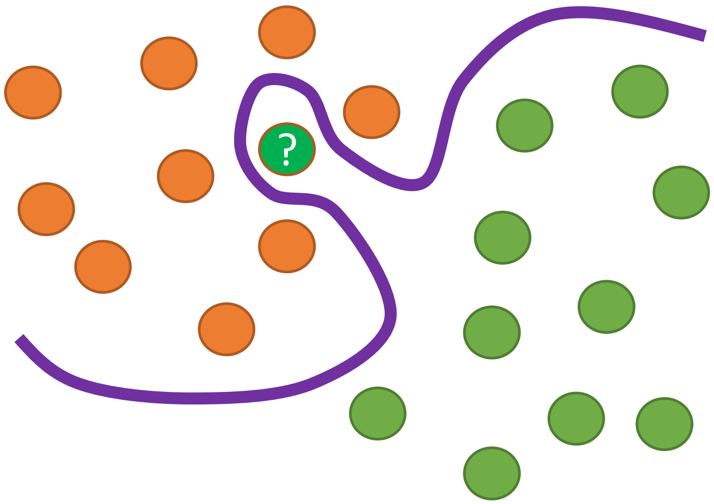
Model complexity and overfitting

- The choice of the "right" model complexity for a problem is crucial
- Overfitting is also an issue for classification problems
- Should this new point be coloured orange or green?



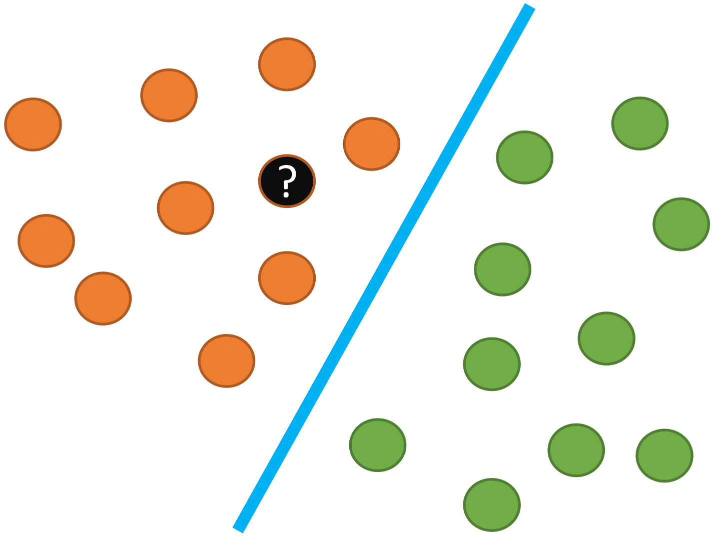
Model complexity and overfitting

- The choice of the “right” model complexity for a problem is crucial
- Overfitting is also an issue for classification problems
- Should this new point be coloured orange or green? Green, it is on the right of this valid decision surface.



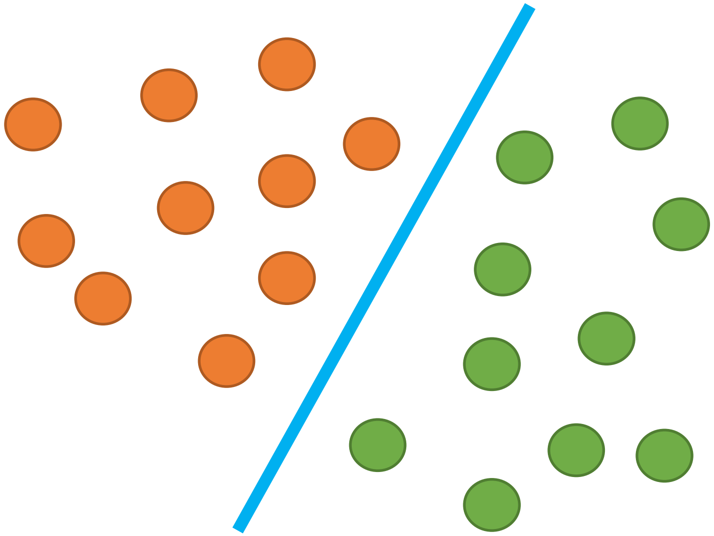
Model complexity and overfitting

- The choice of the “right” model complexity for a problem is crucial
- Overfitting is also an issue for classification problems
- Should this new point be coloured orange or green? Green, it is on the right of this valid decision surface. Or is it?



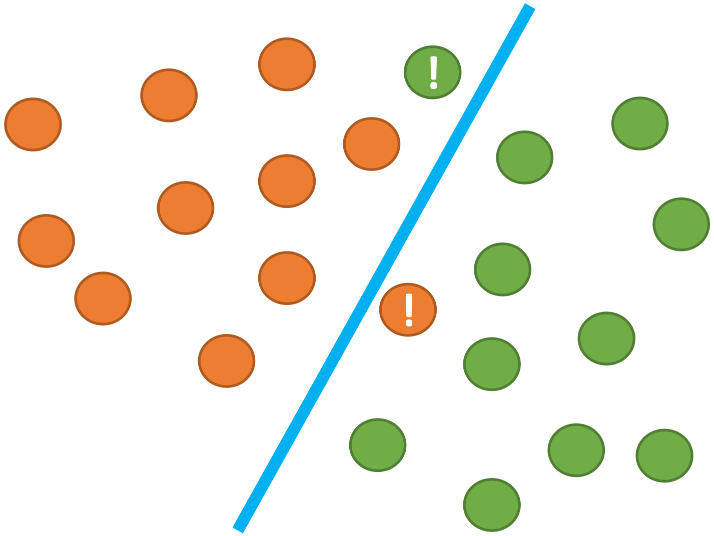
Model complexity and overfitting

- As a general rule, simpler models generalise better



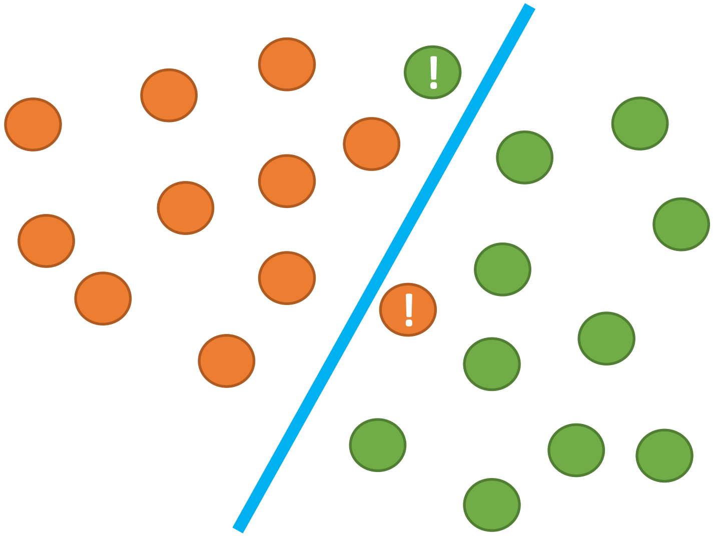
Model complexity and overfitting

- As a general rule, simpler models generalise better
- But what happens if the problem is more complex than the model allows?



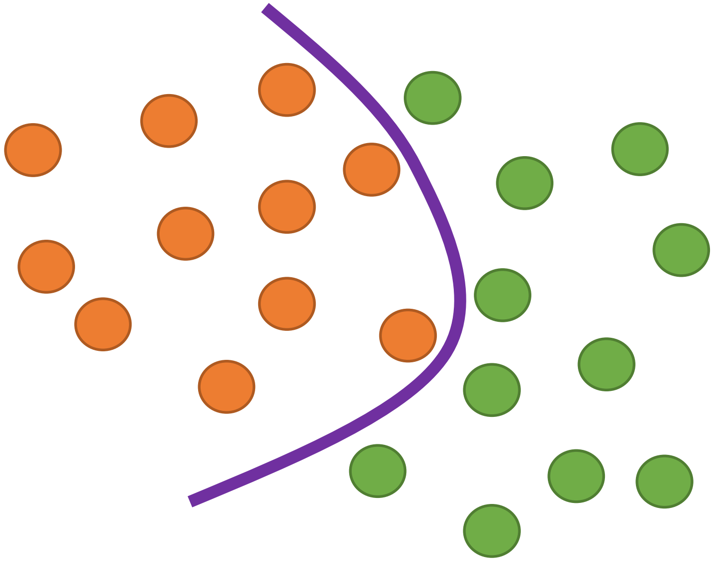
Model complexity and overfitting

- As a general rule, simpler models generalise better
- But what happens if the problem is more complex than the model allows?
- We either accept a small classification error



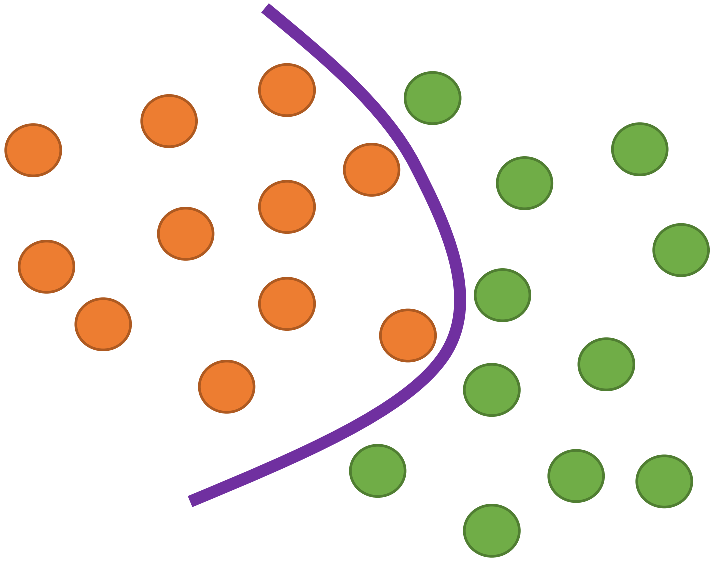
Model complexity and overfitting

- As a general rule, simpler models generalise better
- But what happens if the problem is more complex than the model allows?
- We either accept a small classification error, or we increase the model complexity as required



Model complexity and overfitting

- As a general rule, simpler models generalise better
- But what happens if the problem is more complex than the model allows?
- We either accept a small classification error, or we increase the model complexity as required



- The evaluation step in the learning phase is where these decisions can be analysed and justified

Thank you for your attention