

# Artificial Intelligence

Algorithms and Applications with Python

Chapter 06



Dr. Dominik Jung

[dominik.jung42@gmail.com](mailto:dominik.jung42@gmail.com)



python

# Outline

6

## Machine Learning - Fundamental Algorithms and Concepts

6.1

Machine Learning

6.2

Supervised Learning

6.3

Model Tuning, Combination and Selection

6.4

Unsupervised Learning

6.5

Reinforcement Learning (*forthcoming*)

6.6

Knowledge and Learning (*forthcoming*)

Lectorial 4: Building an AI Microservice for Credit Scoring (*forthcoming*)



Image source: [↗ Pixabay](#) (2019) / [↗ CC0](#)

### ► What we will learn:

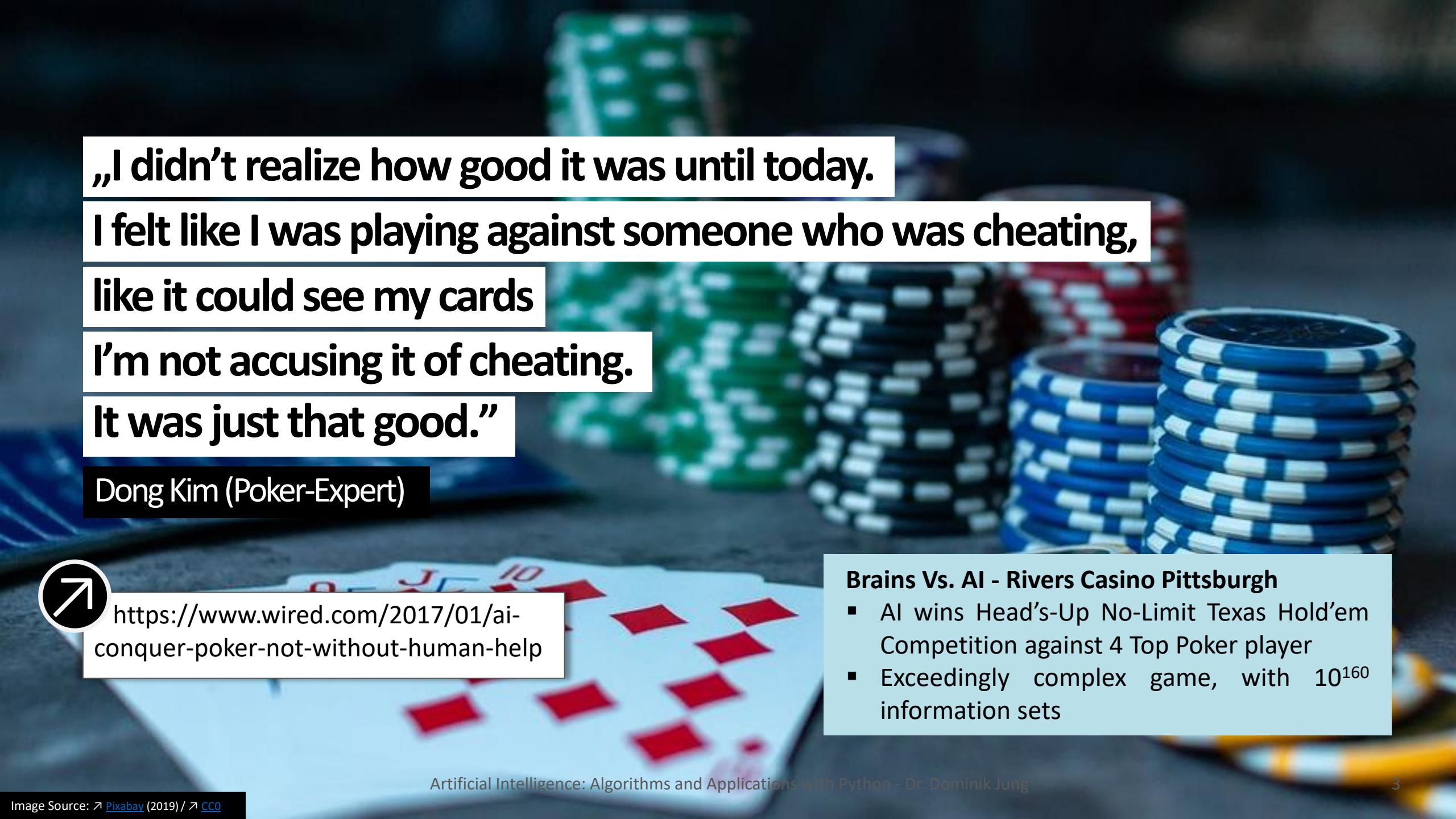
- General concepts of AI modelling and what types of problems match which models
- Get an intuition for which model approach fits best for a particular learning problem
- Know general problems of machine learning and how to optimize learning models

### ► Duration:

- 180 min

### ► Relevant for Exam:

- 6.1 – 6.4



**„I didn't realize how good it was until today.**

**I felt like I was playing against someone who was cheating,**

**like it could see my cards**

**I'm not accusing it of cheating.**

**It was just that good.”**

Dong Kim (Poker-Expert)

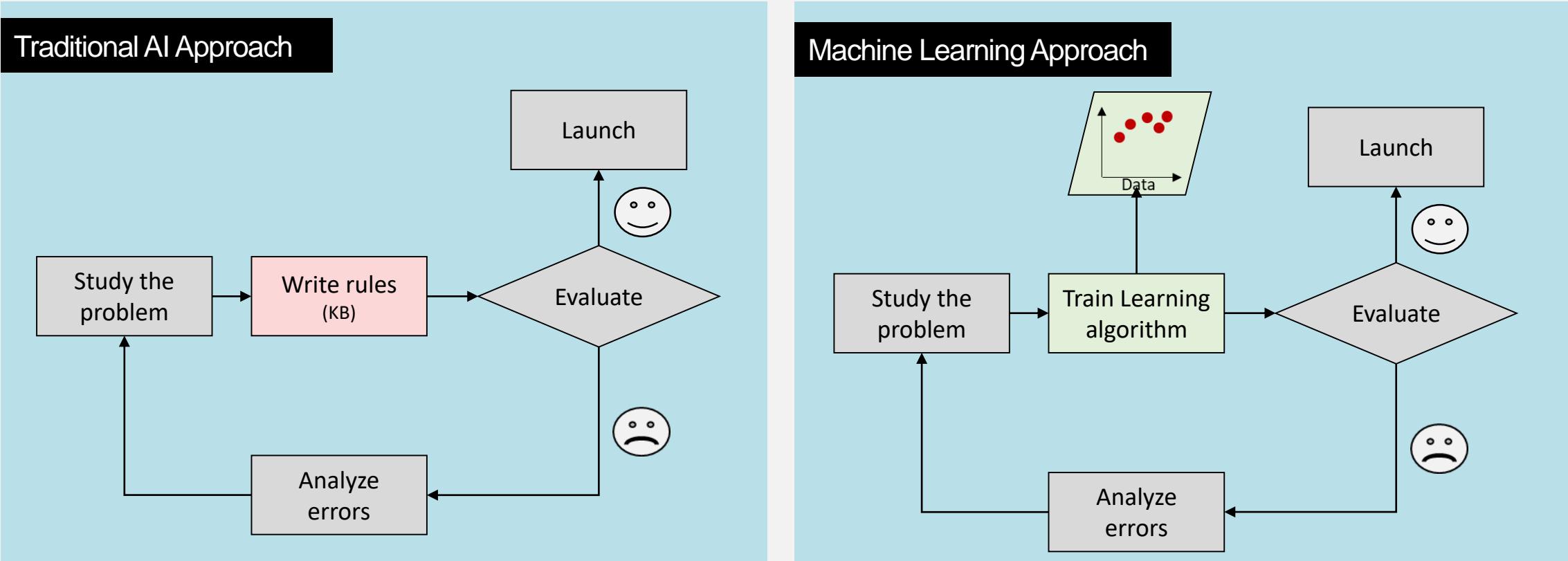


<https://www.wired.com/2017/01/ai-conquer-poker-not-without-human-help>

#### **Brains Vs. AI - Rivers Casino Pittsburgh**

- AI wins Head's-Up No-Limit Texas Hold'em Competition against 4 Top Poker player
- Exceedingly complex game, with  $10^{160}$  information sets

## 6.1 How was This Possible?



- Greater availability of data
- Increase of computational power



Trend away from rule-based and manually specified models to probabilistic data-driven modes

Adapted from Géron, A. (2017) | Image source: Géron, A. (2017)

**„Learning is any process  
by which a system improves  
performance from experience“**

Herber Simon



Image source : ↗ [Chess Programming Wiki](#) (2019)

## 6.1 Formalization of Machine Learning

**Input**

$X$

*Set of possible instances*

**Model**

$f: X \rightarrow Y$

*Unknown target function*

**Output**

$H = \{h \mid h: X \rightarrow Y\}$

*Set of function hypotheses*

**Goal:** Find hypothesis  $h \in H$  that best approximates target function  $f$

so that  $f(X) = \arg \max_y p(Y | X)$



Carefully! Do not get confused by the different meanings of *hypotheses* in science, statistics and machine learning!

**D**

### Machine Learning

A computer program is said to learn from **experience ‘E’**, with respect to some class of **tasks ‘T’** and **performance measure ‘P’** if its performance at tasks in ‘T’ as measured by ‘P’ improves with experience ‘E’. (Mitchell, 1997)

- **Task:** Play Poker
- **Performance:** Measure: Percentage of games/money won?
- **Experience:** Previous games

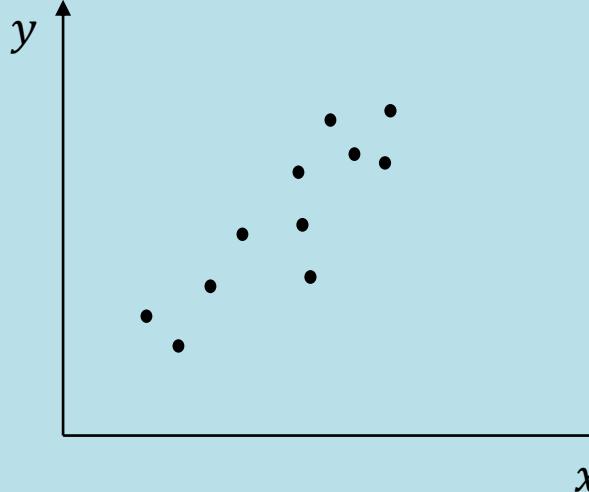


### Machine Learning

A computer program is said to learn from **experience 'E'**, with respect to some class of **tasks 'T'** and **performance measure 'P'** if its performance at tasks in 'T' as measured by 'P' improves with experience 'E'. (Mitchell, 1997)

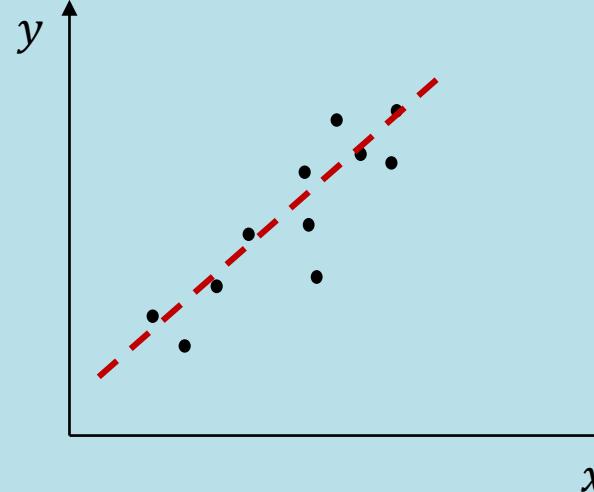
## 6.1 Geometric Interpretation of Machine Learning

1



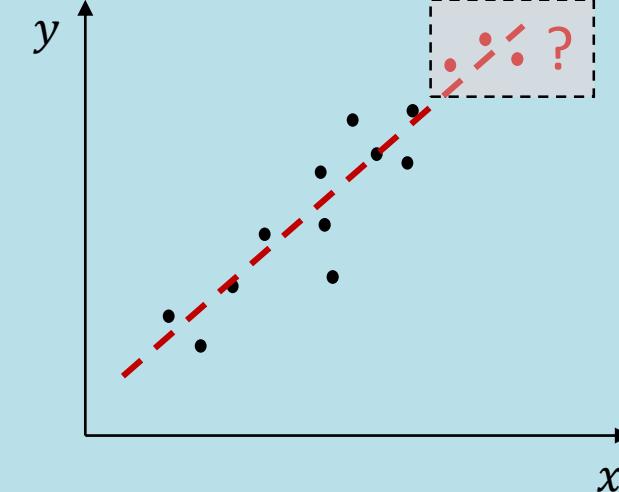
- You have given data  $X$  and want to build a model describes the relation between  $X$  and  $Y$ . Usually, you want to predict the target variable.

2



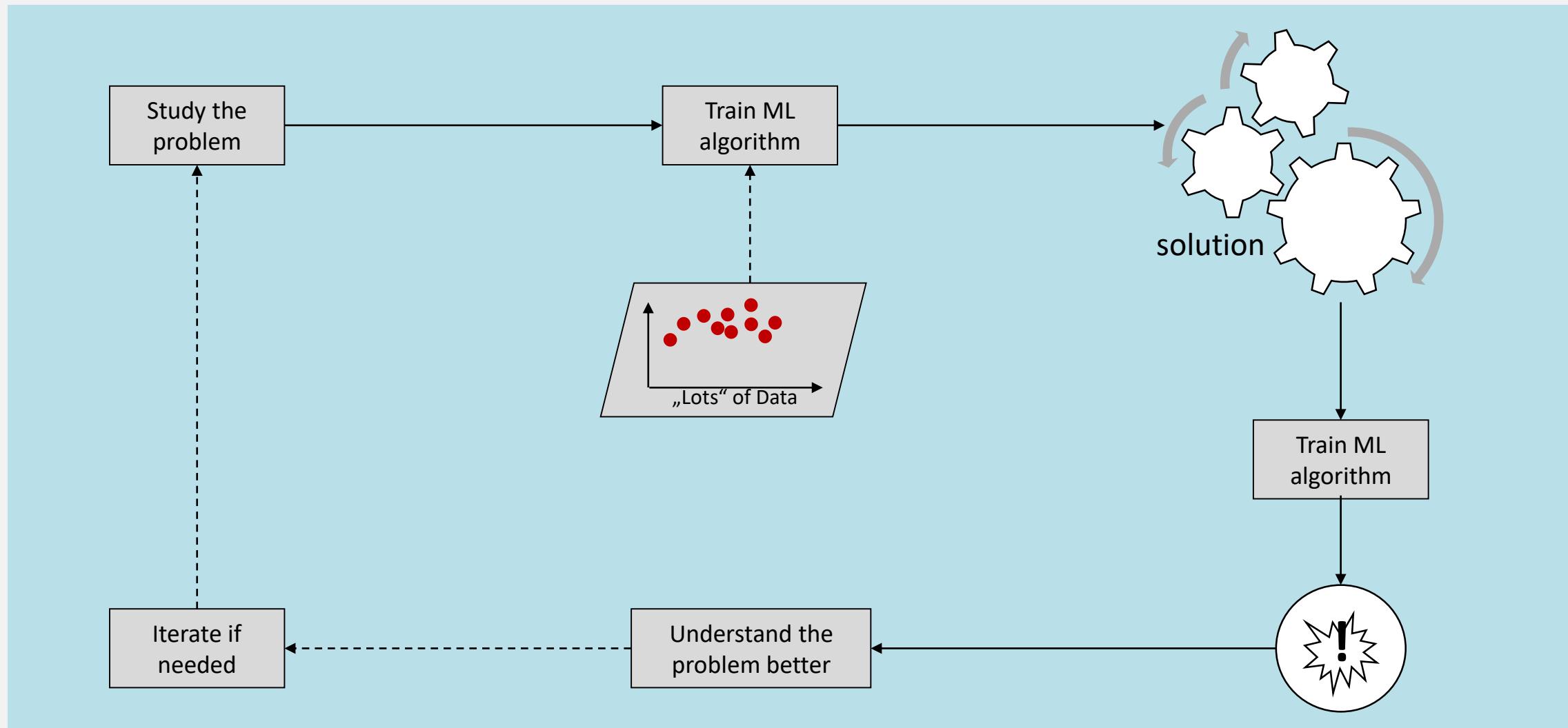
- You can model this problems as a regression problem. Hence you decide to use a linear regression model to fit your data

3



- You can use your model to predict future values of  $Y$

## 6.1 Many New Applications: Use Machine Learning for Data Mining



Adapted from Géron, A. (2017) | Image source: Géron, A. (2017)

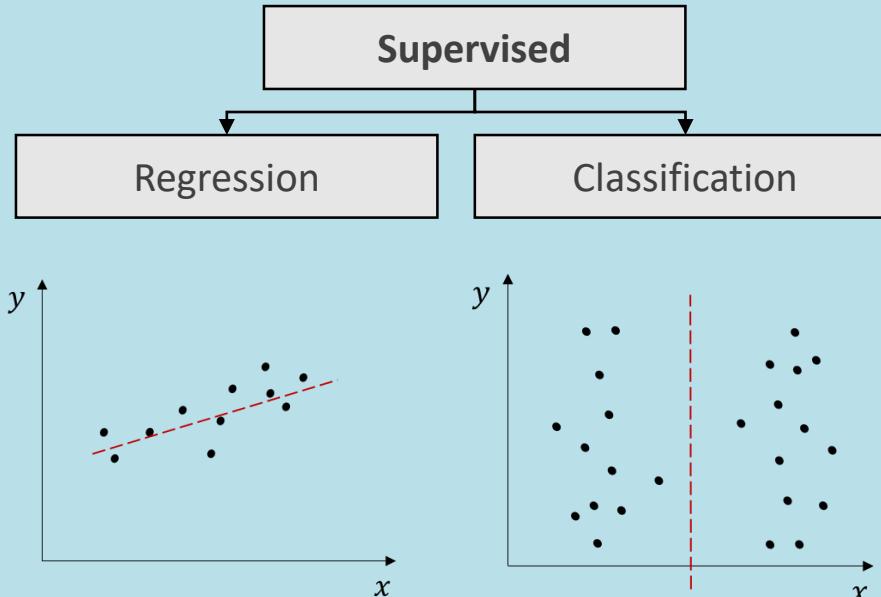
# 6.1 Taxonomy of Machine Learning Types (Simplification)



This taxonomy is just a very simplified representation of the topic machine learning. We will expand this taxonomy step by step in this lecture.

## Supervised Learning

An algorithm uses human-prepared training data to learn the relationship between given inputs and a given outcome.

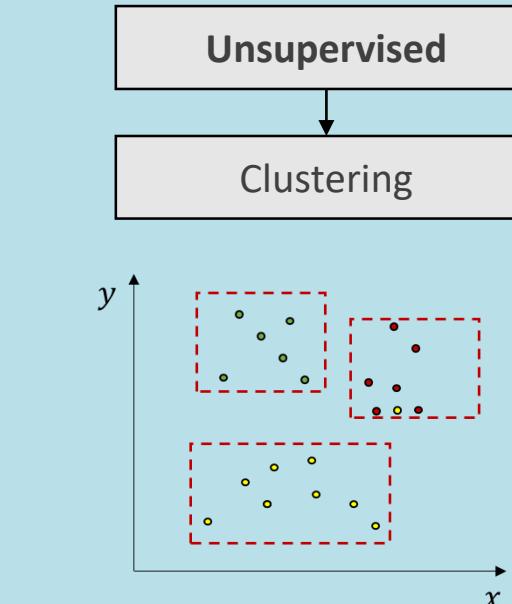


- Predict or explain differences between **continuous** variables

- Predict or explain differences between **categorical** variables

## Unsupervised Learning

An algorithm examines input data without knowing about attributes and possible results.

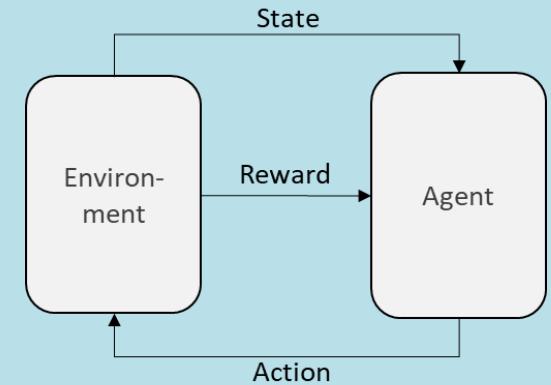


- Cluster observations into (distinct/different) groups

## Reinforcement Learning

An algorithm learns to perform a task by trying to maximize the rewards it receives for its actions.

### Reinforcement



- An AI system learns how to behave in a specific environment

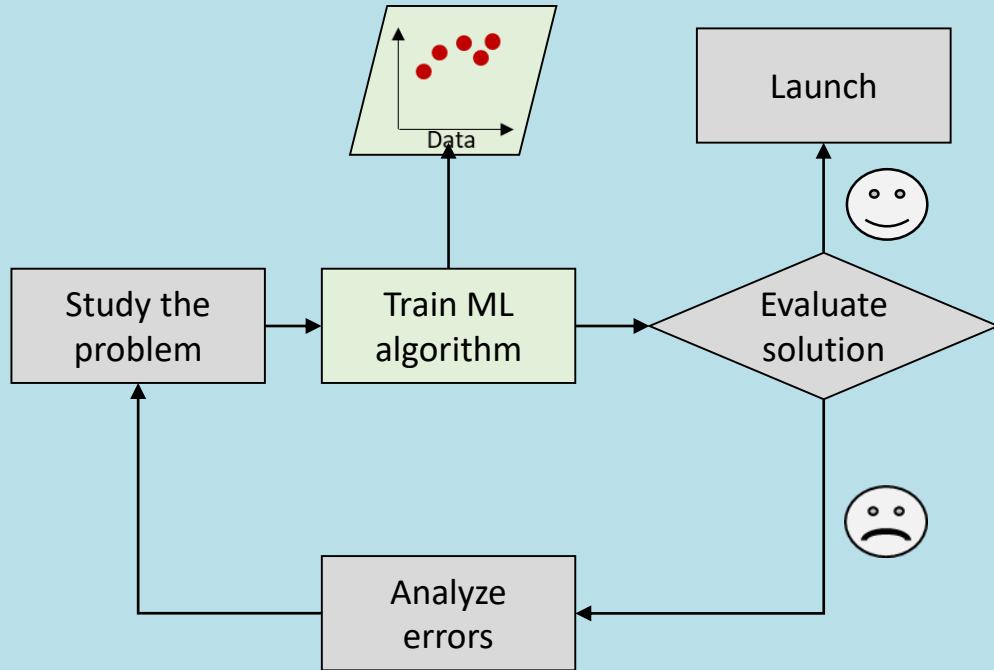
Adapted from Géron, A. (2017)

## 6.1 Other Approaches to Structure Machine Learning

- **Batch and online learning:** whether or not the system can learn incrementally from a stream of incoming data
- **Instance-based vs. model-based learning:** whether the system generalizes to new cases by comparing them to the learned examples or to build a model to make predictions

## 6.1 How to Improve Learning Performance

### Machine Learning Approach

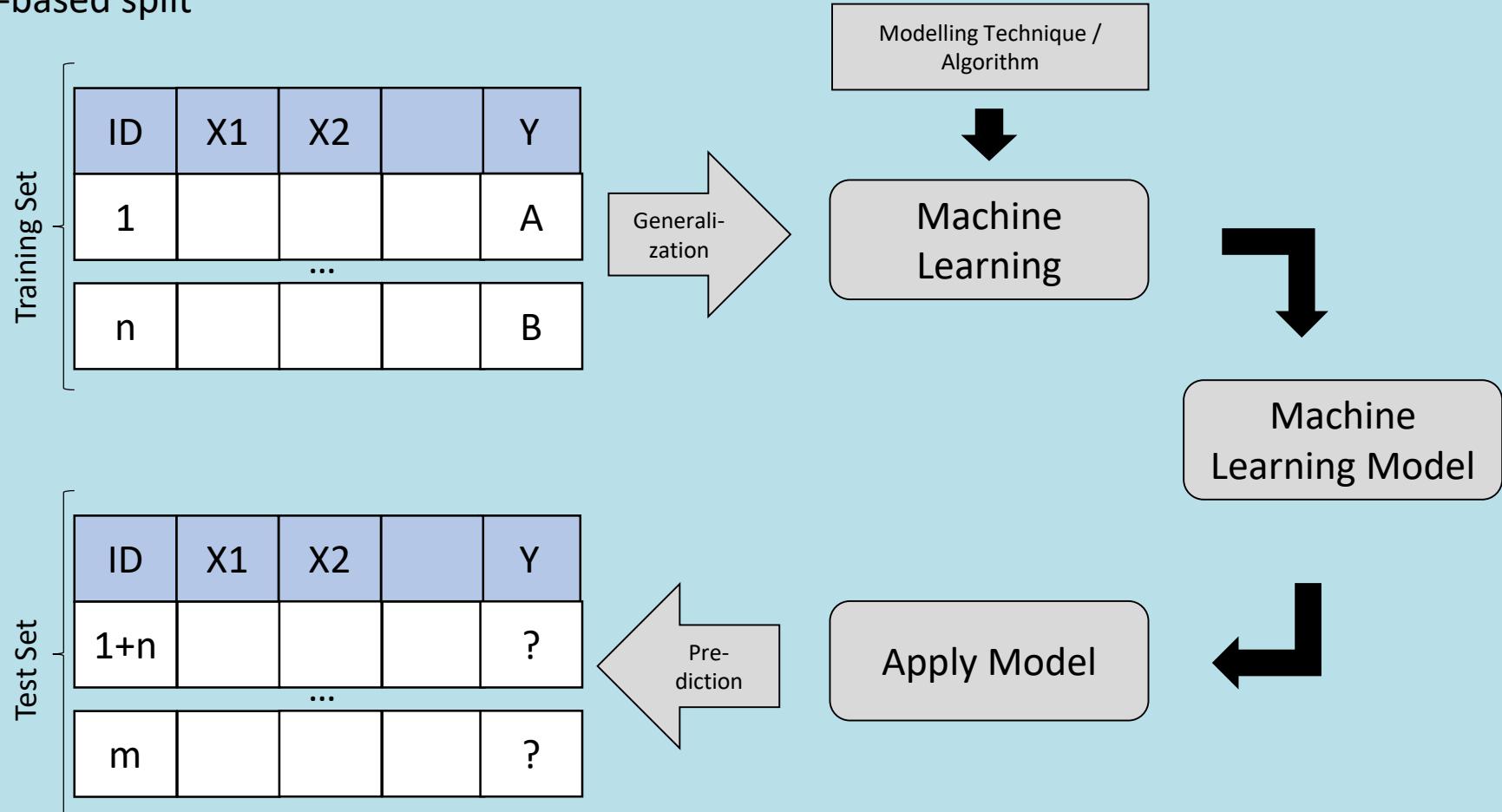


- How can we know how good our agent is learning the current problem?
- What is the correct way to compute our performance measure?
- **Naive approach:** Learning a machine learning model and testing it on the same data
- **Problem:** We measure how good the model is to predict the classes of the test data and not how good it generalizes (the model learnt to solve such problems)

Adapted from Géron, A. (2017)

## 6.1 Split Data into Test and Train to Measure Performance

Simple ID-based split



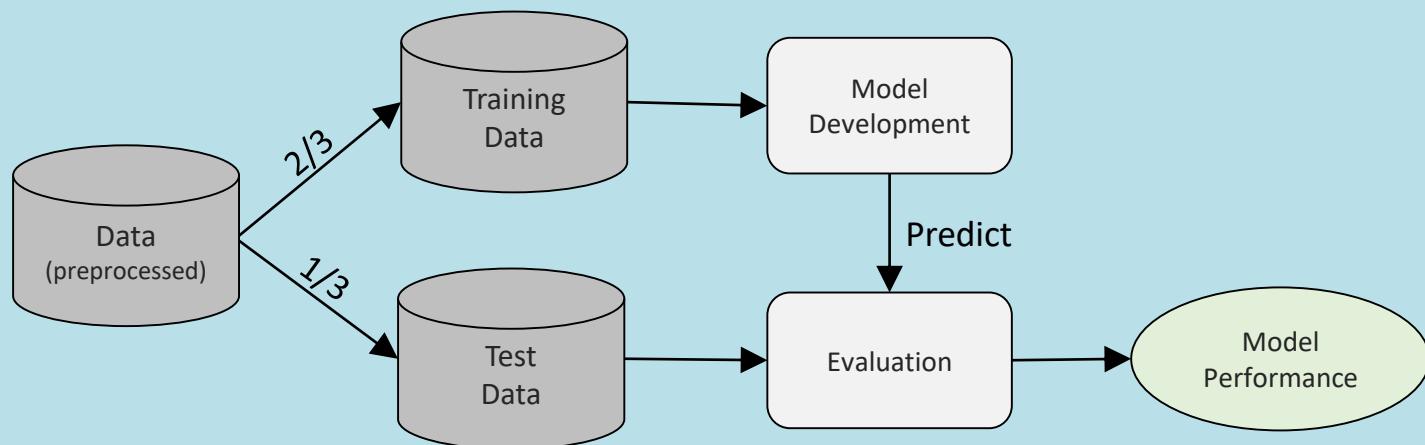
Adapted from Scikit-learn ([www.scikit-learn.org](http://www.scikit-learn.org)); Géron, A. (2017)

## 6.1 Holdout-Method

### Procedure

- Split your labelled data into train and test set, Build your model based on train set, and measure the performance based on the test set
- In problems where we have a sparse dataset we may not be able to afford the “luxury” of setting aside a portion of the dataset for testing
- Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an “unfortunate” split

### Visualization



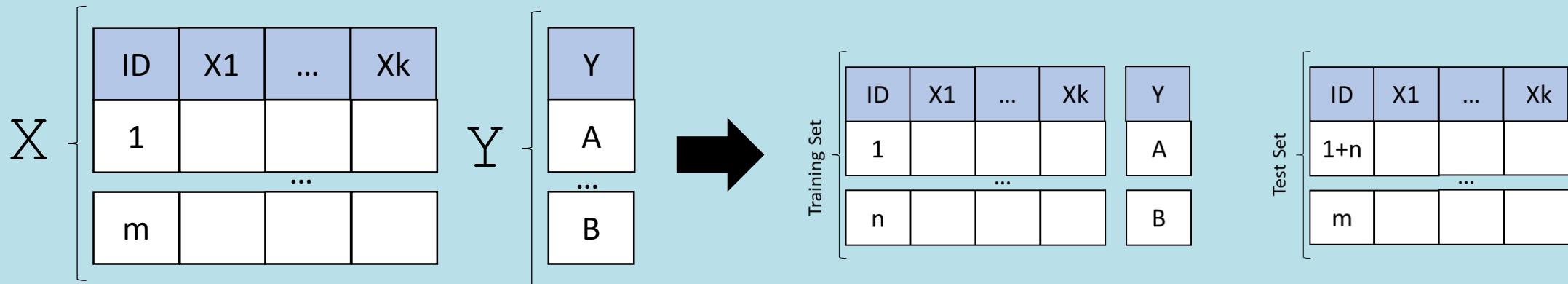
Adapted from Kim, J.-H. (2009); Géron, A. (2017)

## 6.1 Split your Data into train and test Set with Python

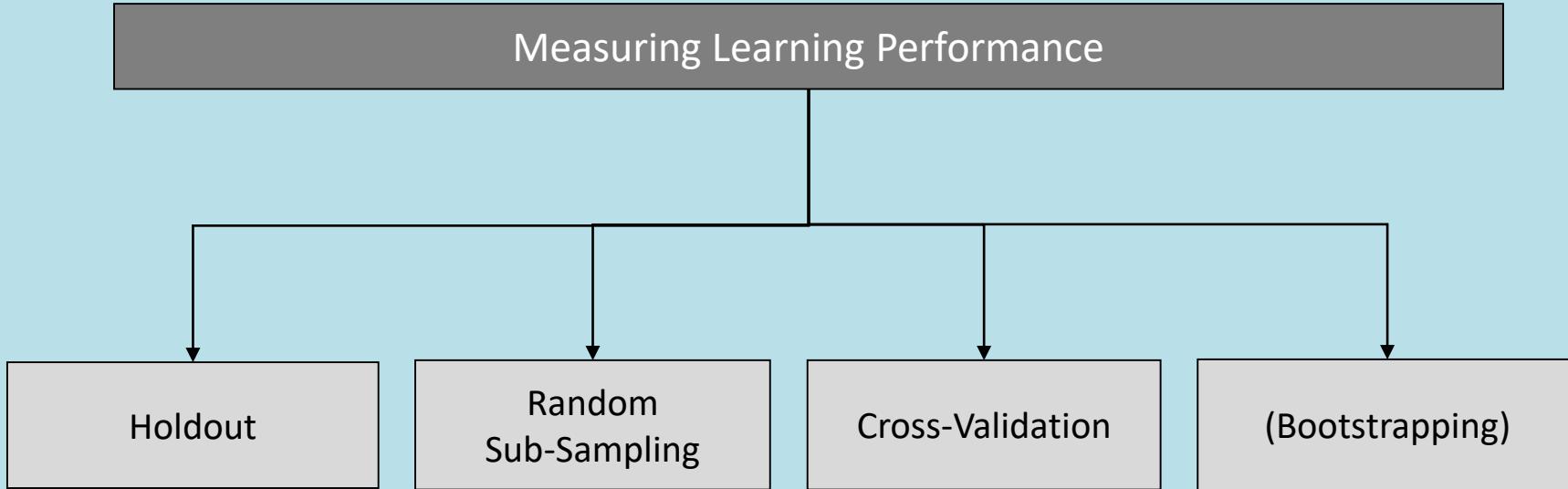
- In scikit-learn a random split into training and test sets can be quickly computed with the `train_test_split` helper function.

```
import numpy as np
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.4, random_state=0)
```



## 6.1 Splitting Paradigms for Measuring Learning Performance



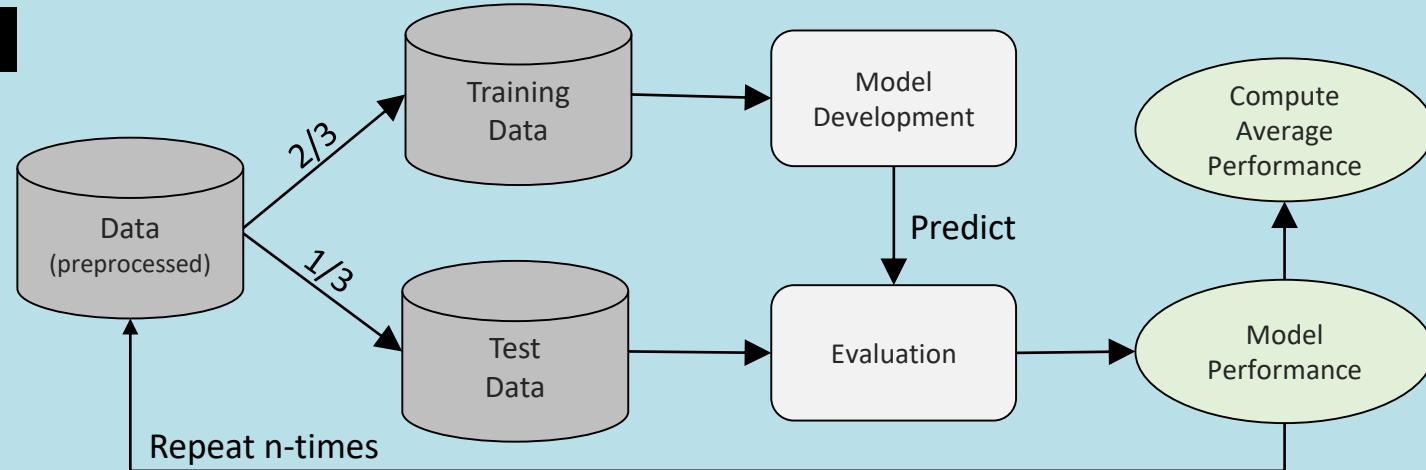
Adapted from Géron, A. (2017)

# 6.1 Random Sub-Sampling

## Procedure

- Repeated holdout with different samples, but measure the average performance
- Multiple models, where you can choose from
- Same disadvantages then simple holdout
- No control how often an observation is used for model building

## Visualization



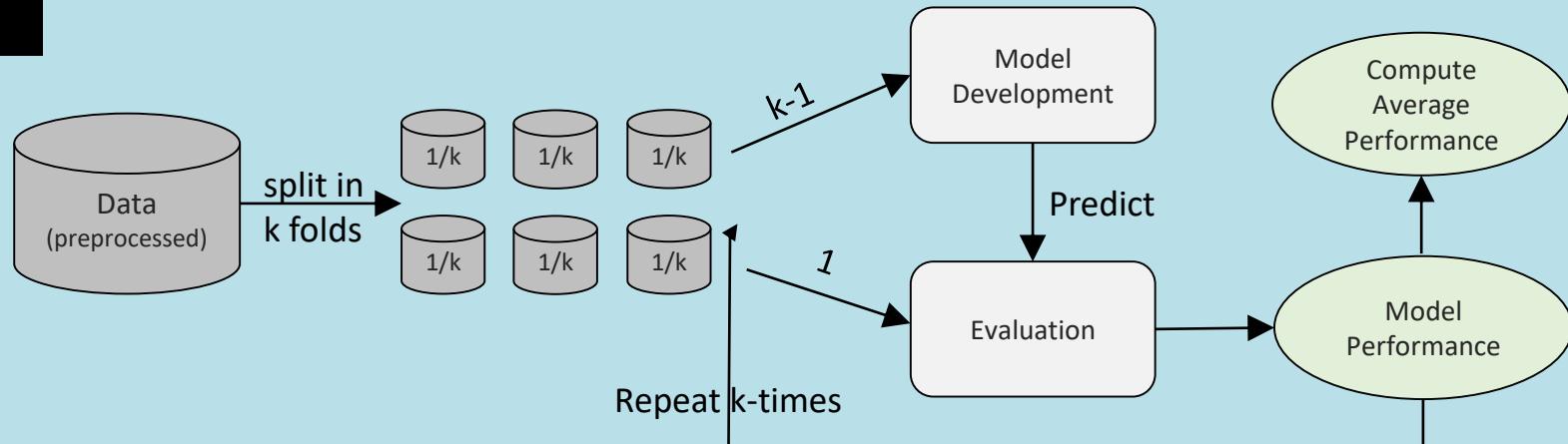
Adapted from Kohavi, R. (1995); Géron, A. (2017)

# 6.1 K-fold Crossvalidation

## Procedure

- Split data into  $k$  same-sized samples, use  $k-1$  samples for training, and 1 sample set for testing. Each observation is used the same time for training
- Benefit is that it uses as many model building examples as possible and test sets disjunct
- High complexity of the process ( $k$ -runs), reliability of the performance statement is weakened, since these statements are derived from only one example.

## Visualization



Adapted from Kim, J.-H. (2009); Géron, A. (2017)

## 6.1 K-fold Crossvalidation with Python

### Crossvalidation()

```
cross_validate(estimator, X, y, cv)
```

### Parameters

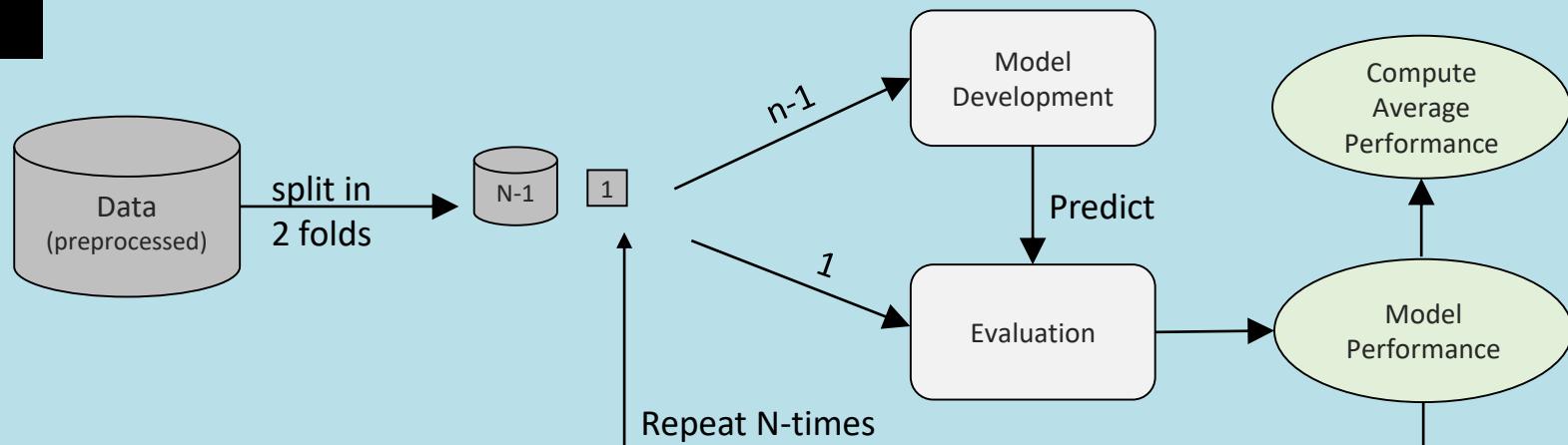
estimator	The object to use to fit the data.
X	The data to fit. Can be for example a list, or an array.
y	The target variable to try to predict in the case of supervised learning.
cv	Determines the cross-validation splitting strategy. Possible inputs for cv are: <ul style="list-style-type: none"><li>▪ None, to use the default 5-fold cross validation,</li><li>▪ int, to specify the number of folds in a (Stratified)KFold,</li><li>▪ CV splitter,</li><li>▪ An iterable yielding (train, test) splits as arrays of indices.</li></ul>

## 6.1 Leave-one Out Cross-Validation

### Procedure

- K-fold cross validation taken to its logical extreme, with K equal to the number of data points in the set (N)
- the model is trained on all the data except for one point and a prediction is made for that point
- As before the average error is computed and used to evaluate the model

### Visualization



Adapted from Kim, J.-H. (2009); Géron, A. (2017)

## 6.1 Leave-one Out Cross-Validation with Python

- In scikit-learn we can use the `LeaveOneOut` helper function to perform leave-one-out crossvalidation.

```
>>> from sklearn.model_selection import LeaveOneOut
>>> X = [1, 2, 3, 4]
>>> loo = LeaveOneOut()
>>> for train, test in loo.split(X):
    ("%s %s" % (train, test))

[1 2 3] [0]
[0 2 3] [1]
[0 1 3] [2]
[0 1 2] [3]
```

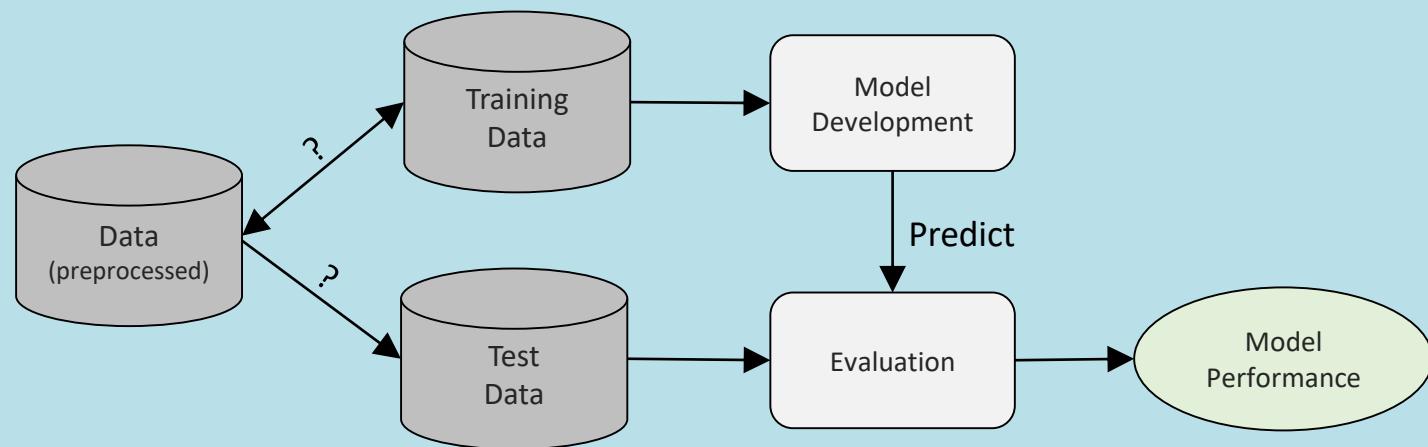
# 6.1 Bootstrapping

Kohavi (1995)

## Procedure

- In the previous procedures, an example was considered several times as a training example (in the same cycle). Here a training set is generated by random selection from the entire set of classified examples (Sampling with replacement).
- Perform sampling with replacement on your original dataset, but use the data points that have not been chosen as the test dataset. Repeat this procedure several times and compute the average score as estimation of your model performance.

## Visualization



Adapted from Kohavi, R. (1995); Géron, A. (2017)

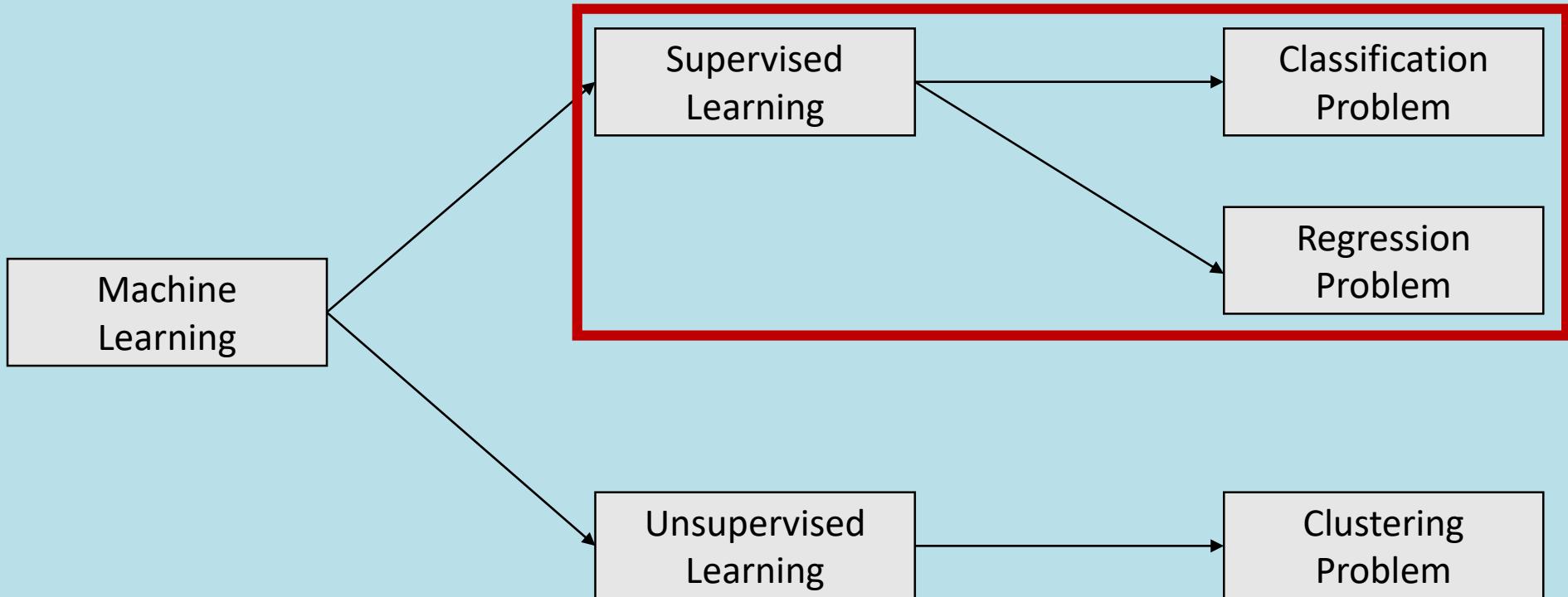
# Your turn!

### Task

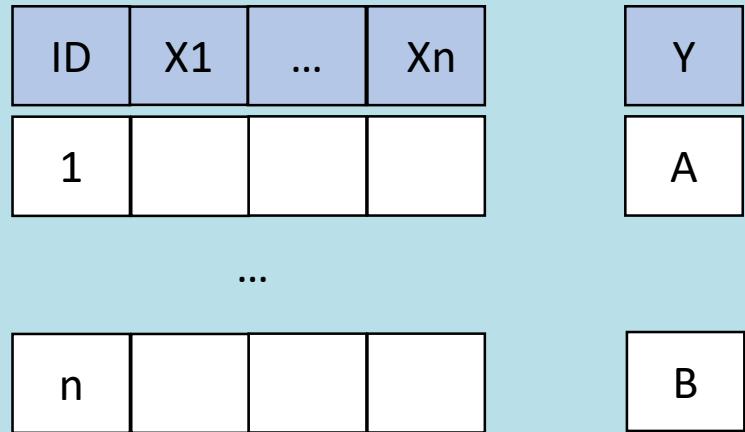
Identify the *task*, *performance measure* and *experience* of the following machine learning problems. Please discuss your results with your neighbor!

- Build a chess-bot to win an online chess competition
- Build an information system that recognizes hand written addresses
- Build an automated spam-mail system for your email server

## 6.2 Problem Types in Machine Learning (High-Level)



## 6.2 Classification Problem and Category Learning



$$f: X \rightarrow Y$$

Assign to one of a given set of finite classes

Common real-life tasks:

- Fraud detection: Credit card applications or transactions
- Spam filtering in email
- Recommend news articles books, movies, music etc.
- Handwritten letters
- Astronomical images

## 6.2 Formalization of Category Learning

- **Task:** Predict if customer is credit worthy
- **Instance:** <price, maintenance, number\_doors>
- **Class** = {positive, negative}

ID	price	maintenance	number_doors	class
1	small	cheap	3	buy
2	high	expensive	5 or more	no buy
3	small	medium	4	buy
...	...	...	...	...
n	medium	expensive	2	no buy

- $\text{price} \in \{\text{small, medium, high}\}$
- $\text{maintenance} \in \{\text{cheap, medium, expensive}\}$
- $\text{number\_doors} \in \{2,3,4,5 \text{ more}\}$

Adapted from Hunt et al. (1966); Russel, S., & Norvig, P. (2016)

## 6.2 Hypothesis Selection

- Possible hypotheses are consistent with the training data set

ID	price	maintenance	number_doors	class
1	small	cheap	3	buy
2	high	expensive	5 or more	no buy
3	small	medium	4	buy
...	...	...	...	...
n	medium	expensive	2	no buy



Do you have any ideas of possible hypotheses based on this training data?

$$H = \{h \mid h: X \rightarrow Y\}$$

1	<i>price = small</i>
2	<i>maintenance = cheap</i> $\cup$ <i>maintenance = medium</i>
3	<i>number of doors = 3</i> $\cup$ <i>number of doors = 2</i>
4	<i>price = small</i> $\cap$ <i>maintenance = cheap</i>

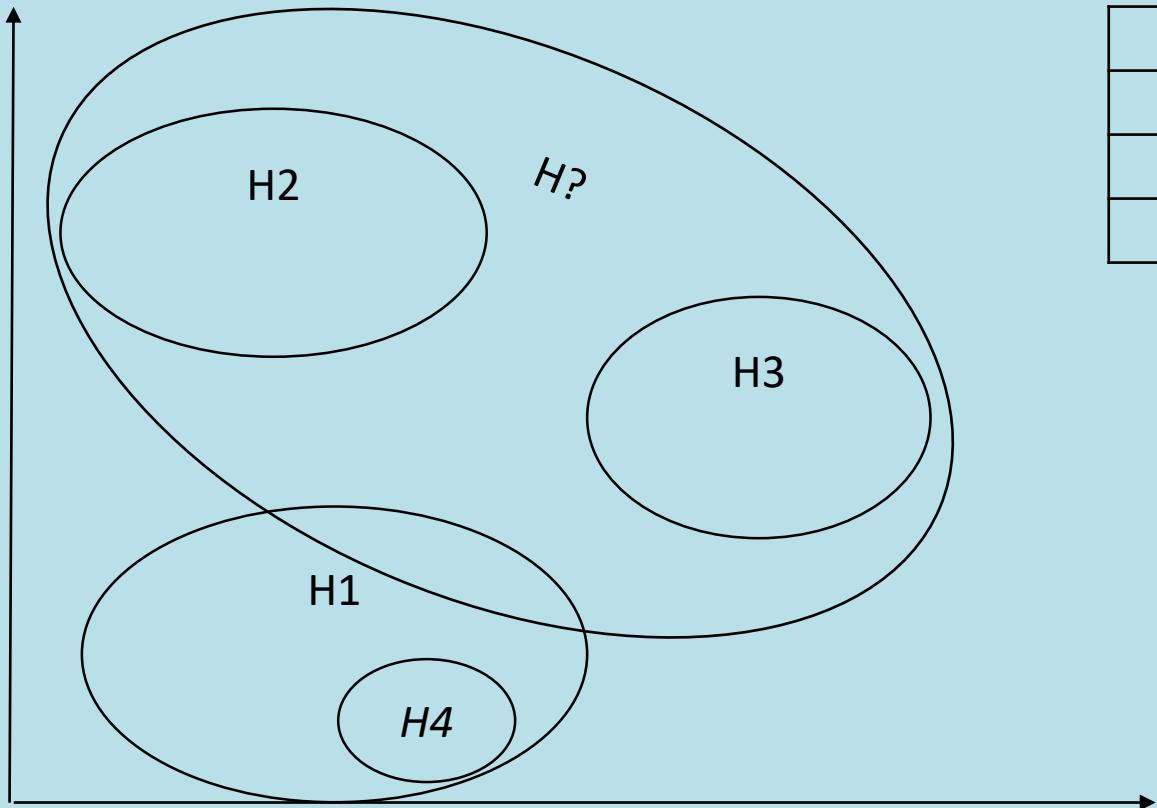


Some older, psychological papers term these kind of hypotheses “concepts”. See e.g. Bruner, J. S., & Austin, G. A. (1986)

Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

## 6.2 Hypothesis Space

$$H = \{h \mid h: X \rightarrow Y\}$$



1	$price = small$
2	$maintenance = cheap \cup maintenance = medium$
3	$number\ of\ doors = 3 \cup number\ of\ doors = 2$
4	$price = small \cap maintenance = cheap$

Adapted from Mitchell, T. M. (1997); Russel, S., & Norvig, P. (2016)

## 6.2 “Interpretation” of the Hypothesis Space

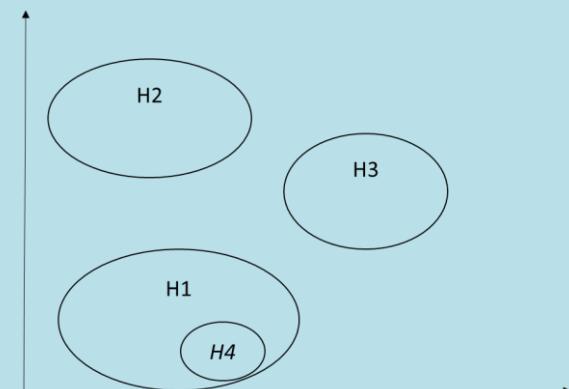
- For learning concepts on instances described by  $n$  discrete-valued features, consider the space of conjunctive hypotheses represented by a vector of  $n$  constraints

$< c_1, c_2, \dots, c_n >$  where each  $c_i$  is either:

- ?, a wild card indicating no constraint on the  $i$ th feature
- A specific value from the domain of the  $i$ th feature
- $\emptyset$  indicating no value is acceptable

Sample conjunctive hypotheses:

- $< \text{price} = \text{small}, \text{doors} = 3, ? >$
- $< ?, ?, ? >$  (most general hypothesis)
- $< \emptyset, \emptyset, \emptyset >$  (most specific hypothesis)



Adapted from Mitchell, T. M. (1997); Russel, S., & Norvig, P. (2016)

## 6.2 Learning and Generalization

- There are many possible hypotheses. Our purpose is to find the hypothesis that is able to predict the training data correctly AND new data correctly

ID	price	maintenance	number_doors	class
1	small	cheap	3	buy
42	small	expensive	4	no buy



Our current best hypothesis  
*"price = small"* can not handle it

- Hypotheses must generalize to correctly classify instances not in the training data
- Simply memorizing training examples is a consistent hypothesis, but it does not generalize.

## 6.2 Inductive Learning Assumption

- **Assumption:** If any function approximates the target concept well on a sufficiently large set of training examples, it also will approximate the target function well on unobserved examples
- We assume that the training and test examples are drawn independently from the same underlying distribution

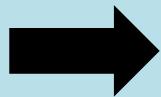
Adapted from Mitchell, T. M. (1997); Russel, S., & Norvig, P. (2016)

## 6.2 Naive Assumption: Category Learning as Search Problem?

- **Intuitive idea:** Category learning model as search problem (see lecture 2)
- **Problem:** All reasonable hypothesis spaces are too large (or even infinite)

ID	Feature
1	{true, false}
n	{true, false}

Instance space with n binary  
{true, false} features



Target hypothesis can be any  
of the  $2^{2^n}$  features

## 6.2 Learning by Enumeration

- **Other idea:** Category learning model as search problem (see lecture 2)

**Algorithm:** Learning by Enumeration

***persistent:***

*H, hypothesis space with h hypotheses*

*D, training data*

***For each h in H :***

*If h is consistent with D*

*return h*

- This algorithm is guaranteed to terminate with a consistent hypothesis if one exists; however, it is obviously computationally intractable for almost any practical problem.

## 6.2 More Efficient Learning?

- Is there a way to learn conjunctive concepts more efficient to solve our classification problems?
- How do humans learn hypotheses about the world (conjunctive concepts)?

Adapted from Mitchell, T. M. (1997); Russel, S., & Norvig, P. (2016)

## 6.2 Concept Learning System (CLS)

► Built a model to predict or classify future observations by divide- and-conquer based on a set of decision rules

- Most decision trees and rule learning algorithms are based on Hunt's fundamental algorithm of concept learning
- The algorithm embodies a method used by humans when learning simple concepts, namely finding key distinguishing features between two categories, represented by positive and negative (Training) examples
- Based on a divide and conquer strategy

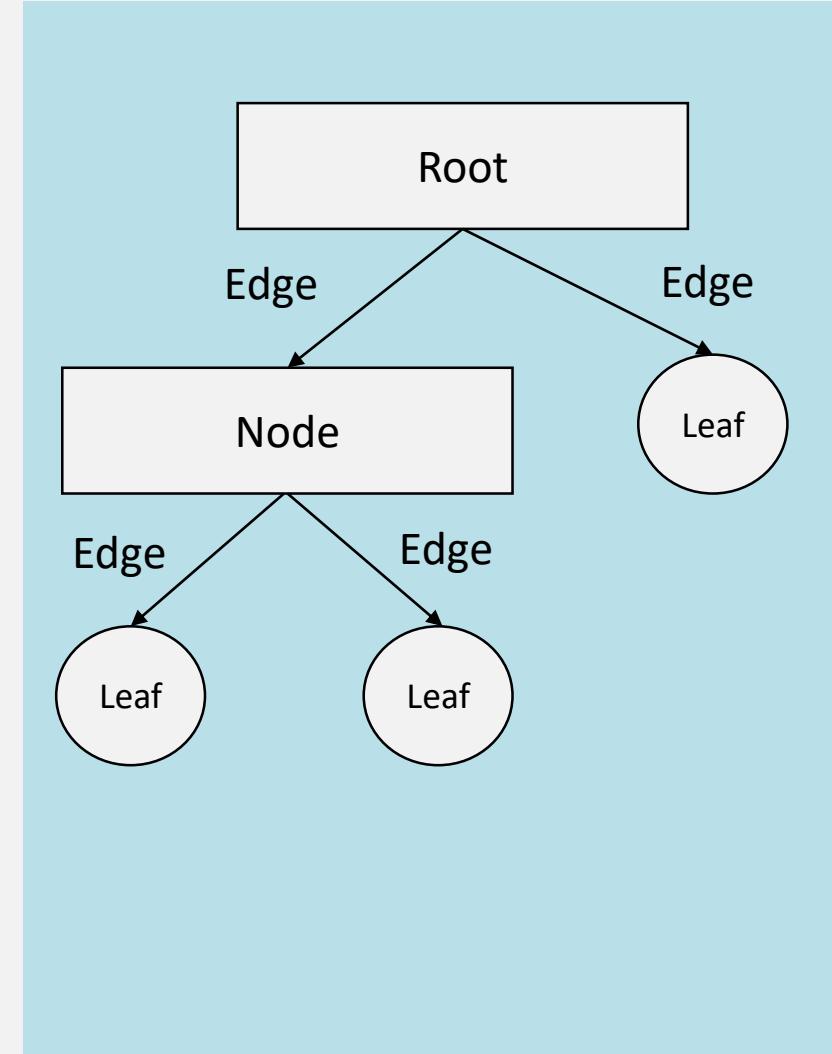
### Procedure

1. If all data objects belong to the same class, you form a leaf in the tree.
2. Otherwise:
  - a) An attribute is selected,
  - b) The number of data objects is partitioned according to the attribute values that occur for this attribute, which represent each successive node and
  - c) The algorithm is recursively applied to these successor nodes and the Remaining attributes

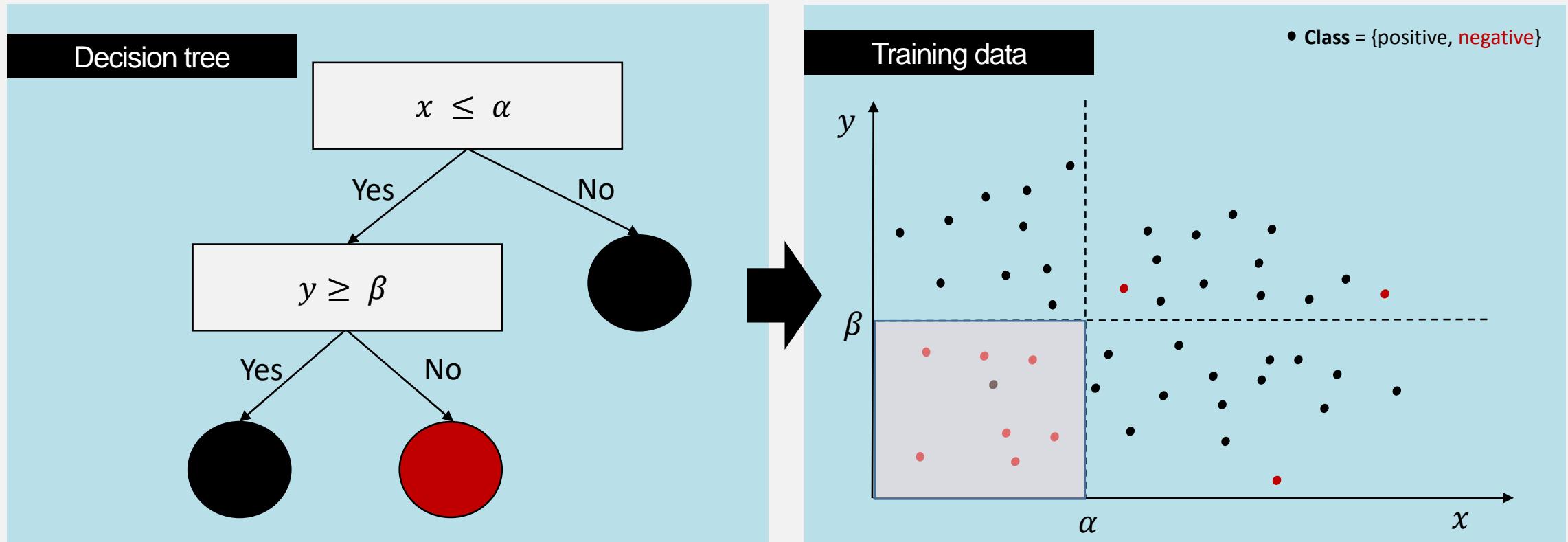
Adapted from Hunt et al. (1966); Russel, S., & Norvig, P. (2016)

## 6.2 Decision Trees

- A decision tree is made of nodes (representing attributes), edges (representing possible values), and leaves (representing categories or classes)
- Each path in a decision tree from the root to a leaf can be rewritten as a set of rules, i.e. disjunctive normal form (DNF)
- Can represent arbitrary conjunction and disjunction



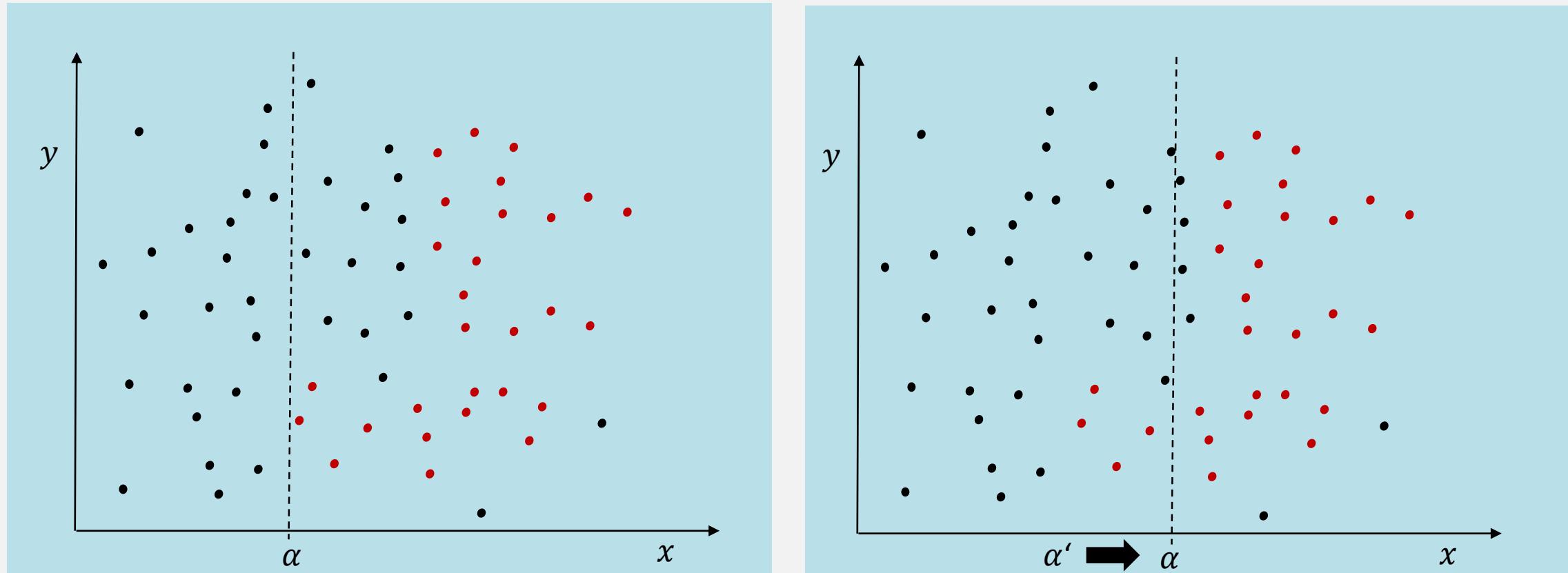
## 6.2 Geometric Interpretation: Decision Tree



**Top-Down Decision Tree Induction:** Decision trees divide the decision space into subspaces. Recursively build the decision tree top-down by divide and conquer. Objects falling into a certain space are classified as one respective group.

Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

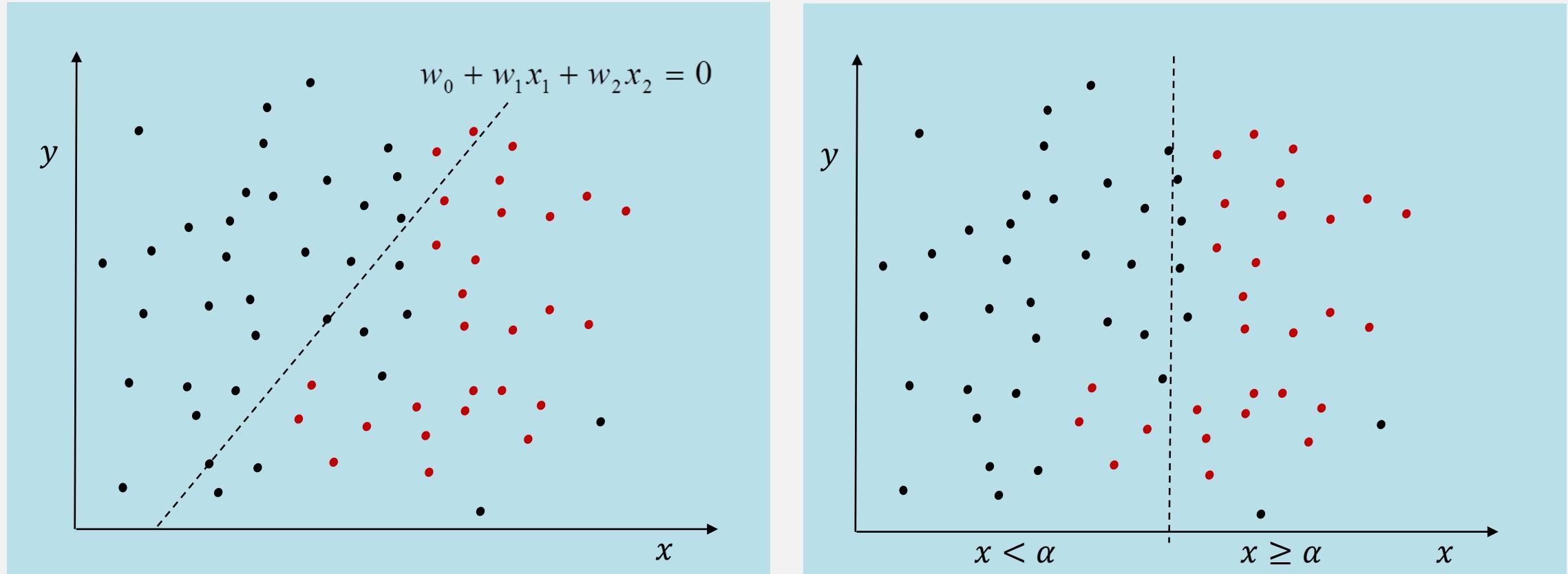
## 6.2 Geometric Interpretation: Linear Classification Problem



Which one of these would be better?

Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

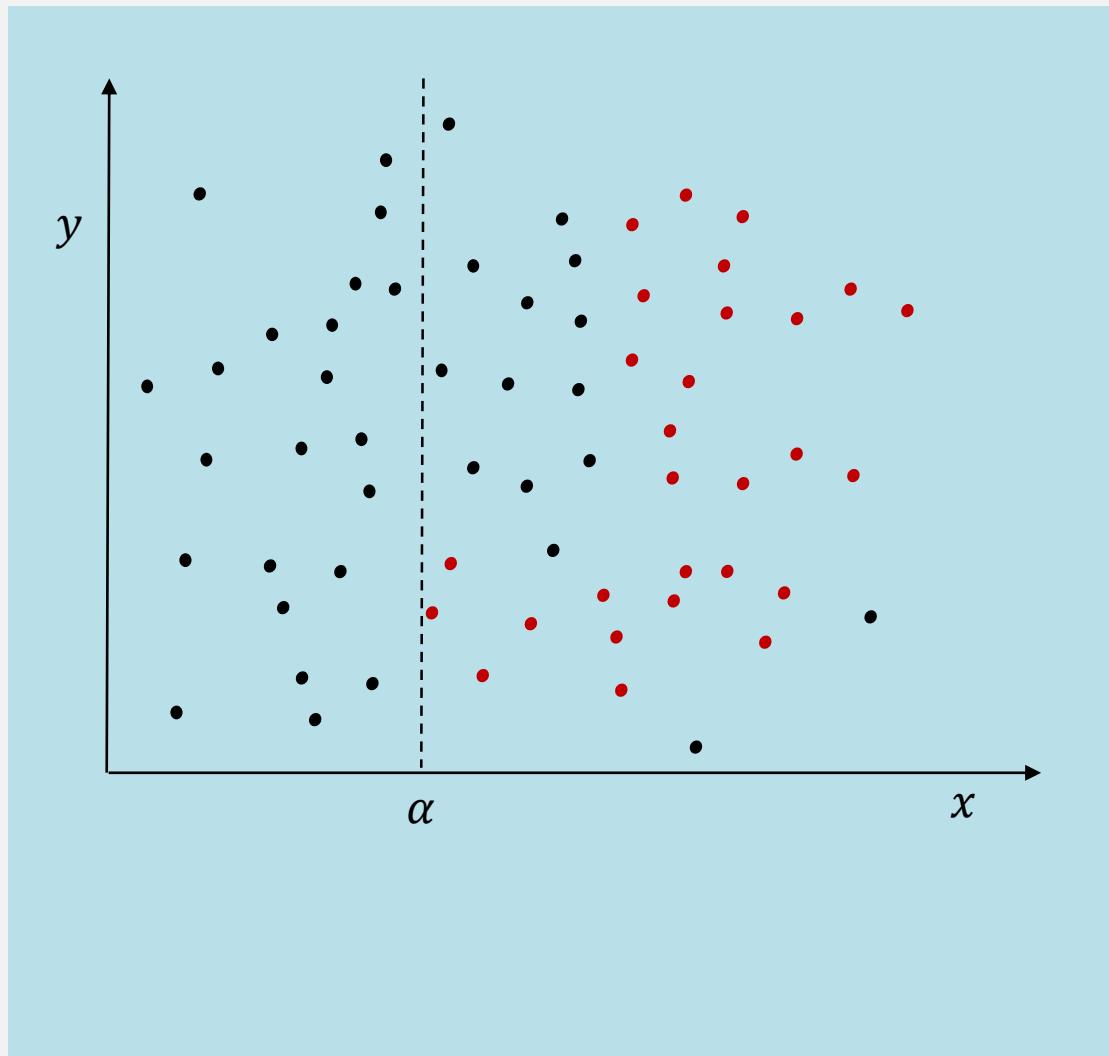
## 6.2 Geometric Interpretation: Decision Trees and Linear Classifiers



Where is the difference between these two? Advantage? Disadvantage?

Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

## 6.2 Entropy as a Measure of Impurity



- We need a function that describes the “*impurity*” of a subset of data

- **Solution:** *Entropy*

$$H(S) = - \sum_{i=1}^n p_i \cdot \log_2 p_i$$

Here:  $H(S) = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$

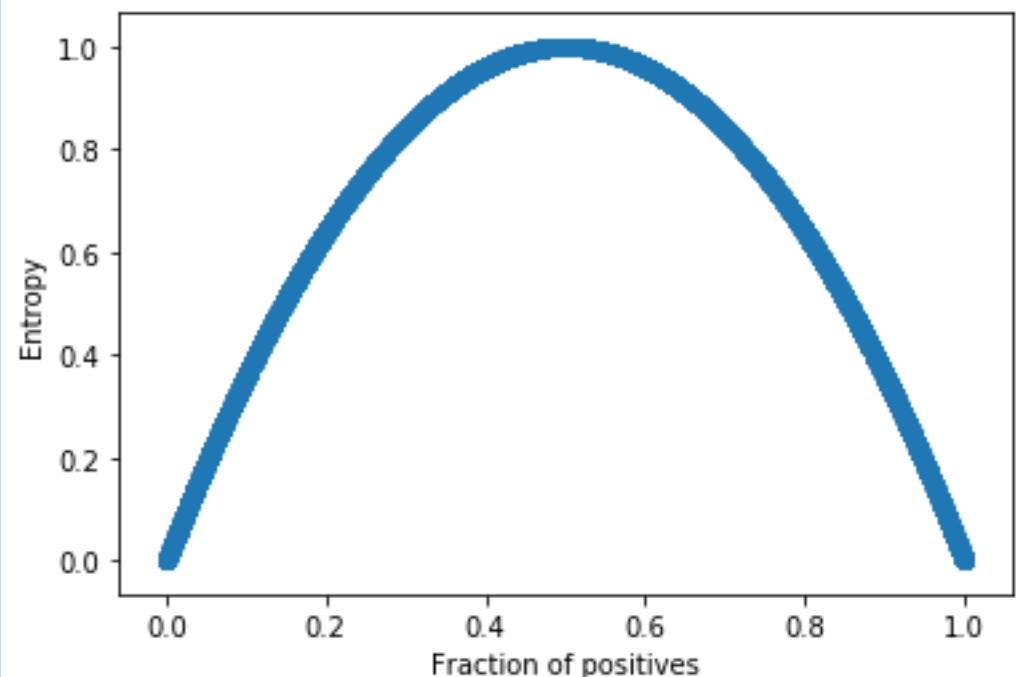
Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

## 6.2 Entropy as a Measure of Impurity

- Entropy (disorder, impurity) of a set of examples,  $S$ , relative to a binary classification is:

$$\text{Entropy}(S) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

- where  $p_1$  is the fraction of positive examples in  $S$  and  $p_0$  is the fraction of negatives.
- If all examples are in one category, entropy is zero (we define  $0\log(0) = 0$ )
- If examples are equally mixed ( $p_1 = p_0 = 0.5$ ), entropy is a maximum of 1



Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

## 6.2 Multi-Class Entropy

- Entropy can be viewed as the number of bits required on average to encode the class of an example in  $S$  where data compression (e.g. Huffman coding) is used to give shorter codes to more likely cases.
- For multi-class problems with  $c$  categories, entropy generalizes to:

$$\text{Entropy}(S) = \sum_{i=1}^c -p_i \log_2(p_i)$$

## 6.2 Picking a Good Split Feature

- Goal is to have the resulting tree be as small as possible (remember our generalization discussion about learning each single instance)
- Finding a minimal decision tree (nodes, leaves, or depth) is an NP-hard optimization problem
- Top-down divide-and-conquer method does a greedy search for a simple tree but does not guarantee to find the smallest
- Want to pick a feature that creates subsets of examples that are relatively “pure” in a single class so they are “closer” to being leaf nodes

## 6.2 Information Gain and Gini Impurity

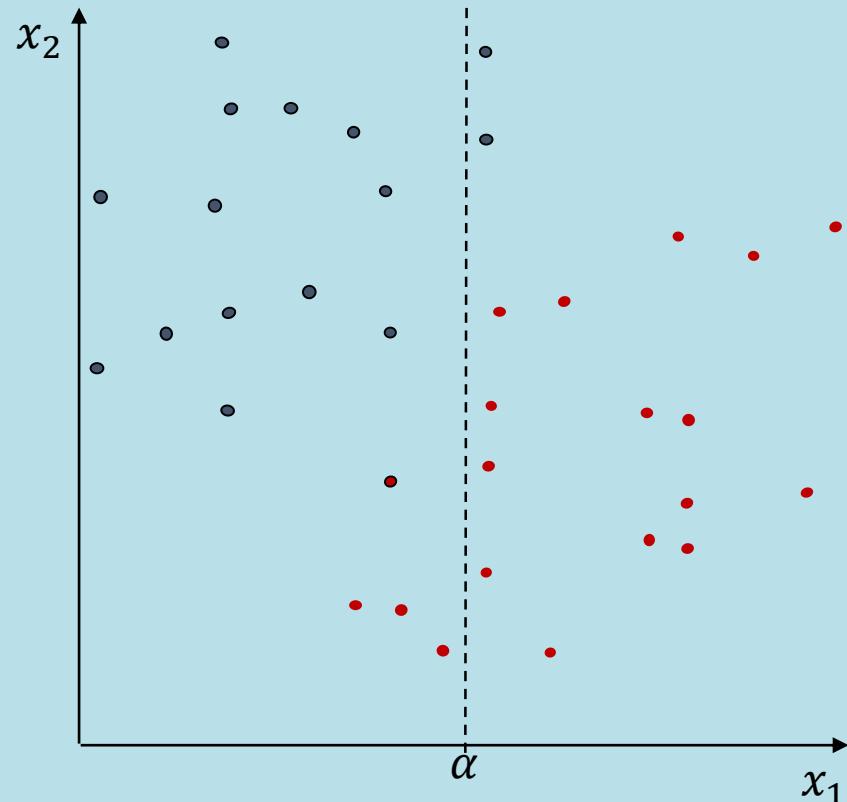
- Besides there are a variety of metrics that can be used as heuristic for picking a good split, like information gain of class label S and attribute A or *gini impurity*

$$Gini = 1 - \sum_{i=1}^C (p_i)^2$$

$$Gain(S, A) = Entropy(S) - Entropy(S, A)$$

## 6.2 Entropy and Information Gain

Entire population (34 instances)



Entire population (34 instances) “Parent entropy”:

$$-\left(\frac{15}{34} \log_2 \frac{15}{34}\right) - \left(\frac{19}{34} \log_2 \frac{19}{34}\right) = 0.989$$

“Child entropy” of left side (17 instances):

$$-\left(\frac{13}{17} \log_2 \frac{13}{17}\right) - \left(\frac{4}{17} \log_2 \frac{4}{17}\right) = 0.787$$

“Child entropy” of right side (17 instances):

$$-\left(\frac{2}{17} \log_2 \frac{2}{17}\right) - \left(\frac{15}{17} \log_2 \frac{15}{17}\right) = 0.523$$

Entropy of children:

$$\left(\frac{17}{34} 0.787\right) + \left(\frac{17}{34} 0.523\right) = 0.655$$

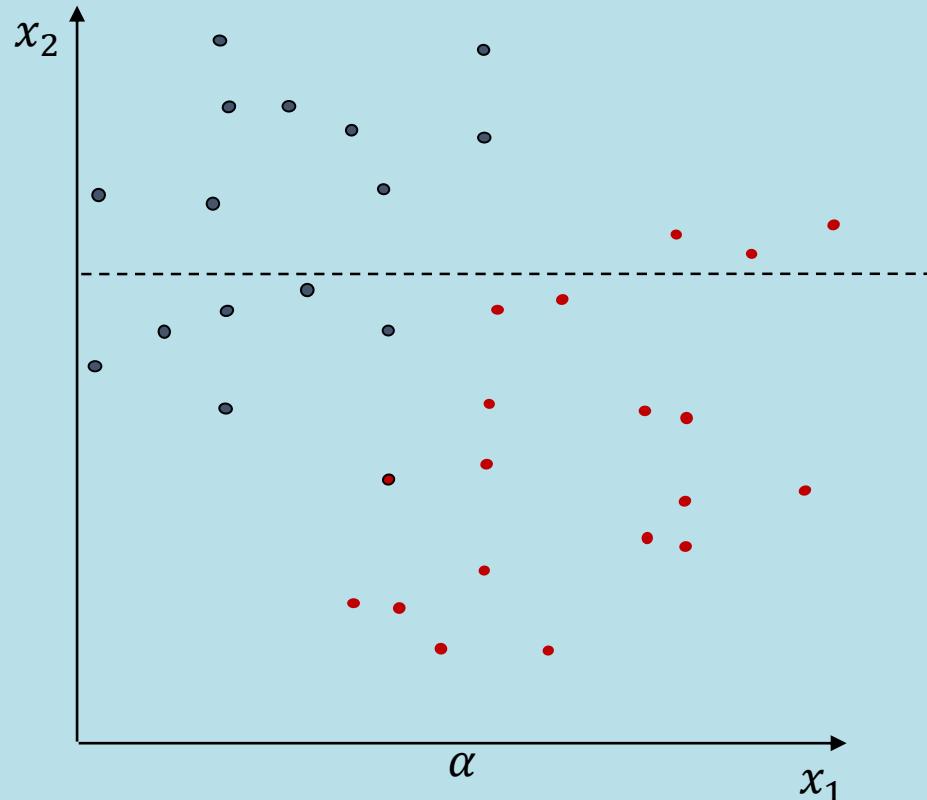
*Information gain:*

$$0.989 - 0.655 = 0.334$$

Adapted from Russell, S., & Norvig, P. (2016)

## 6.2 Entropy and Information Gain

Entire population (34 instances)



Entire population (34 instances):

$$-\left(\frac{15}{34} \log_2 \frac{15}{34}\right) - \left(\frac{19}{34} \log_2 \frac{19}{34}\right) = 0.989$$

$$\left(\frac{12}{34} 0.811\right) + \left(\frac{22}{34} 0.845\right) = 0.833$$

Entropy above (12 instances):

$$-\left(\frac{9}{12} \log_2 \frac{9}{12}\right) - \left(\frac{3}{12} \log_2 \frac{3}{12}\right) = 0.811$$

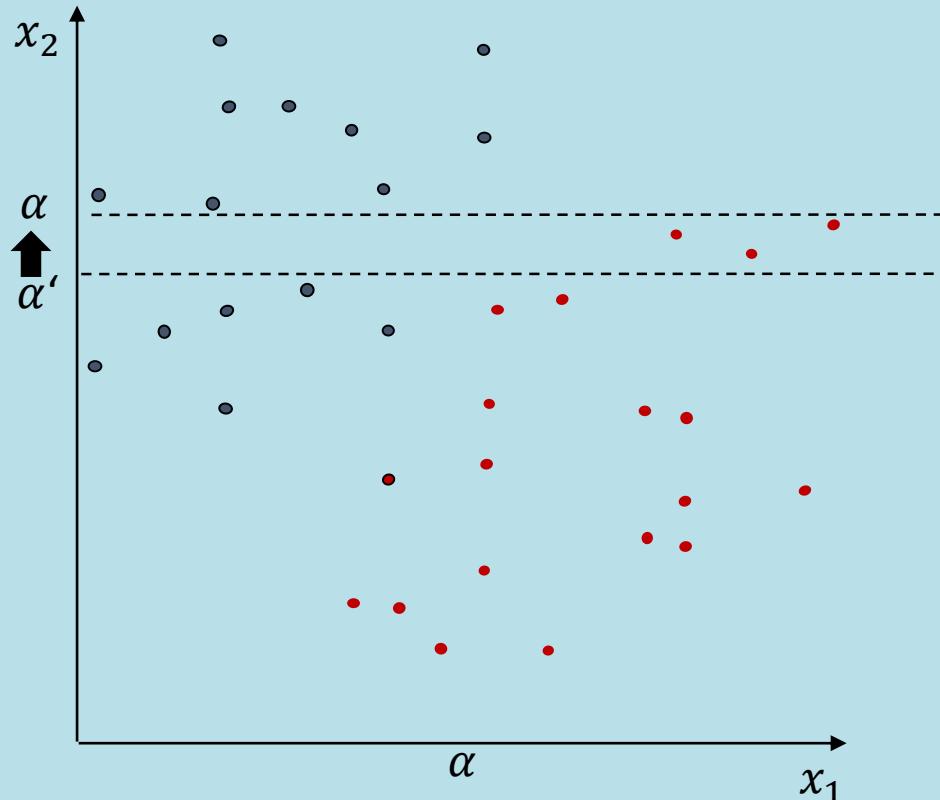
Entropy below (22 instances):

$$-\left(\frac{6}{22} \log_2 \frac{6}{22}\right) - \left(\frac{16}{22} \log_2 \frac{16}{22}\right) = 0.845$$

Adapted from Russell, S., & Norvig, P. (2016)

## 6.2 Entropy and Information Gain

Entire population (34 instances)



Entire population (34 instances):

$$-\left(\frac{15}{34} \log_2 \frac{15}{34}\right) - \left(\frac{19}{34} \log_2 \frac{19}{34}\right) = 0.989$$

$$\left(\frac{12}{34} 0.811\right) + \left(\frac{22}{34} 0.845\right) = 0.833$$

Entropy above (9 instances):

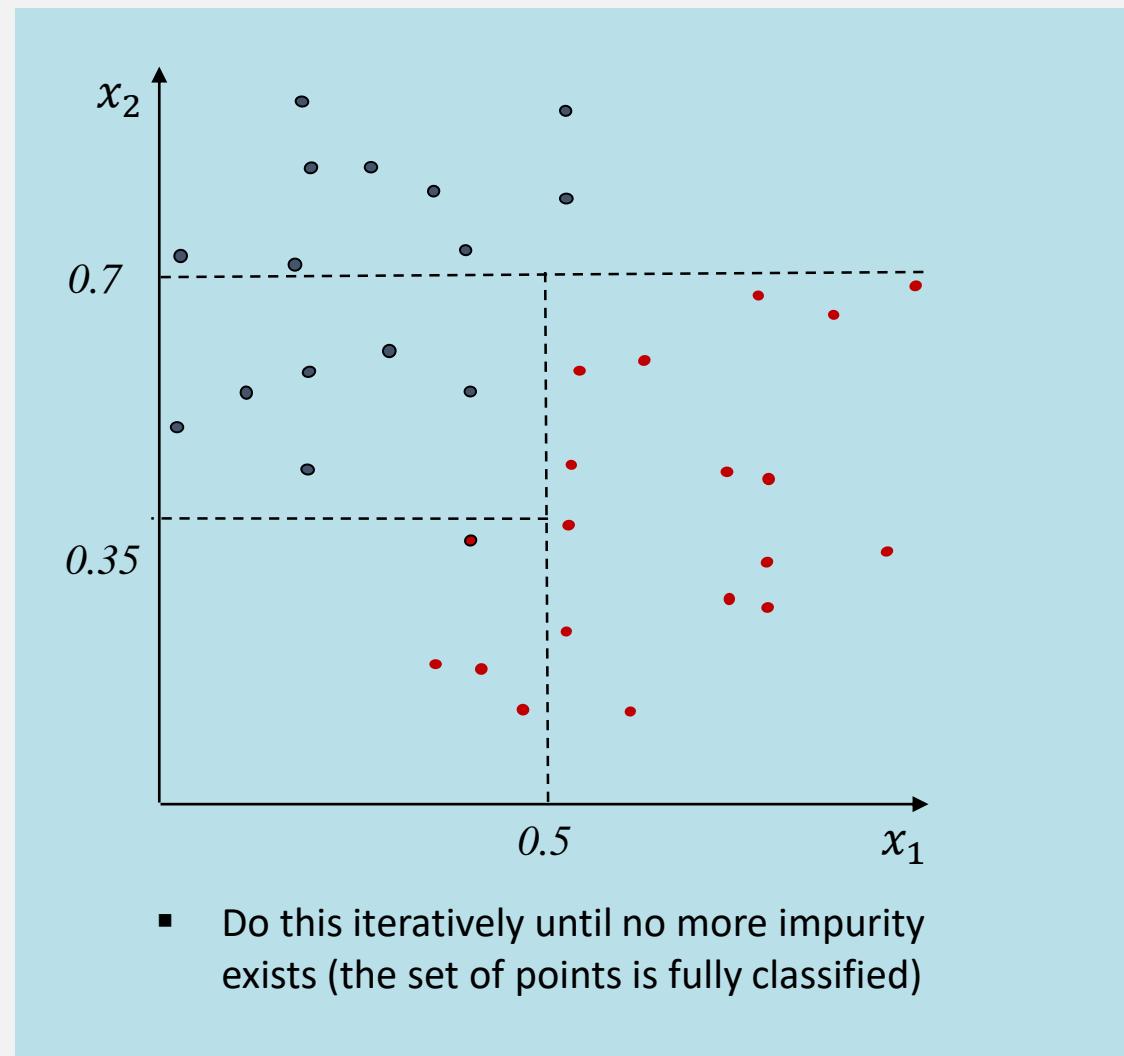
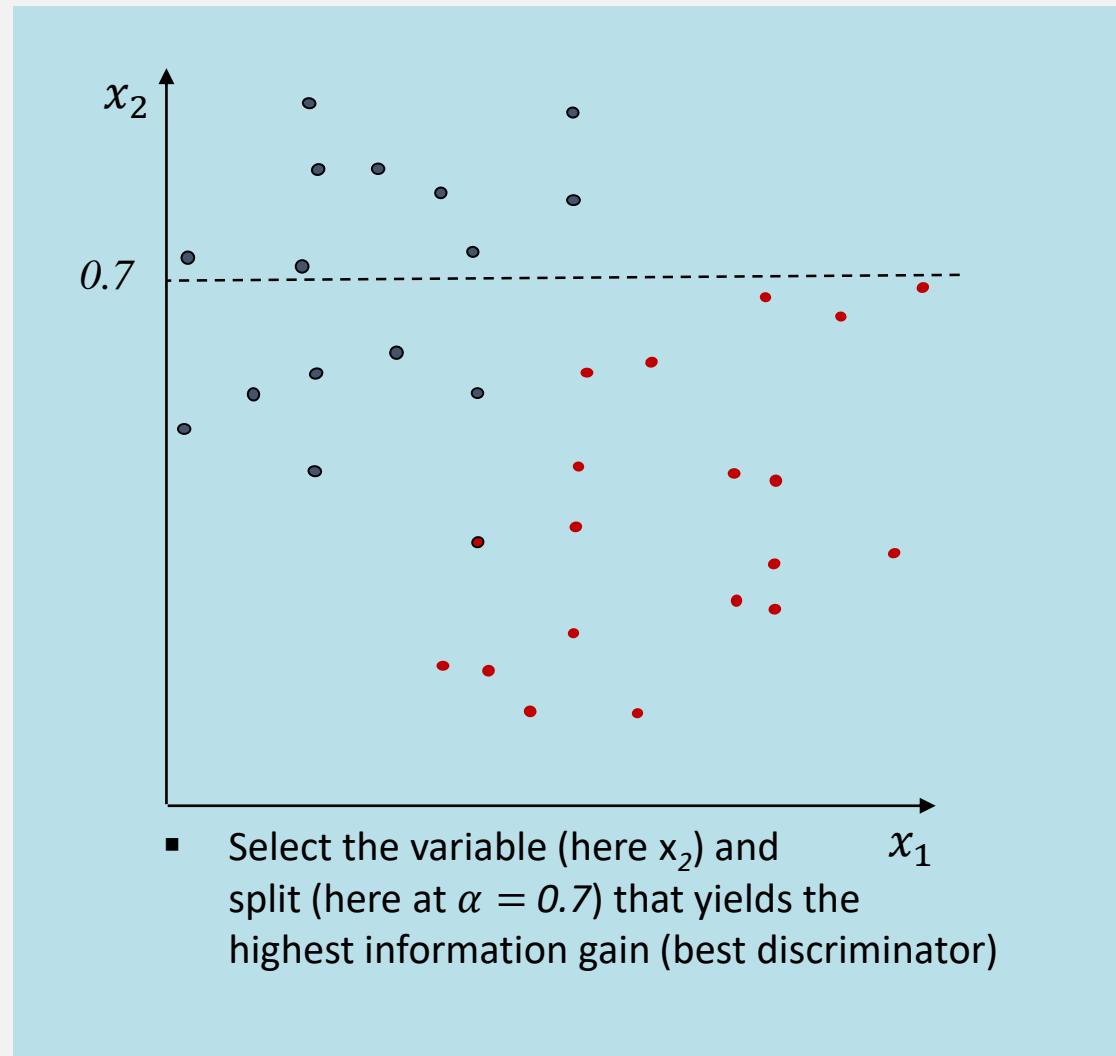
$$-\left(\frac{9}{9} \log_2 \frac{9}{9}\right) - \left(\frac{0}{9} \log_2 \frac{0}{9}\right) = 0$$

Entropy below (25 instances):

$$-\left(\frac{6}{25} \log_2 \frac{6}{25}\right) - \left(\frac{19}{25} \log_2 \frac{19}{25}\right) = 0.795$$

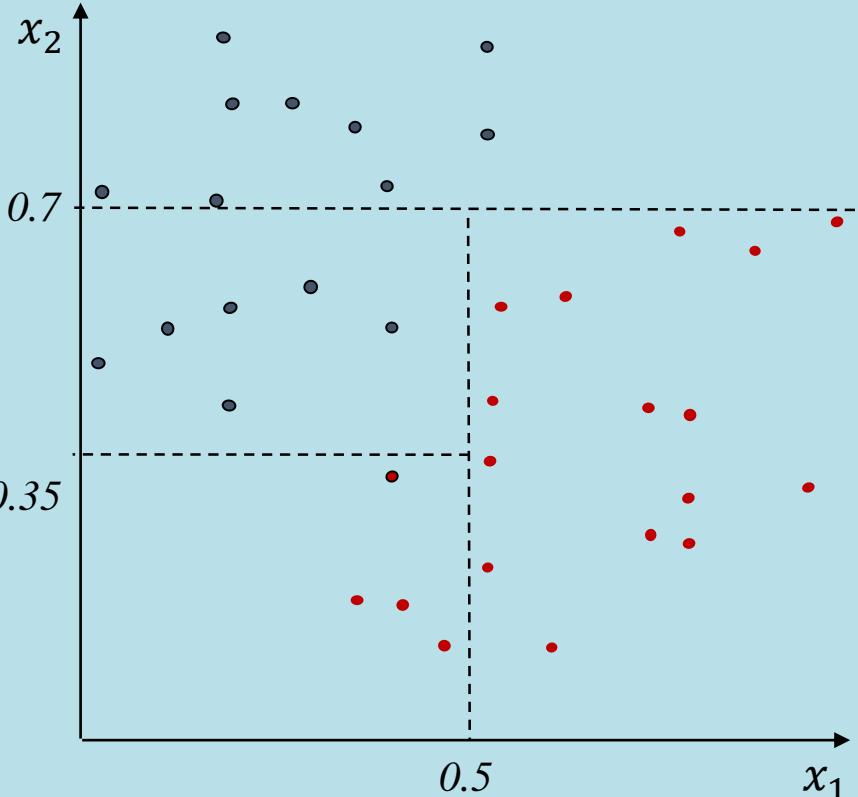
Adapted from Russell, S., & Norvig, P. (2016)

## 6.2 Entropy and Information Gain



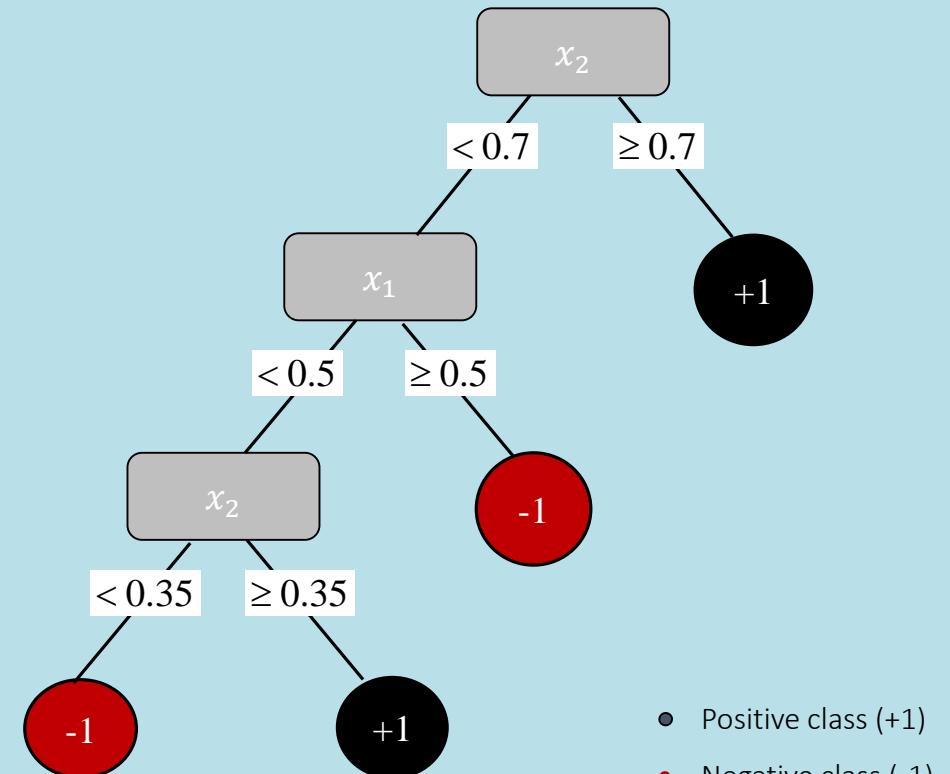
Adapted from Russell, S., & Norvig, P. (2016)

## 6.2 Final Decision Tree



- Do this iteratively until no more impurity exists (the set of points is fully classified)

- This iterative process of splitting can be expressed as a “*decision tree*”:



Adapted from Russell, S., & Norvig, P. (2016)

## 6.2 DecisionTreeClassifier in Python



DecisionTreeClassifier()

DecisionTreeClassifier(criterion, max\_depth, )

### Parameters

criterion	The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain.
max_depth	You can set an individual type of index for the row labels, the default index if no index is passed is np.arange(n).

## 6.2 Simple Decision Tree with Python

- Build your own decision tree with Python

```
from sklearn.tree import DecisionTreeClassifier  
  
clf = DecisionTreeClassifier(random_state=0)  
clf = clf.fit(X_train, Y_train)
```

- After being fitted, the model can then be used to predict the class of samples

```
clf.predict(X_test, Y_test)
```

## 6.2 Plot Decision Trees with the `plot_tree` Function

```
X[1] <= 2.5  
gini = 0.5  
samples = 4  
value = [2, 2]
```

```
gini = 0.0  
samples = 2  
value = [0, 2]
```

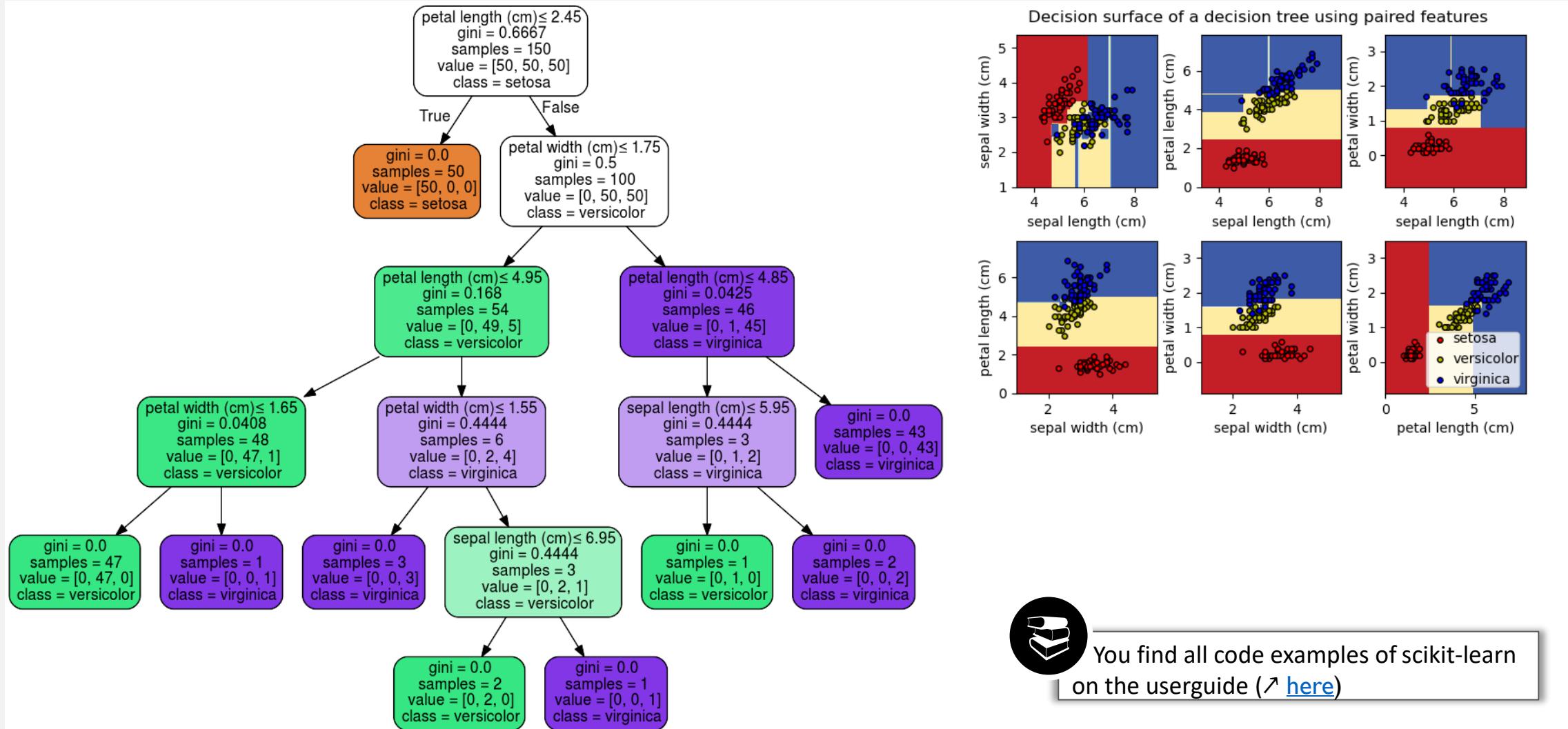
```
gini = 0.0  
samples = 2  
value = [2, 0]
```

```
X = array([[1, 1, 3],  
          [3, 3, 5],  
          [1, 2, 5],  
          [2, 3, 2]])  
  
Y = array([1, 0, 1, 0])  
  
clf = tree.DecisionTreeClassifier()  
clf = clf.fit(X, Y)  
  
tree.plot_tree(clf)
```



You find all code examples in the ↗ code folder of the lecture repository

## 6.2 Scikit-learn has Many Tools to Visualize Decision Trees



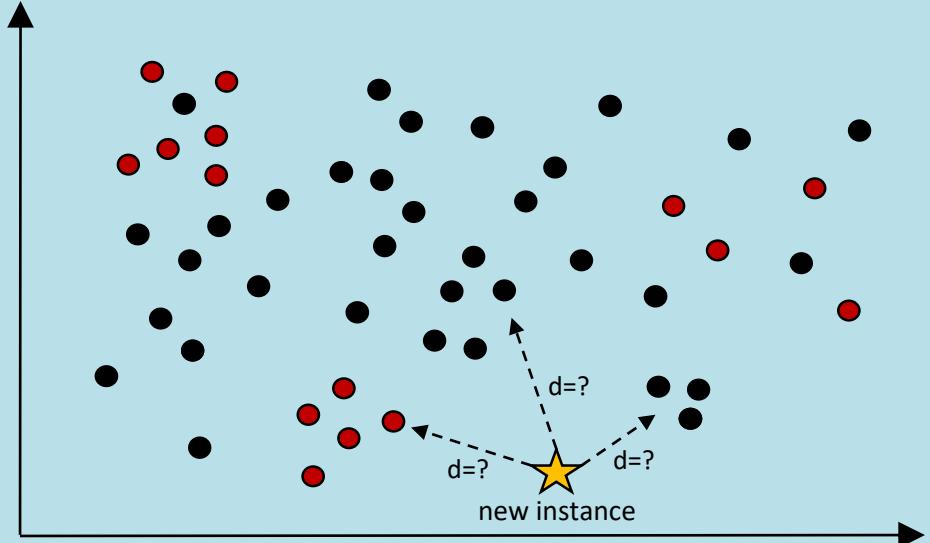
- The scikit-learn module provides many useful decision tree related functions

<code>cost_complexity_pruning_path(X, y [, ...])</code>	Compute the pruning path during Minimal Cost-Complexity Pruning.
<code>decision_path(X[, check_input])</code>	Return the decision path in the tree.
<code>fit(X, y[, sample_weight, check_input, ...])</code>	Build a decision tree classifier from the training set (X, y).
<code>get_depth()</code>	Return the depth of the decision tree.
<code>get_n_leaves()</code>	Return the number of leaves of the decision tree.
<code>predict(X[, check_input])</code>	Predict class or regression value for X.
<code>predict_proba(X[, check_input])</code>	Predict class probabilities of the input samples X.
<code>score(X, y[, sample_weight])</code>	Return the mean accuracy on the given test data and labels.

## 6.2 Decision Tree Algorithms

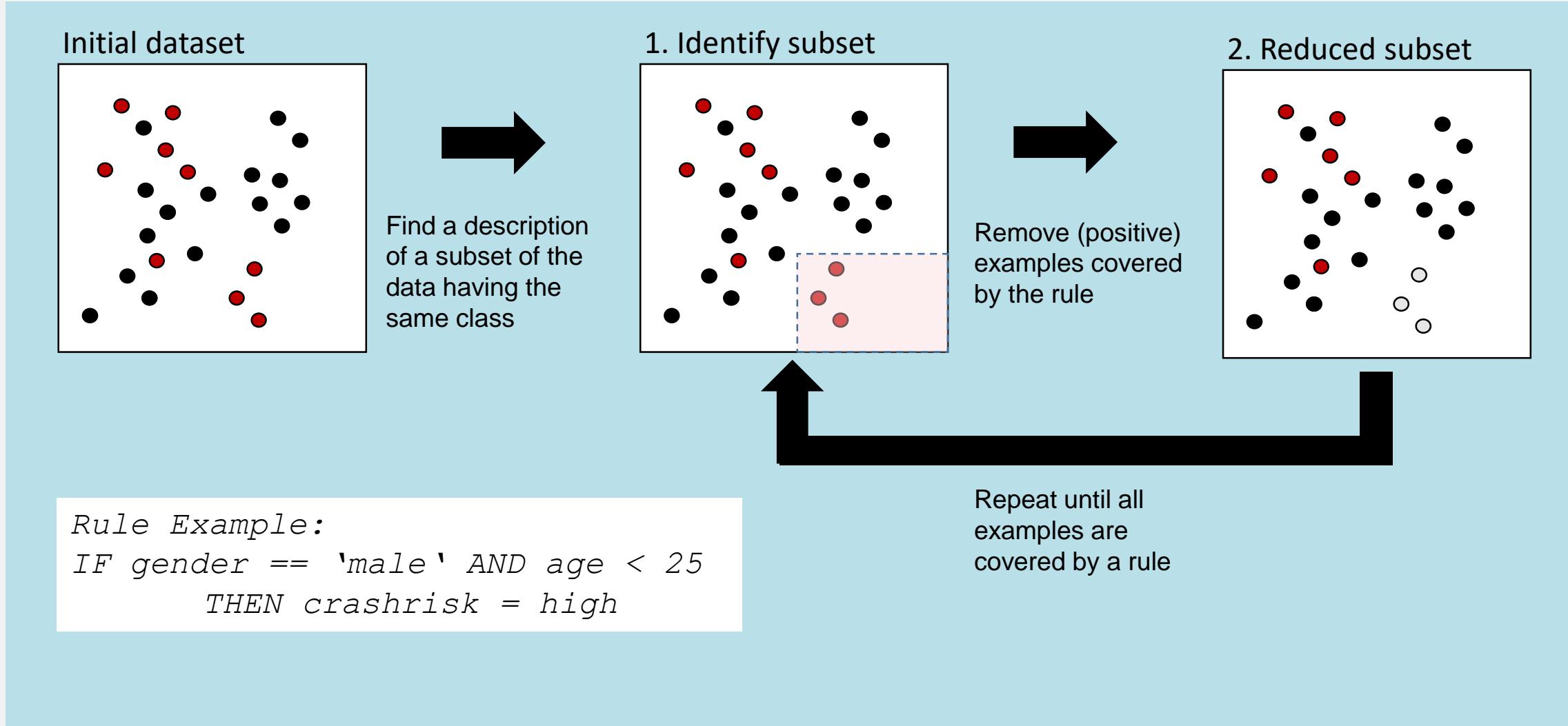
Algorithm	Authors	Characteristics
Chi <sup>2</sup> -Automatic Interaction Detection (CHAID)	Sonquist and Morgan, 1994	<ul style="list-style-type: none"><li>▪ One of the first decision tree learning algorithms, building on Chi<sup>2</sup>-test as split criterion</li><li>▪ Only nominal attributes</li></ul>
Classification and Regression Tree (CART)	Breiman et al., 1984	<ul style="list-style-type: none"><li>▪ Only nominal and numeric attributes with binary splits based on Gini-Index</li><li>▪ Considers all possible subsets for nominal attributes</li></ul>
Iterative Dichotomiser 3 (ID3)	Quinlan, 1986	<ul style="list-style-type: none"><li>▪ Simple decision tree with multiple splits per node</li><li>▪ Information Gain as split criterion</li></ul>
C4.5	Quinlan, 1993	<ul style="list-style-type: none"><li>▪ Building on ID3, numeric attributes are accepted as splits</li><li>▪ Pruning is possible</li></ul>
C5.0	Quinlan, 1997	<ul style="list-style-type: none"><li>▪ Increased performance (accuracy, speed, and less memory)</li></ul>

## 6.2 K-Nearest-Neighbor (KNN)



- Find rules by separate-and-conquer describing different subsets of the dataset
- Expands existing hand-crafted rule-based systems with a computed knowledge base
- Human-readable set of rules derived from the train dataset
- Rules are constructed based on equality
- Works best with nominal attributes

## 6.2 (Conjunctive) Rule Learning Procedure



Adapted from Russell, S., & Norvig, P. (2016)

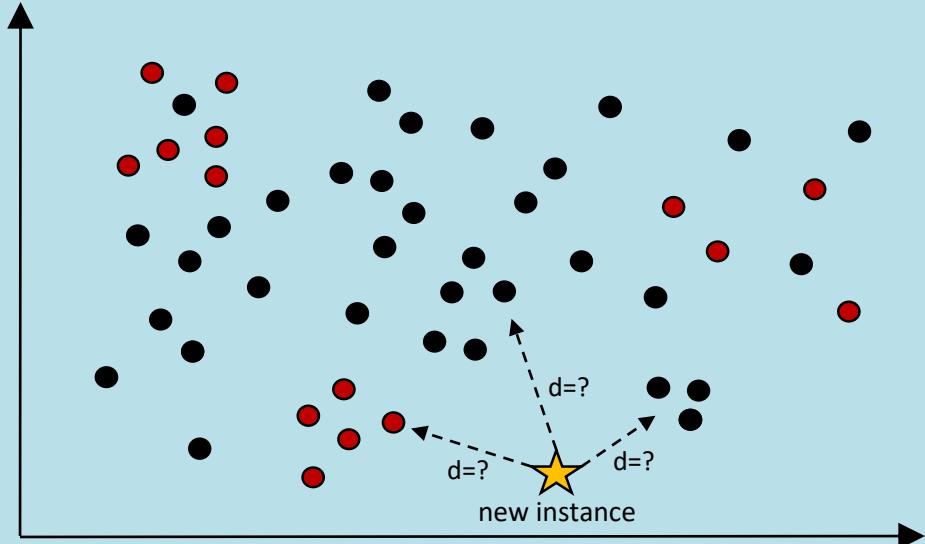
## 6.2 Problems of Rule Learning

- Must check consistency with negative examples. If inconsistent, **no** conjunctive rule exists
- If a concept does not have a single set of necessary and sufficient conditions, conjunctive learning fails

## 6.2 Rule Learner Algorithms

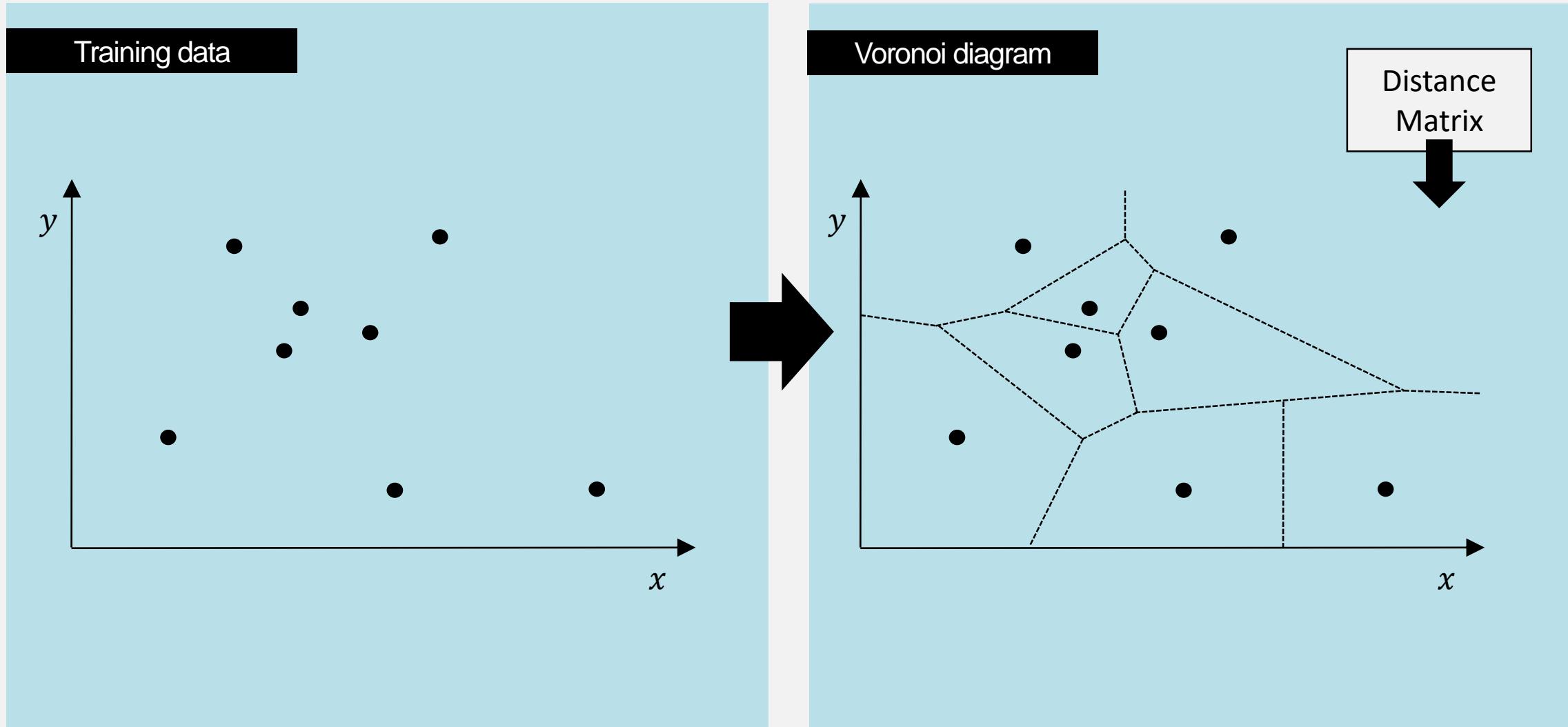
Algorithm	Authors	Characteristics
First-Order Inductive Learning (FOIL)	Quinlan, 1990	<ul style="list-style-type: none"><li>▪ Rule improvement is evaluated based on its predecessor</li><li>▪ Minimum description length as stopping criterion</li></ul>
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)	Cohen, 1995	<ul style="list-style-type: none"><li>▪ Reduces overfitting of previous rule learners</li><li>▪ Post-processing: Remove rule, and try to learn it again</li></ul>
Optimized Pruning for Unordered Search (OPUS)	Webb, 1995	<ul style="list-style-type: none"><li>▪ Search all possible rules</li><li>▪ Order results to evaluate a rule multiple times</li></ul>
FPGrowth	Han, 2000	<ul style="list-style-type: none"><li>▪ Association Rule learner</li><li>▪ Ordered search</li></ul>

## 6.2 K-Nearest-Neighbor (KNN)



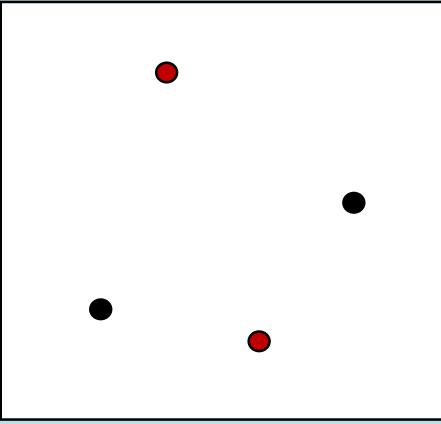
- Classify an object by his  $k$  nearest neighbours
- K-Nearest-Neighbor approach is no „real model“ ist just a type of simple classification heuristic (there is no learning)
- When a new instance comes in find the  $k$  closest neighbors and use the majority class to classify the new instance

## 6.2 Geometric Interpretation: Voronoi diagram



## 6.2 K-Nearest Neighbor Procedure

Initial dataset



Compute the  
distance of each  
point-to-point relation  
(use heuristics to  
speed up)

1. Distance Matrix

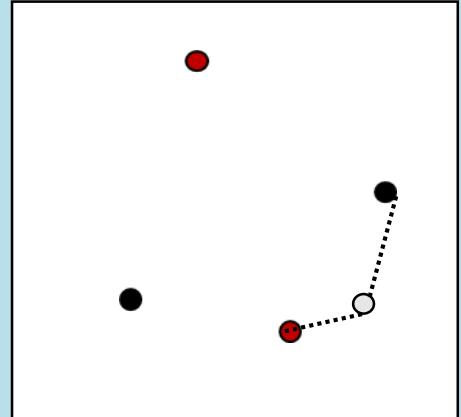
Object	O1	O2	On
O1	0	1	4
O2	2	0	3
...	...	...	...

A diagram below the matrix shows a horizontal dotted line connecting the cell containing '1' (representing the distance between O1 and O2) to the cell containing 'O2' in the row header. Another dotted line connects the cell containing '4' (representing the distance between O1 and On) to the cell containing 'On' in the row header.



Use distance matrix  
to label new objects

*Label new subsets*



The distance between two objects (e.g.  $X, Y$ ) is calculated by different types of distance

Adapted from Géron, A. (2017); Russel, S., & Norvig, P. (2016)

## 6.2 KNN with Python

- Build your own KNN-classifier with Python

```
>>> import numpy as np
>>> from sklearn.impute import KNNImputer

>>> nan = np.nan
>>> X = [[1, 2, nan], [3, 4, 3], [nan, 6, 5], [8, 8, 7]]

>>> imputer = KNNImputer(n_neighbors=2, weights="uniform")
>>> imputer.fit_transform(X)

array([[1., 2., 4.],
       [3., 4., 3.],
       [5.5, 6., 5.],
       [8., 8., 7.]])
```

- Classification accuracy (percentage of instances classified correctly).
  - Measured on an independent test data.
- Training time (efficiency of training algorithm).
- Testing time (efficiency of subsequent classification).

## 6.2 Evaluation of Classification Models

Confusion Matrix	
True class	
Positive	Negative
Positive	True Positive Count (TP)
Negative	False Negative Count (FN)
Negative	True Negative Count (TN)

D

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{True Positive Rate} = \frac{TP}{TP + FN}$$

$$\text{True Negative Rate} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

## 6.2 Example Confusion Matrix

**Task:** Car classification, *positive class = Taycan*

	Taycan	911	718
Taycan	4	2	2
911	2	2	8
718	1	0	3



	Taycan	911	718
Taycan	TP (4)	FN (4)	
911		FP (3)	TN (13)
718			

**Accuracy:** Amount of correct model predictions

$$A = \frac{(TP + TN)}{Total} = \frac{4 + 13}{4 + 4 + 3 + 13} = 70,83\%$$

- 4 actually positive cases were correctly assigned
- 13 other cases were correctly assigned as negative
- The model is correct in 70% of the Taycan detection tasks

Adapted from Géron, A. (2017); Rusell, S., & Norvig, P. (2016)

## 6.2 Confusion Matrix in Python

- In Python you can compute confusion matrices easily with scikit-learn

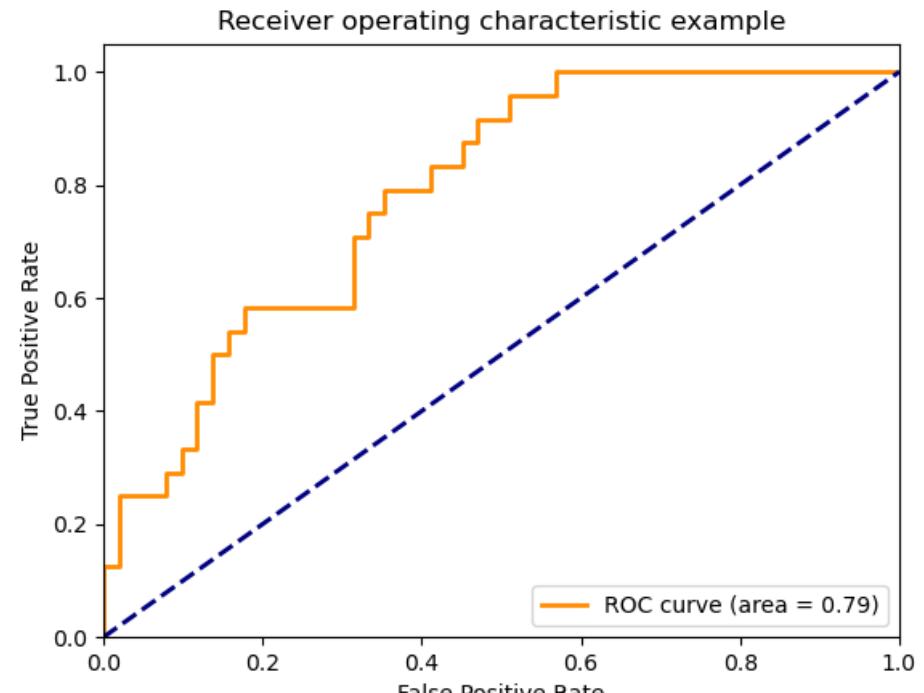
```
>>> from sklearn.metrics import confusion_matrix

>>> y_true = ["Taycan", "911 ", "Taycan ", "Taycan ", "911", "718"]
>>> y_pred = ["911", "911", "Taycan ", "Taycan ", "911", "Taycan "]

>>> confusion_matrix(y_true, y_pred, labels=["911 ", "718", "Taycan"])

array([[2, 0, 0],
       [0, 0, 1],
       [1, 0, 2]]))
```

## 6.2 Visualization of the Accuracy – Receiver Operating Characteristic



```
from sklearn import metrics  
metrics.roc_curve()
```

- Receiver Operating Characteristic (ROC) is used to assess the accuracy of a continuous measurement for predicting a binary outcome
- ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis
- The top left corner of the plot is the “ideal” point - a false positive rate of zero, and a true positive rate of one

Adapted from Géron, A. (2017); Russel, S., & Norvig, P. (2016)

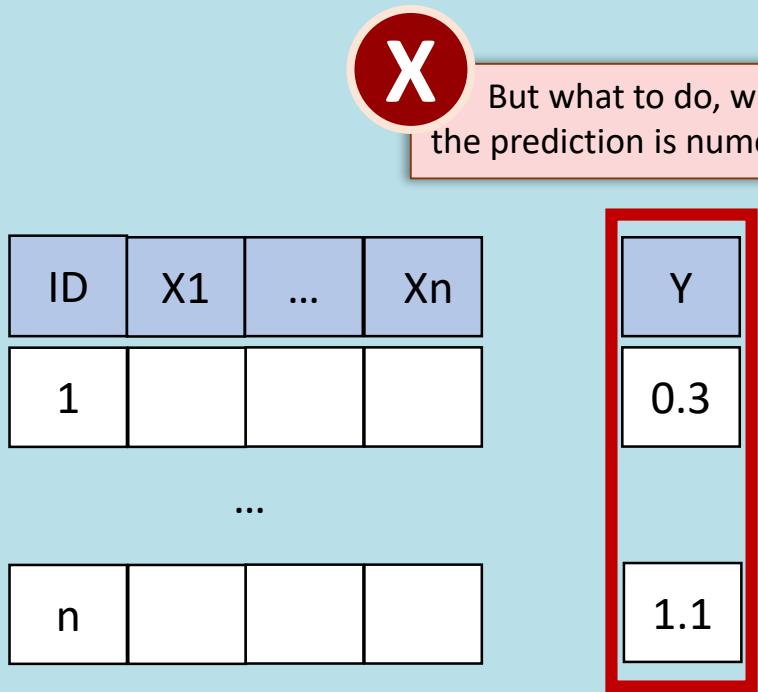
## 6.2 Area Under the Curve (AUC)

- The `roc_auc_score` function computes the area under the receiver operating characteristic curve, which is also denoted by AUC or AUROC.

```
>>> import numpy as np  
>>> from sklearn.metrics import roc_auc_score  
  
>>> y_true = np.array([0, 0, 1, 1])  
>>> y_scores = np.array([0.1, 0.4, 0.35, 0.8])  
>>> roc_auc_score(y_true, y_scores)  
  
0.75
```

- By computing the area under the roc curve, the curve information is summarized in one number

## 6.2 Regression Problem

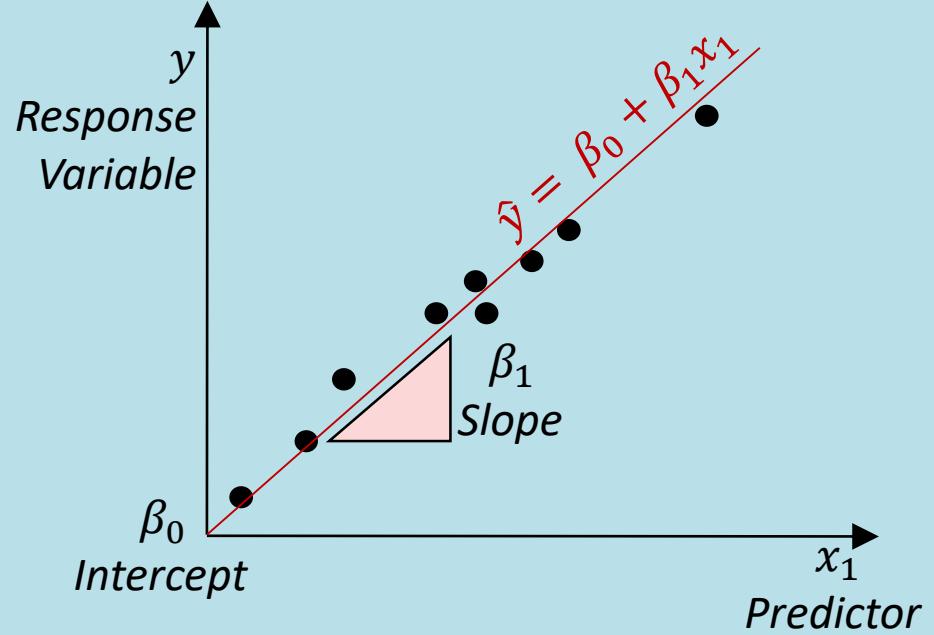


Estimate a continuous value

Common real-life problems:

- Compute scores: Fraud detection
- Predict values in finance and investing like product prices, stock prices etc.
- Estimate demographic characteristics in marketing like salary, age, size etc.

## 6.2 Linear Regression



- Linear regression fits a linear model with coefficients  $w = (w_1, \dots, w_n)$  to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation
- **Univariate (single attribute)**  
$$\bar{y} = \beta_0 + \beta_1 x_1$$
- **Multivariate (many attributes)**  
$$\bar{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

## 6.2 Non-Linear Regression Methods

Power

Exponential

Logarithmic

Polynomial

### Method properties

- Analogous to linear regression model
- Specify nonlinear function must be preselected (risk of misspecification)
- Computationally more complex than linear regression
- Sensitive to outliers

### Data requirements

- Complete numeric, independent predictors

Adapted from Géron, A. (2017); Rusell, S., & Norvig, P. (2016)

## 6.2 Evaluation of Regressions Models

### Regression Error

Regression Error / Residuals: Measures how far off the predicted value is from the actual known value

### Loss Function

Error:  $e_t = y_t - \hat{y}_t$

Absolute Error:  $|y_t - \hat{y}_t|$

Squared Error:  $(y_t - \hat{y}_t)^2$

Observation:  $y_t$

Prediction:  $\hat{y}_t$

### Accuracy Measures

Measure	Formula
Mean Absolute Error	$MAE = \text{average}( e_t )$
Mean Squared Error	$MSE = \text{average}( e_t^2 )$
Mean Absolute Percentage Error	$MAPE = 100 \cdot \text{average}\left(\left \frac{e_t}{y_t}\right \right)$
Mean Absolute Scaled Error	$MASE = \frac{MAE}{Q}$

Q: Scaling constant

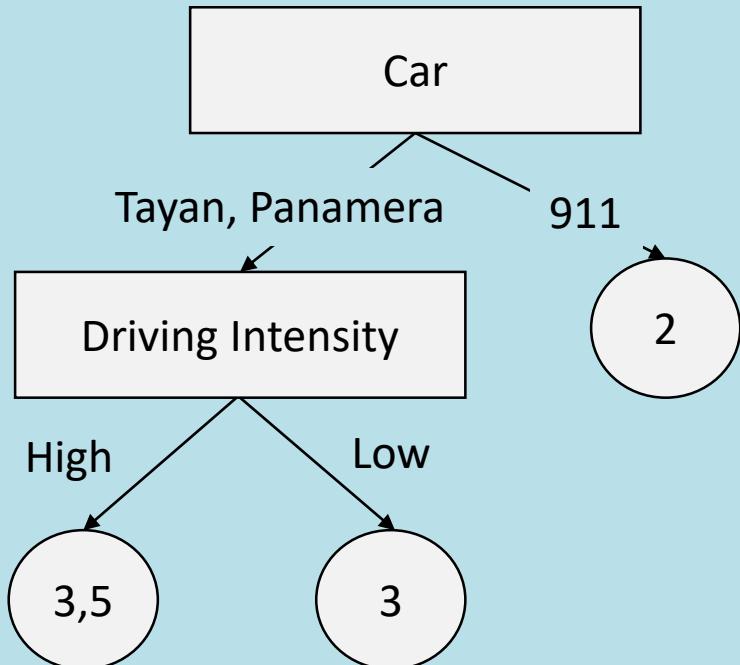
Adapted from Géron, A. (2017); Russel, S., & Norvig, P. (2016)

## 6.2 Decision Tree Learning and Continuous Data (Classes or Features)

- Continuous features can be handled by allowing nodes to split a real valued feature into two ranges based on a threshold (e.g.  $\text{length} < 3$  and  $\text{length} \geq 3$ )
- Classification trees have discrete class labels at the leaves, regression trees allow real-valued outputs at the leaves.

## 6.2 Basic Idea of Regression Trees

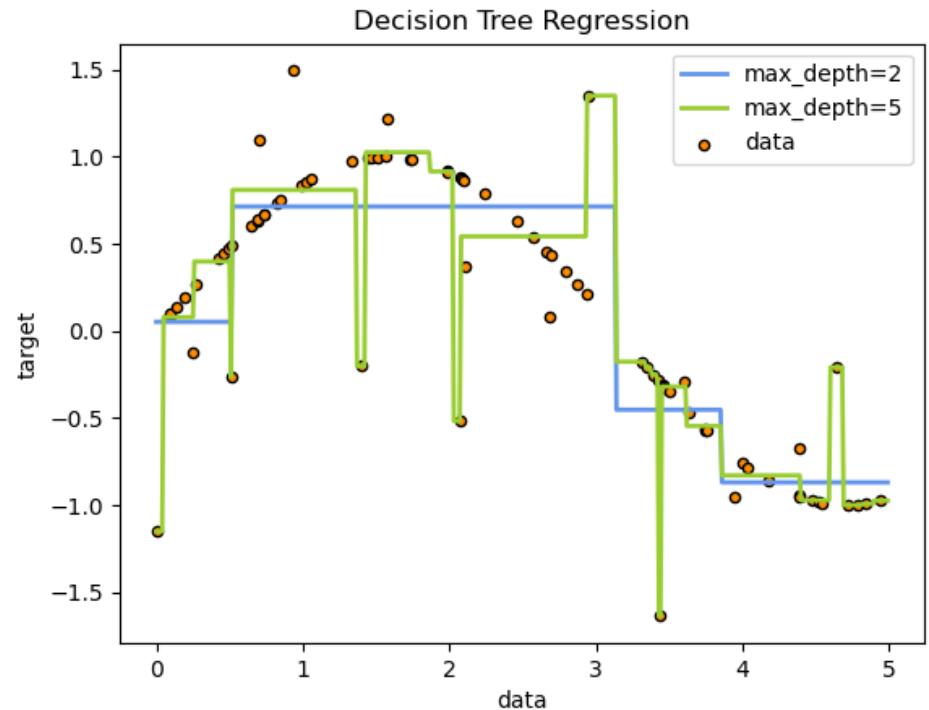
Concept similar to decision tree, but store continuous value at each leaf representing the prediction



Car	Driving_intensity	Num_of_defects
Taycan	High	3
Taycan	High	4
911	Low	2
911	High	2
Panamera	Low	3
...	...	...

Adapted from Géron, A. (2017); Rusell, S., & Norvig, P. (2016)

## 6.2 Regression Trees with Python



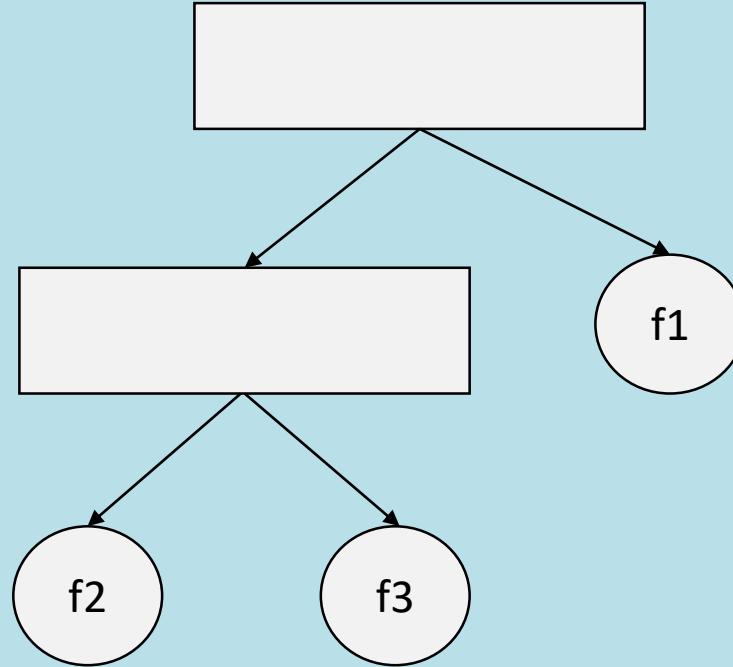
```
>>> from sklearn import tree  
  
>>> X = [[0, 0], [2, 2]]  
>>> y = [0.5, 2.5]  
  
>>> clf = tree.DecisionTreeRegressor()  
>>> clf = clf.fit(X, y)  
>>> clf.predict([[1, 1]])  
  
array([0.5])
```

- As in the classification setting, the fit method will take as argument arrays X and y,
- y is expected to have floating point values instead of integer values

Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

## 6.2 Basic Idea of Model Trees

- Concept similar to decision tree, but store regression models in the leaf node
- Basic algorithm named M5



Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

## 6.2 More Model Evaluation Metrics on Scikit-learn

### 3.3.1.1. Common cases: predefined values

For the most common use cases, you can designate a scorer object with the `scoring` parameter; the table below shows all possible values. All scorer objects follow the convention that **higher return values are better than lower return values**. Thus metrics which measure the distance between the model and the data, like `metrics.mean_squared_error`, are available as `neg_mean_squared_error` which return the negated value of the metric.

Scoring	Function	Comment
<b>Classification</b>		
'accuracy'	<code>metrics.accuracy_score</code>	
'balanced_accuracy'	<code>metrics.balanced_accuracy_score</code>	
'average_precision'	<code>metrics.average_precision_score</code>	
'neg_brier_score'	<code>metrics.brier_score_loss</code>	
'f1'	<code>metrics.f1_score</code>	for binary targets
'f1_micro'	<code>metrics.f1_score</code>	micro-averaged
'f1_macro'	<code>metrics.f1_score</code>	macro-averaged
'f1_weighted'	<code>metrics.f1_score</code>	weighted average
'f1_samples'	<code>metrics.f1_score</code>	by multi-sample
'neg_log_loss'	<code>metrics.log_loss</code>	requires
'precision' etc.	<code>metrics.precision_score</code>	suffixes apply as with 'f1'

Adapted from Géron, A. (2017); Russel, S., & Norvig, P. (2016)



<https://scikit-learn.org/>

# Your turn!

### Task

Please construct and draw a decision tree to predict “Buy = yes” with minimal leaf size 1.

- Try out the scikit-learn `DecisionTreeClassifier()` to train your own decision tree in Python

Brand	Model	Engine	Buy
Porsche	Taycan	Electric	y
Mercedes	S-Class	Gasoline	n
Tesla	Model 3	Electric	n
VW	Golf	Gasoline	n
Porsche	Cayenne	Hybrid	y
Renault	Zoe	Electric	n
Porsche	911	Gasoline	n

## 6.3 Machine Learning and Learning

- Machine Learning in a nutshell: Build a statistical model to make predictions about your dataset (the reality) with as few **errors** as possible
- The prediction errors of your final model can be decomposed into two main subcomponents:
  - **Error due to Bias:** difference between the expected prediction and the true value
  - **Error due to Variance:** variability of a model prediction for a given data point



What do you think: What is the relationship between error, bias and variance?

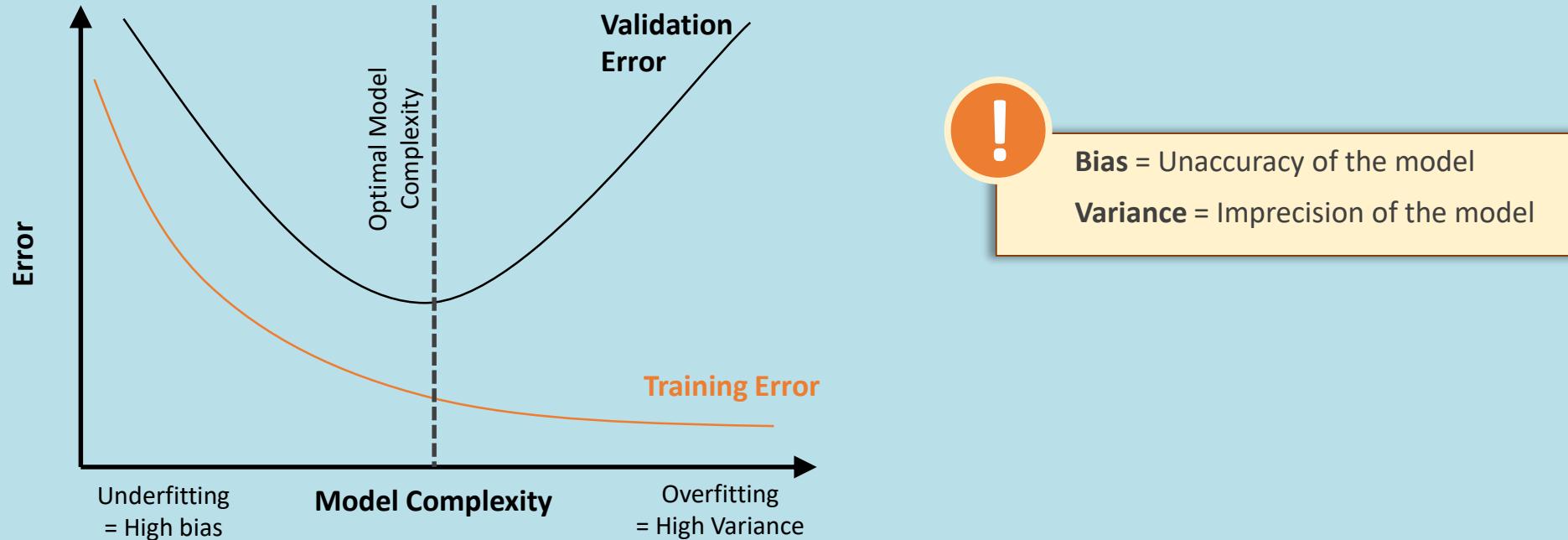
## 6.3 Fitting Dilemma

- We have seen bias and variance influence each other: the bias decreases with the complexity of the model, while the variance increases with the complexity of the model
- As a consequence, there are two opposite effects:
  - **Underfitting:** When the model has high bias and low variance, i.e. is too general (high total error)
  - **Overfitting:** When the model has low bias and high variance, i.e. is too specific (high total error)

Adapted from Géron, A. (2017); Russel, S., & Norvig, P. (2016)

## 6.3 Variance Bias Trade-off

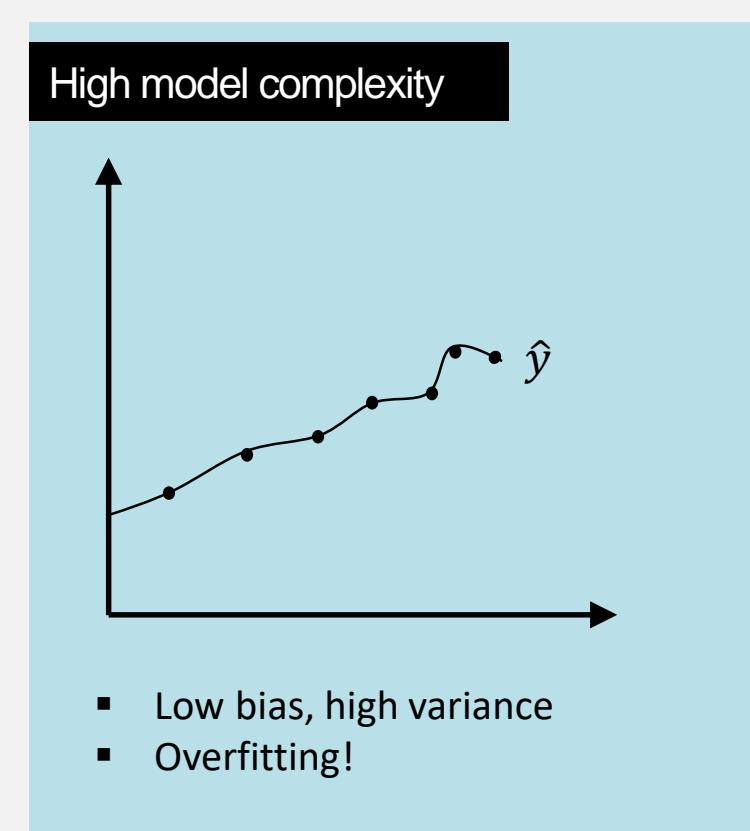
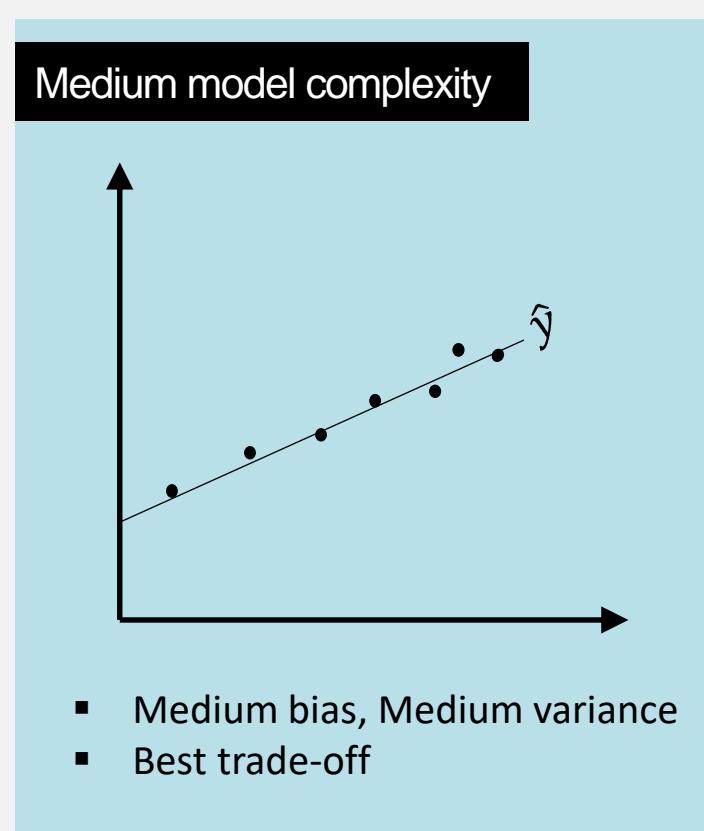
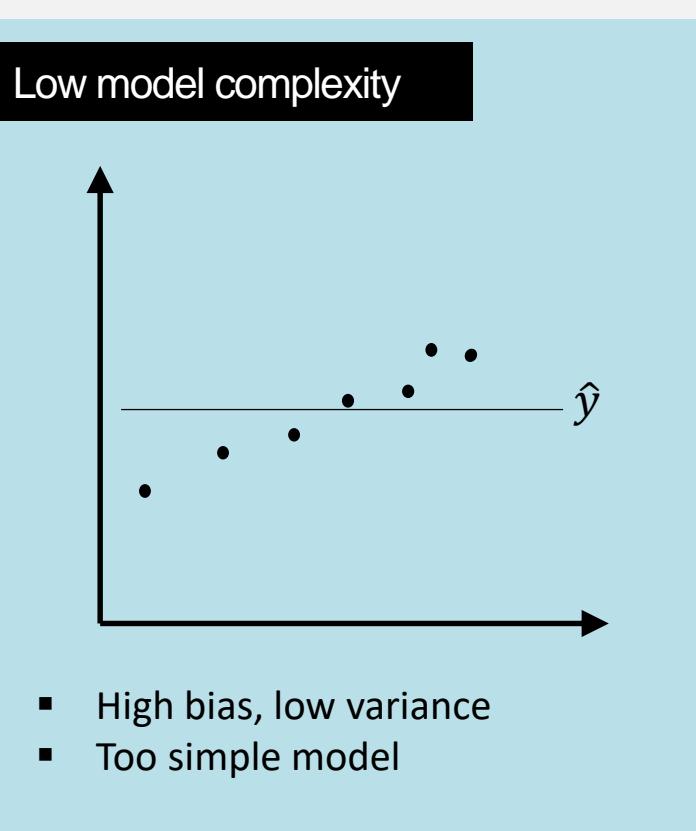
- We need to make a trade-off between “too specific” and “too general”



Adapted from Géron, A. (2017); Russel, S., & Norvig, P. (2016)

## 6.3 Bias, Variance and Model Complexity

- We can vary our model's complexity by adjusting the number of features, number of parameters, model type etc.



Adapted from Géron, A. (2017); Rusell, S., & Norvig, P. (2016)

## 6.3 Variance Bias Trade-off

- The error, we discussed on the previous slide deals with errors based on labelled observations. However, the precise value of our labels are often unknown.
- Real-world data is usually noisy, the “real” value is often influenced by its generation or measurement:

$$y = l(x) + \varepsilon$$

- There can be more observations with same attribute values but different labels or the other way round.



In this phase, we do not care where the noise  $\varepsilon$  comes from

## 6.3 From Modelling to Model Complexity

- We talked a lot about errors and increasing accuracy adjusting your model based on different performance measurement and sampling techniques, but there is also an other side...
- **Problem:** What do you do when you have two suitable models, and both of your models starts overfitting?
- **Solution:** Consider the model complexity in your evaluation of the model → Penalize too complexe models

## 6.3 Occam's Razor



Image source: [William of Ockham, from stained glass window at a church in Surrey](#) (2007) by Moscarlo from Wikimedia [CC-BY-SA-3.0](#)

*“Entia non sunt multiplicanda praeter necessitatem”*

- *More things should not be used than are necessary*

- If you have two models with the same generalization error, the simpler model (with fewer nodes, elements, predictors, etc.) is preferable.

Adapted from Géron, A. (2017); Russel, S., & Norvig, P. (2016)

## 6.3 Minimum Description Length Principle

- For each misclassification a measure is added to the binary coding to punish the complexity of the model.

$$cost(fit, data) = cost(data|fit) + cost(fit)$$

*Example: for 16 observations (4 bits) and 3 errors, cost (data | fit) = 3\*4 = 12*

- Mostly used for decision trees

## 6.3 Pessimistic Error Rate $e_g(T)$

- **Example:** Decision tree
- To the sum of all misclassifications  $e(t_i)$  at the leaf nodes above the training data one adds a malus ("penalty")  $\Omega(t_i)$  for each leaf node  $t_i$  in the tree and refers the result to the number of observations in the training data.



### Pessimistic Error Rate

$$e_g(T) = \frac{\sum_{i=1}^k [e(t_i) + \Omega(t_i)]}{\sum_{i=1}^k n(t_i)} = \frac{e(T) + \Omega(T)}{N_i}$$

Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

## 6.3 Regularization to reduced Overfitting

- Penalize to complex models
  - Example: linear regression
- Regularization term  $\Omega(w)$
- Used to enforce small weights:  $\Omega(w) = \|w\|_2^2$ ; many weights 0:  $\Omega(w) = \|w\|_1$
- Parameter  $\lambda$  controls strength of regularization

D

### Regularization

$$w^* = \arg \min_{w \in R^d} \sum_{i=1}^n \|w^T x_i - y_i\|_2^2 + \lambda \Omega(w)$$

Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

## 6.3 Parameter Tuning

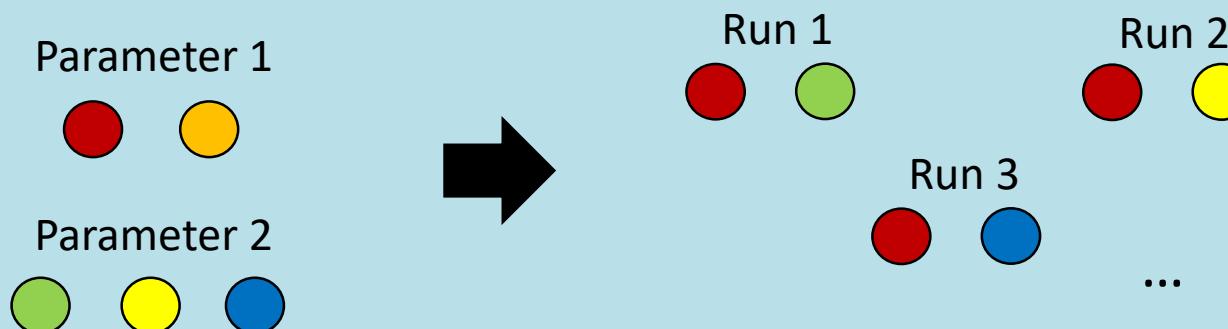
- Optimize your model's learning parameter to control the learning process (also known as hyperparameter) to reduce error.
- The most popular tuning methods are:
  - Random search
  - Grid search
  - Evolutionary search

## 6.3 Parameter Tuning with Random Search

- Compute  $n$  random samples of hyperparameters, repeat a fixed number of times
- Efficient in high dimensional feature spaces

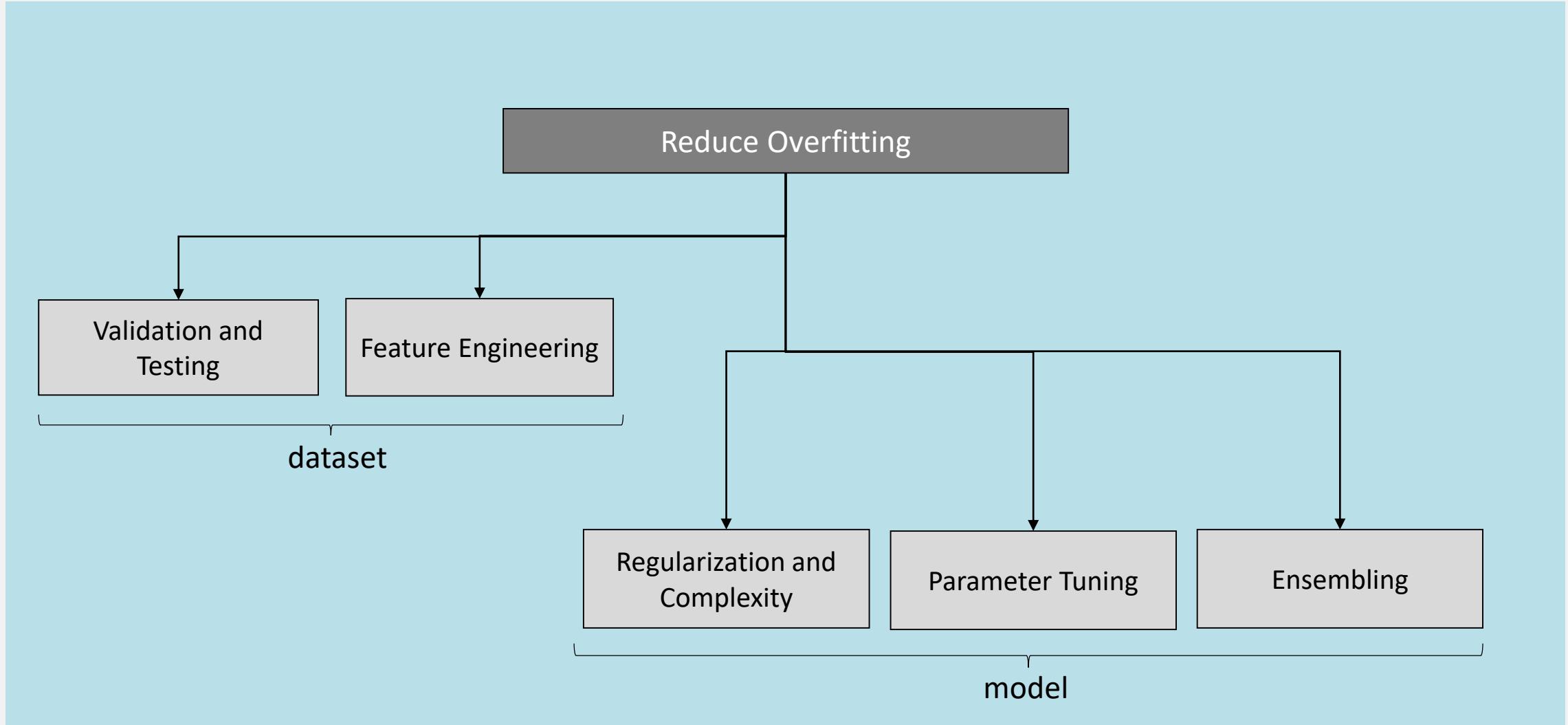
## 6.3 Parameter Tuning with Grid Search

- Simple, systematic search for optimal parameters
- Test all combinations for  $n$  parameters and  $k$  attributes in a specific range which has to be set manually
- Computational expensive, results in a total of  $n^k$  combinations



Adapted from Géron, A. (2017); Russel, S., & Norvig, P. (2016)

## 6.3 Summary: How to handle Overfitting



Adapted from Géron, A. (2017); Russel, S., & Norvig, P. (2016)

## 6.3 Ensemble Methods

- In the previous part, we discussed different methods for robust model validation
- In the next step we will take a look at different methods for robust model combination:
  - Bagging: Train independent model ensembles
  - Boosting: Train dependent model ensembles

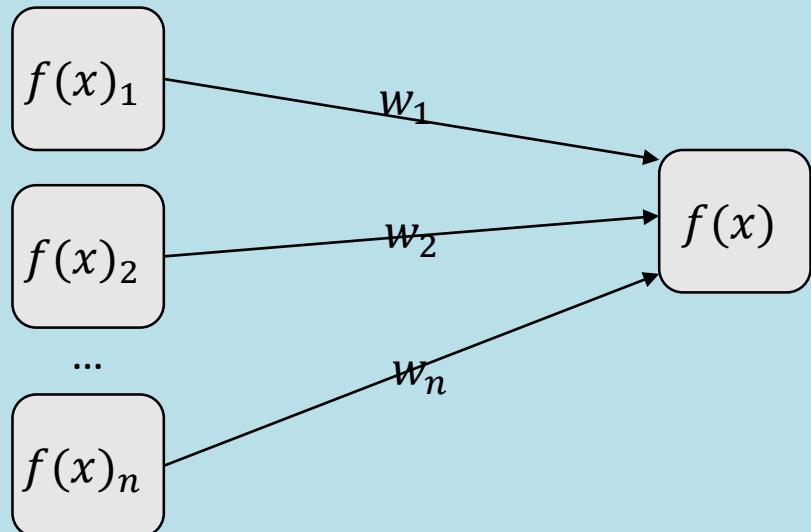
## 6.3 Ensemble Methods

Dietrich TG (2000)

### D

### Ensembling

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions (Dietrich TG, p.1, 2000)



- By combining the predictions from multiple models  $f(x)_1, \dots, f(x)_n$  together we can make more accurate predictions than any individual model  $f(x)_i$

Adapted from Géron, A. (2017); Russel, S., & Norvig, P. (2016)

### Training

- In each iteration
  - Randomly sample with replacement N samples from training set
  - Train a chosen “base” model (e.g. decision tree) on the samples

### Testing

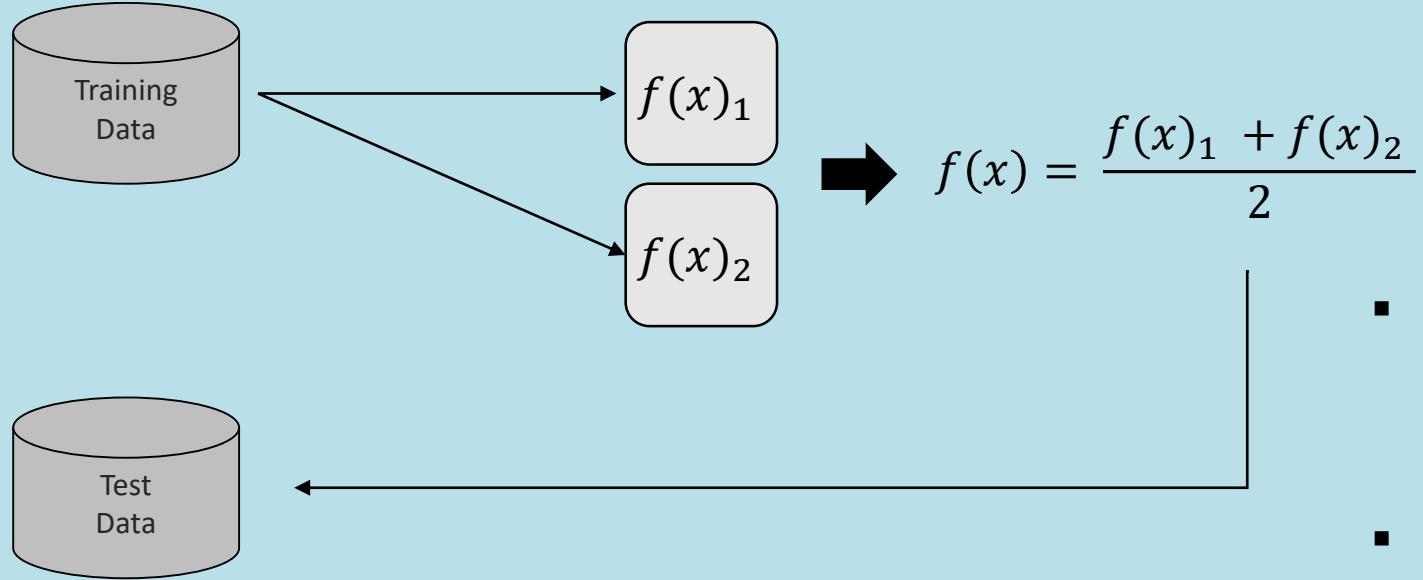
- For each test example
  - Start all trained base models
  - Predict by combining results of all trained models
  - Voting: Regression → Averaging, Classification → majority vote

### Training

- Train a sequence of  $T$  base models on  $T$  different sampling distributions defined over the training set ( $D$ )
  - A sample distribution  $D(t)$  for building the model  $t$  is constructed by modifying the sampling distribution  $D(t-1)$  from the  $(t-1)$ th step
  - Examples classified incorrectly in the previous step receive higher weights in the new data (attempts to cover misclassified samples)
- Classify according to the weighted majority of classifiers

## 6.3 Example: Random forest algorithm with $i = 2$

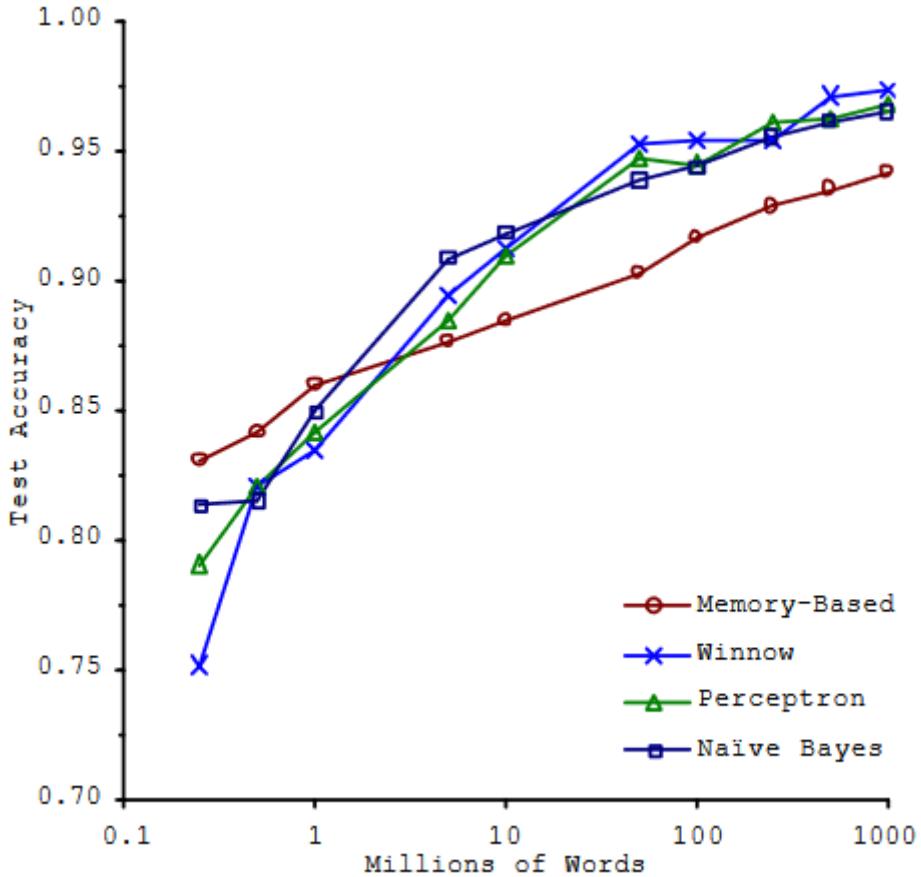
Dietrich TG (2000)



- The random forest algorithm makes use of the concept of ensemble learning
- It predicts by averaging the different learners (ensemble of weak learners)

Adapted from Géron, A. (2017); Russell, S., & Norvig, P. (2016)

## 6.3 The Unreasonable Effectiveness of Data



Alon Halevy, Peter Norvig, and Fernando Pereira, Google

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"<sup>1</sup> examines why so much of physics can be neatly explained with simple mathematical formulas

such as  $f = ma$  or  $e = mc^2$ . Meanwhile, sciences that involve human beings rather than elementary particles have proven more resistant to elegant mathematics. Economists suffer from physics envy over their inability to neatly model human behavior. An informal, incomplete grammar of the English language runs over 1,700 pages.<sup>2</sup> Perhaps when it comes to natural language processing and related fields, we're doomed to complex theories that will never have the elegance of physics equations. But if that's so, we should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data.

One of us, as an undergraduate at Brown University, remembers the excitement of having access to the Brown Corpus, containing one million English words.<sup>3</sup> Since then, our field has seen several notable corpora that are about 100 times larger, and in 2006, Google released a trillion-word corpus with frequency counts for all sequences up to five words long.<sup>4</sup> In some ways, this corpus is a step backwards from the original Brown Corpus, taken from unfiltered Web pages complete sentences, spelling errors, and all sorts of other re-



The unreasonable Effectiveness of Data:

Online available from ↗ [Institute of Electrical and Electronics Engineers \(ieeexplore\)](#)

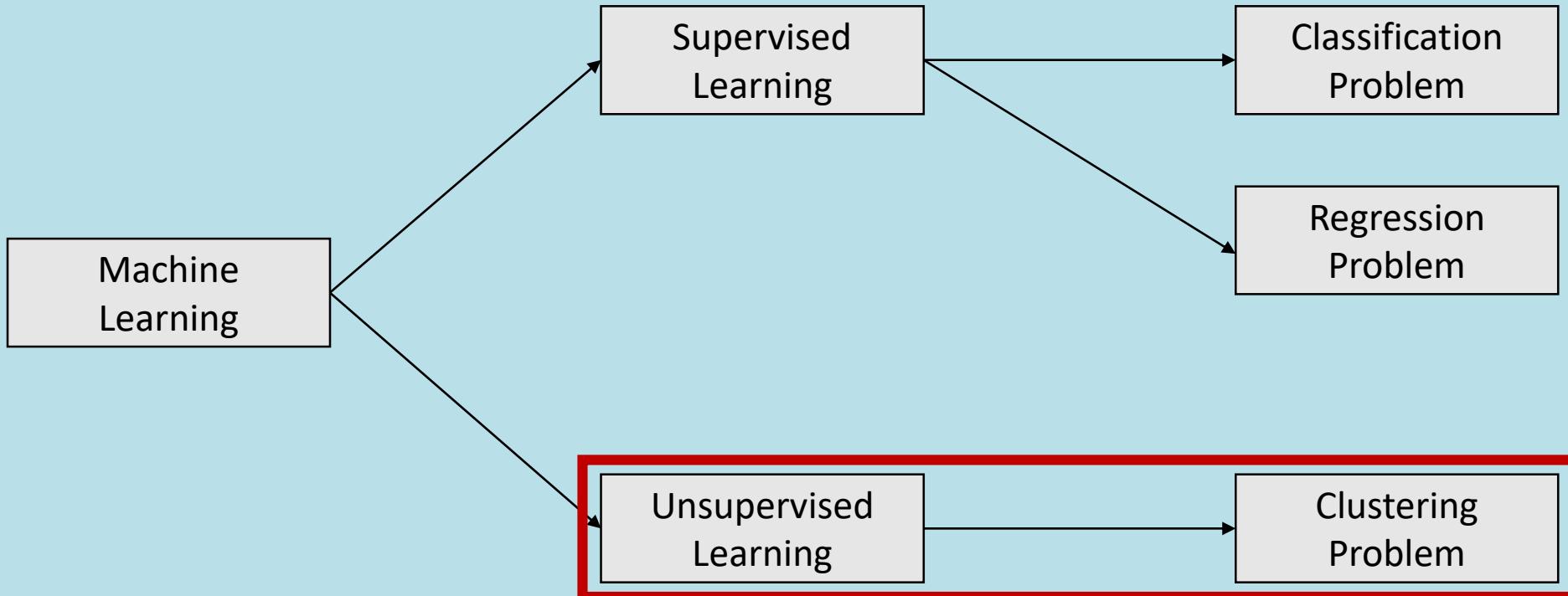
Adapted from Banko, M., & Brill, E. (2001, July); Halevy, A., Norvig, P., & Pereira, F. (2009); Russell, S., & Norvig, P. (2016) |Image source: ko, M., & Brill, E. (2001, p. 2);

# Your turn!

### Task

One of your colleagues says “I heart that boosting relays on dependent model ensembles. But I don’t get it why!”. Is this true? Why are the models dependent from each other?

## 6.4 Problem Types in Machine Learning (High-Level)



## 6.4 What is clustering about?

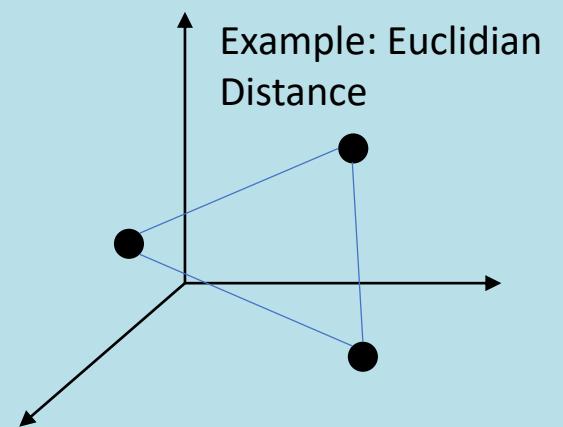
- Grouping a set of data objects into clusters
- Cluster: a collection of data objects, similar to one another within the same cluster, and dissimilar to the objects in other clusters
- Clustering = unsupervised “classification” (no predefined classes)
- Typical usage in analytics:
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

## 6.4 What is Similarity?



D

- **Nonnegativity:**  $\forall xy: d(x, y) \geq 0$
- **Symmetry:**  $d(x, y) = d(y, x); s(x, y) = s(y, x)$
- **Triangle inequality:**  $d(x, z) \leq d(x, y) + d(y, z)$



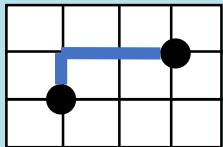
## 6.4 Most Popular Metric: Minkowski Distance

$$d(x, y) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{\frac{1}{r}}$$

- Generalization of Euclidian Distance

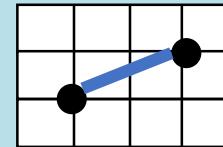
### Mannhattan Distance

- $r = 1$
- Named Cityblock distance, or  $L_1$ -norm
- For binary attributes also named Hamming Distance
- Sum of distance about all dimensions 1...n



### Euclidian Distance

- $r = 2$
- $L_2$ -norm
- For numeric variables (ordinal or rational)
- Describes the geometric distance between two points



### Supremum

- $r = \infty$
- $L_\infty$ -norm
- Biggest difference in the dimensions

## 6.4 Difference L1 and L2 Norm

- $L_1$ : Impact of a difference proportional to difference itself
- $L_2$ : Higher relative impact of larger distances!
  - Example:  $x = (3,3), y = (4,5)$
  - L1 norm distance:  $|3 - 4| + |3 - 5| = 1 + 2 = 3$
  - L2 norm distance:  $\sqrt{|3 - 4|^2 + |3 - 5|^2} = \sqrt{1 + 4} = 2.2$
- $L_2$  norm is smaller than  $L_1$  but the individual difference has more relative weight

## 6.4 Beyond Minkowski

### Binary Objects

- Simple Matching Coefficient (SMC)
- Jaccard Coefficient

### Correlation

- Pearson
- Spearman

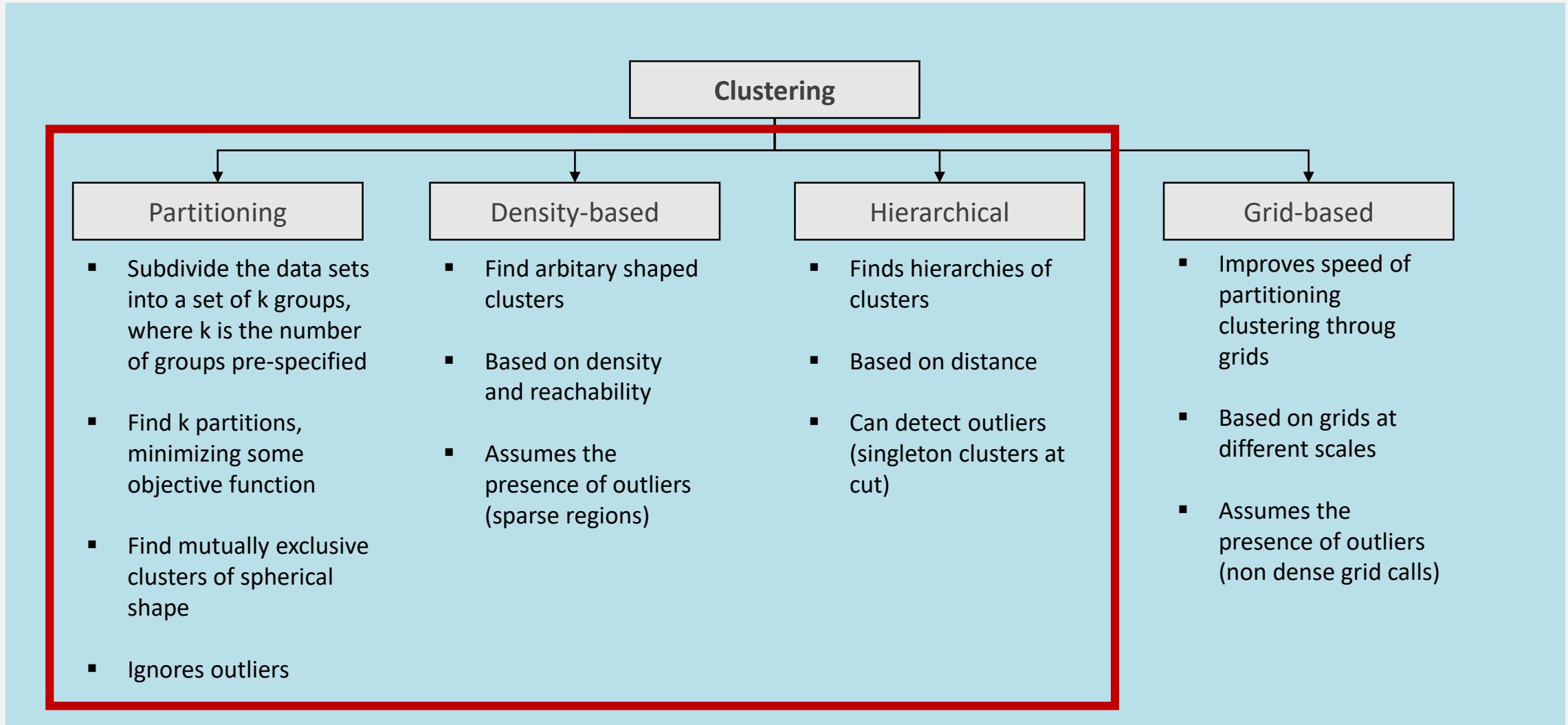
### Vectors

- Simple Matching
- Cosinus Coefficient
- Dice Coefficient
- Tanimoto Coefficient
- Overlap Coefficient



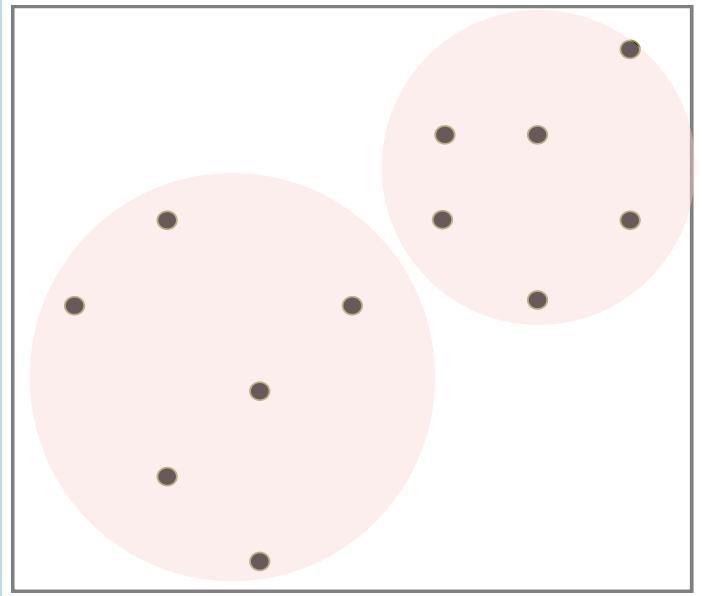
You find a comprehensive overview of further similarity concepts at „Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions”

## 6.4 Fundamental Clustering Approaches



## 6.4 Partitioning Clustering: K-Means

- ▶ Forming groups of objects in a way that the same group (clusters) are more similar to each other than to those in other groups



### Example

- Cluster different log-files etc. for further analysis

- Most popular clustering technique today
- Based on centroids, clusters instances based on the closest centroid
- Generally used for exploratory rather than confirmatory analysis

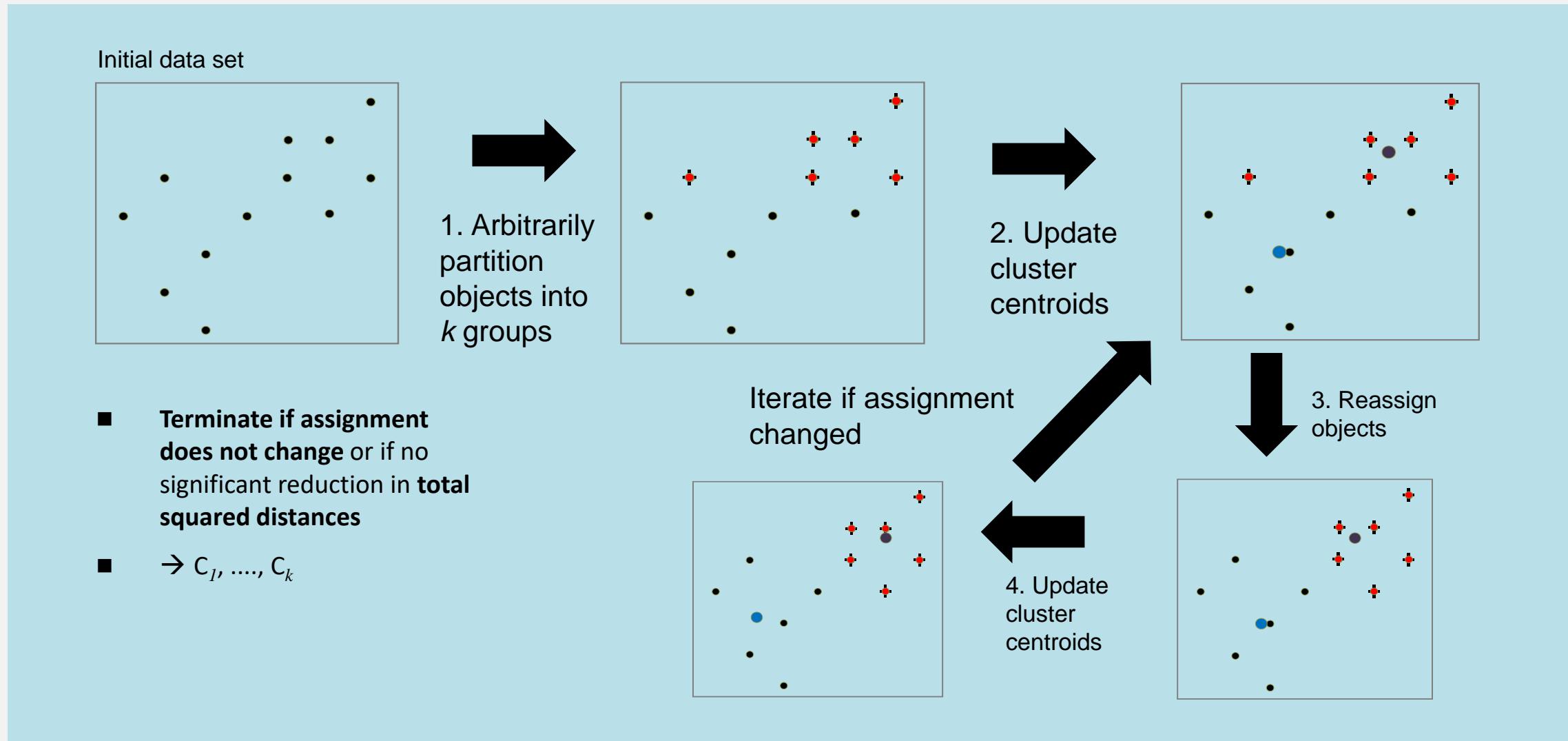


Only one parameter ( $k$ ), execution is very fast, resulting clusters are convex



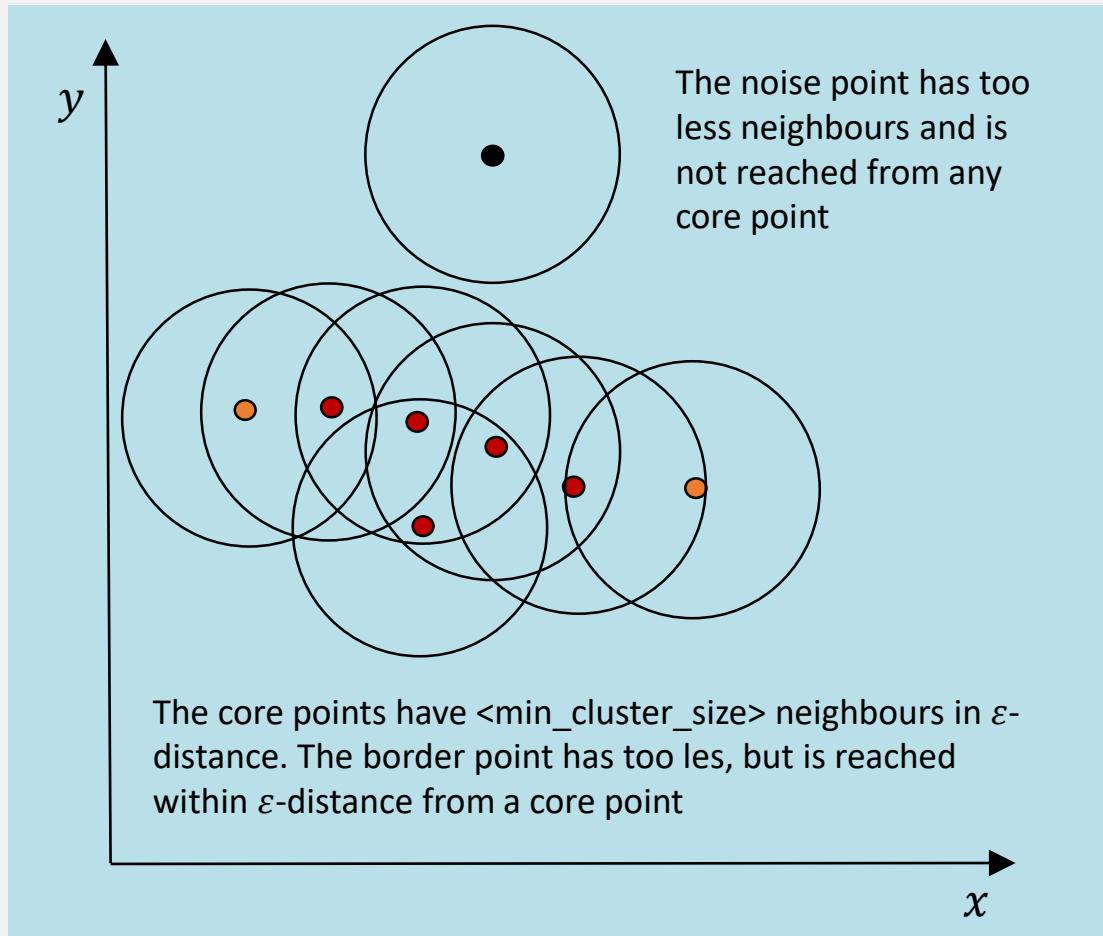
Applicable only when mean is defined, outliers have a strong influence on the result

## 6.4 K-means procedure (with k = 2)



## 6.4 Density-based Clustering

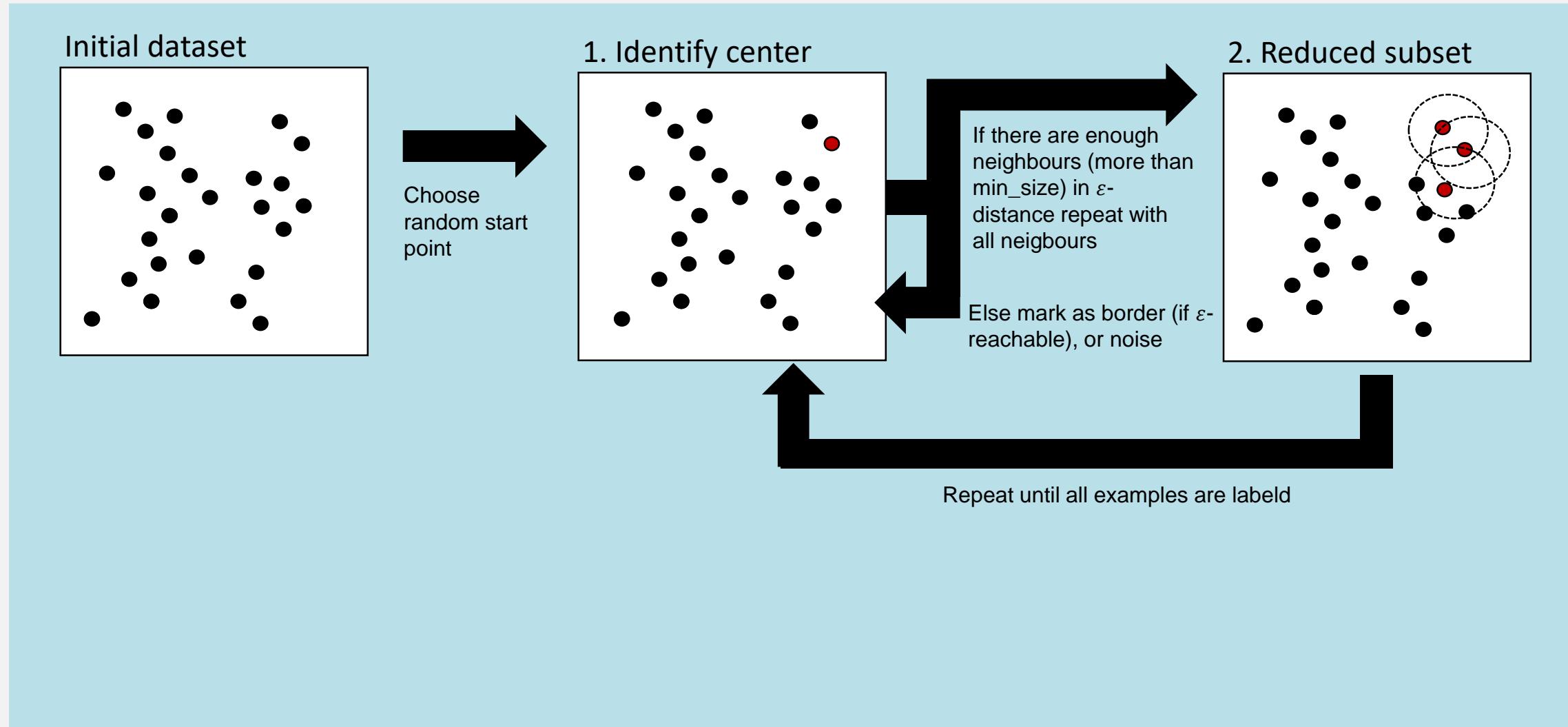
► Find arbitrary shaped clusters based on density and reachability



- Based on  $\varepsilon$ -reachability. A point is  $\varepsilon$ -reachable from another one, if they lie not more than  $\varepsilon$ -distance away
- There are core points, border points and noise points

- + Number of clusters can be unspecified. Clusters are arbitrary-shaped. Can handle noise (assumes there is), hence very outlier resistance
- Clusters can not have different densities, parameter tuning for  $\varepsilon$  is difficult

## 6.4 Density-based Clustering procedure

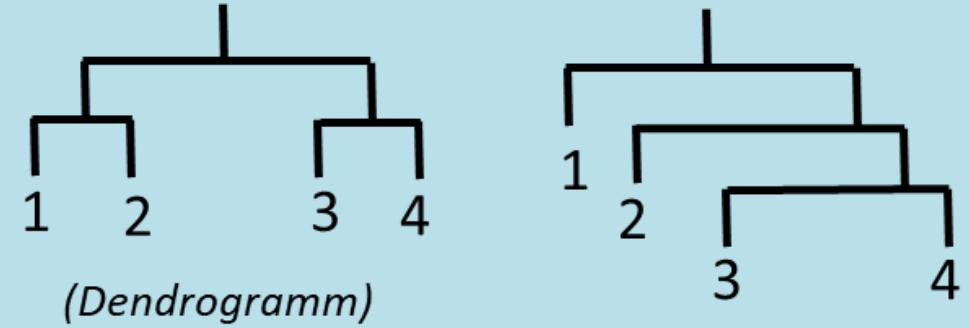


## 6.4 Hierarchical Clustering

### ► Create a hierarchical decomposition of the set of objects

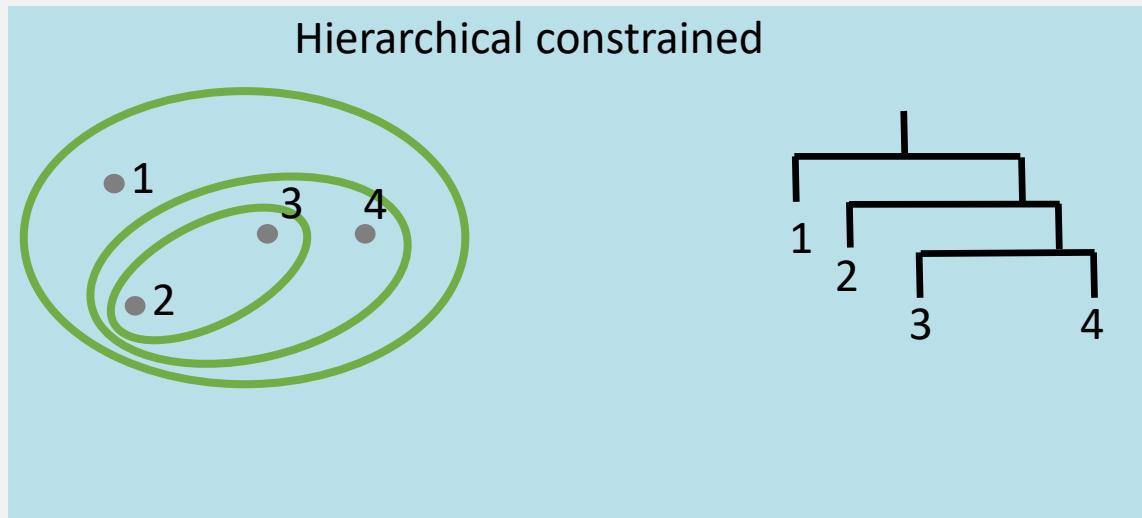
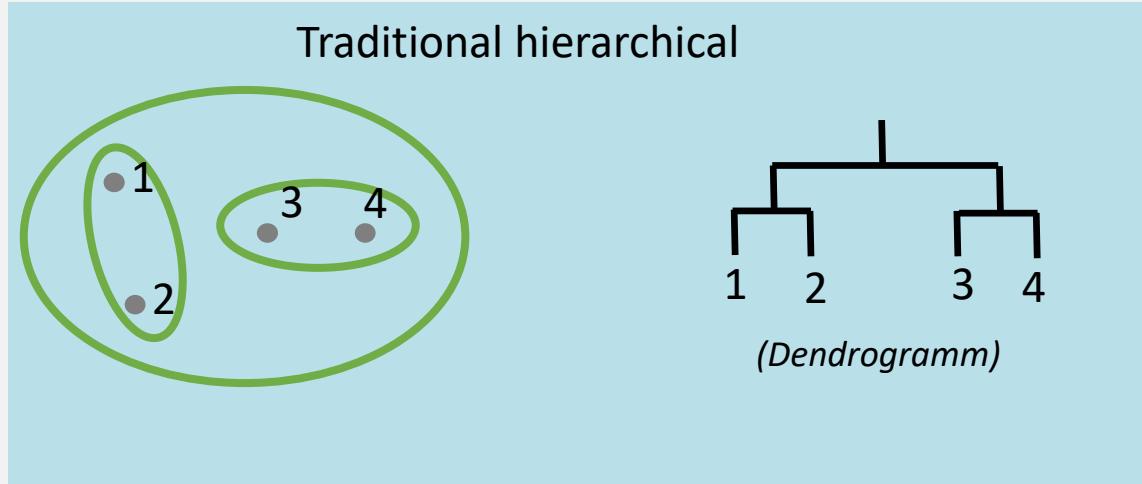
Method of cluster analysis which seeks to build a hierarchy of clusters

- **Agglomerative:** "bottom up" approach; each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy
- **Divisive:** "top down" approach; all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy



- + Do not have to assume any particular number of clusters, any desired number of clusters can be obtained by 'cutting' the dendrogram
- Can not handle complex shapes, new points can not be added without recalculation of the model, single-link-effect ("chains" depending on the metric)

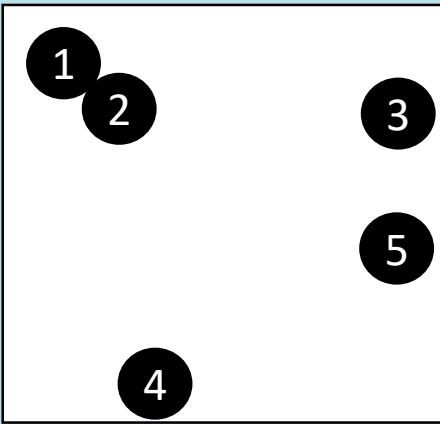
## 6.4 Traditional hierarchical vs. Hierarchical constrained



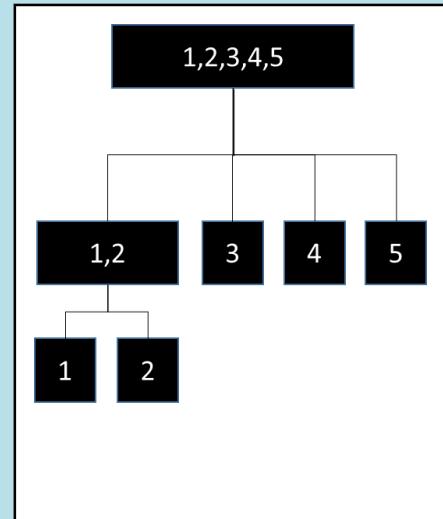
- Produces a set of nested clusters organized as hierarchical tree
- Can be visualized as a dendrogram

## 6.4 Agglomerative Clustering

Initial dataset

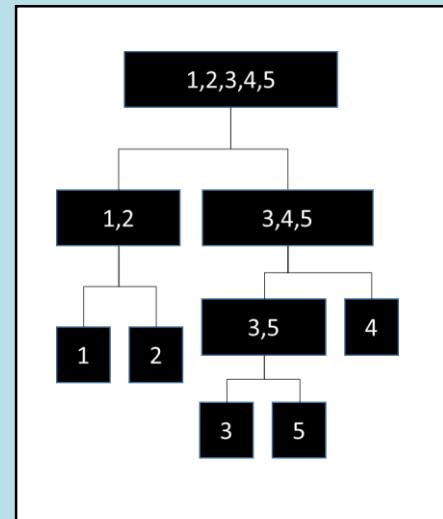


Define distance measure and put every observation in a single cluster



Until only one cluster left:

1. Select the two clusters with minimum distance
2. Join these two clusters



## 6.4 Linkage Criteria

### Single

- Minimal distance between two elements of each cluster

$$D(X, Y) := \min_{x \in X, b \in B} \{d(x, y)\}$$

### Complete

- Maximum distance between two elements of each cluster

$$D(X, Y) := \max_{x \in X, b \in B} \{d(x, y)\}$$

### Average

- Average distance between two elements of each cluster

$$D(X, Y) := \frac{1}{|X||Y|} \sum_{x \in X, b \in B} \{d(x, y)\}$$

### Ward

- Growth of cluster variance

$$D(X, Y) := \frac{d(\bar{a}, \bar{b})^2}{\frac{1}{|A|} + \frac{1}{|B|}}$$

$\bar{a}, \bar{b}$  are cluster centroids

## 6.4 Evaluation of Clustering Models: Silhouette-Value

Silhouette Value [-1,1]:

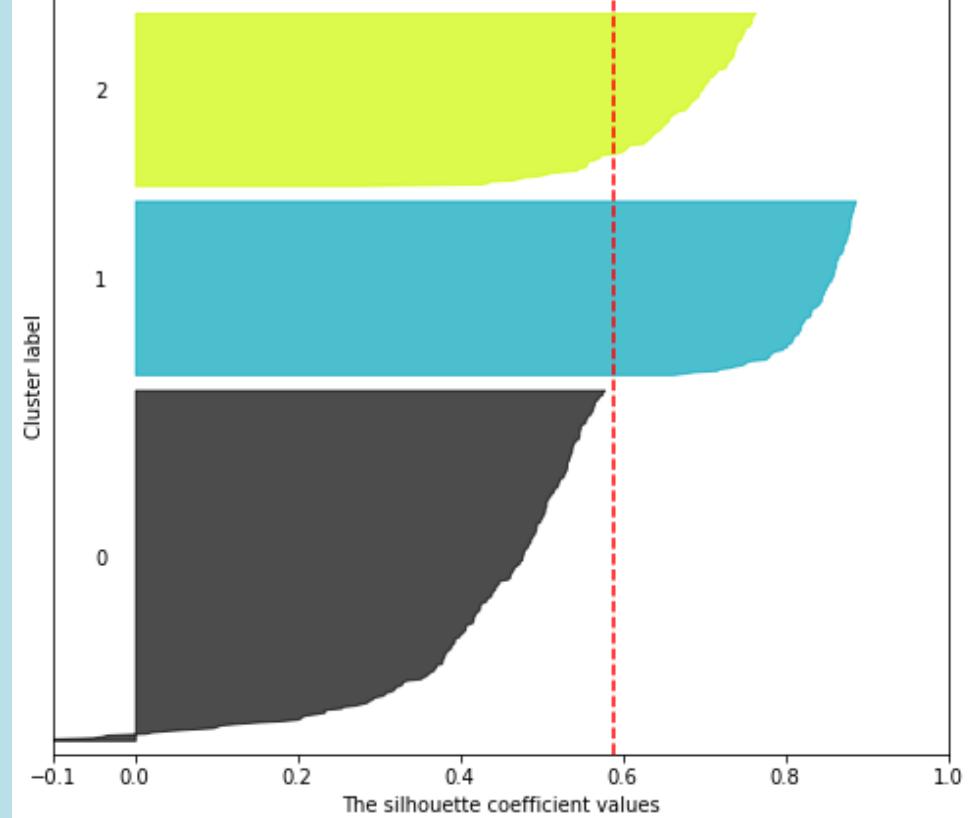
$$S(v_k) = \frac{b(v_k) - a(v_k)}{\max\{a(v_k), b(v_k)\}}$$

D

- $a(v_k)$ : mean dissimilarity of the silhouette value to all other elements of the same clusters
- $b(v_k)$ : mean dissimilarity of the silhouette value to all other elements to the nearest cluster
- A negative silhouette value indicates that an element would better fit into another cluster
- The greater the value the better the model

```
Import sklearn  
silhouette_score(x, cluster_labels)
```

The silhouette plot for the various clusters.

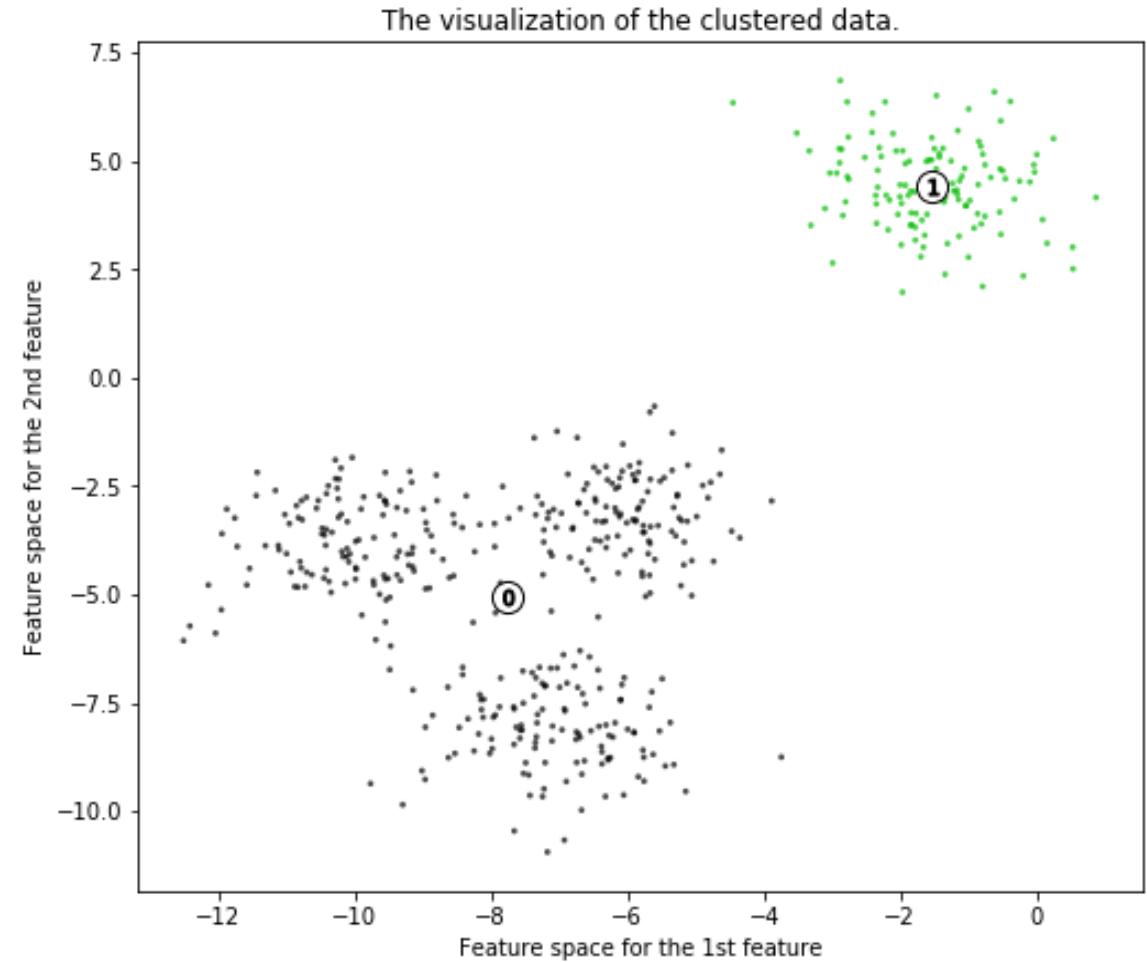
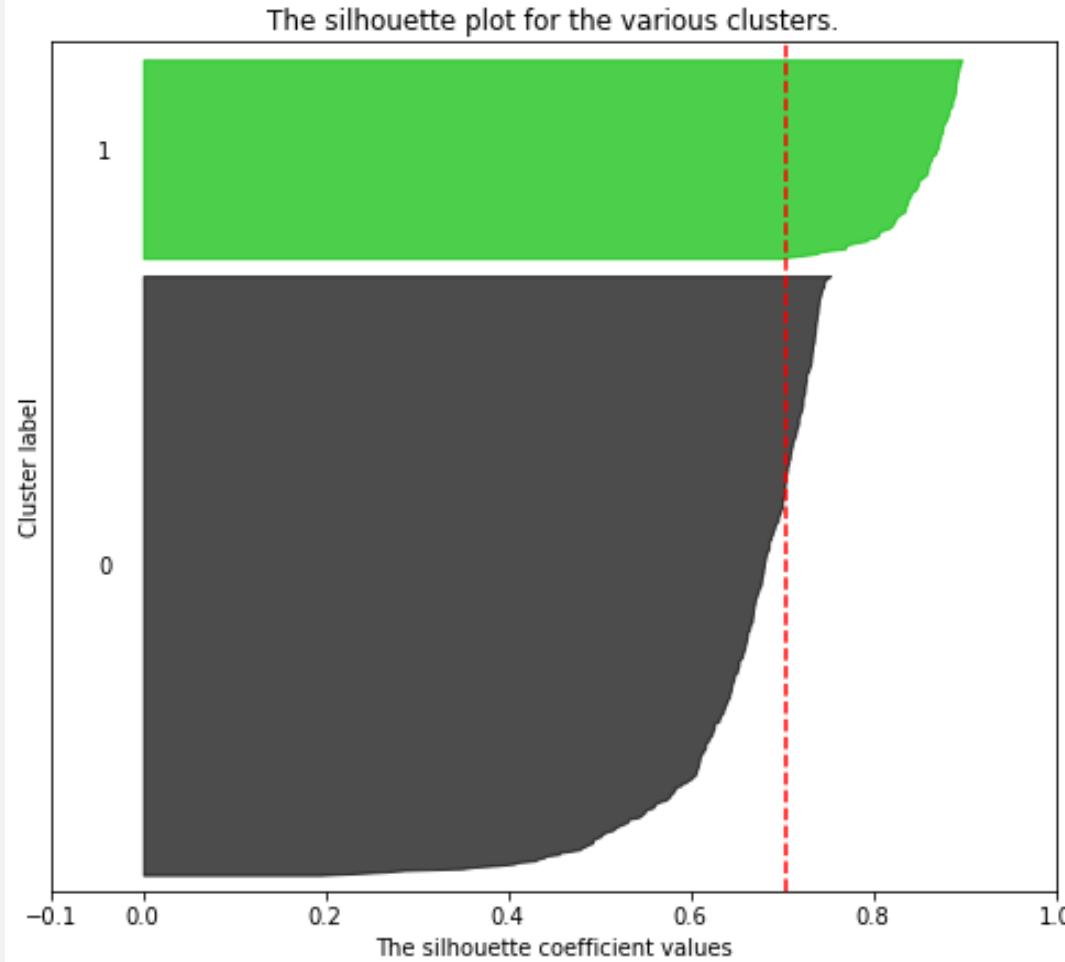


Silhouette plot: Visualizes the silhouettes for all data and the mean value

## 6.4 Compare Silhouette-Values with Python

For n\_clusters = 2 The average silhouette\_score is : 0.7049787496083262

**Silhouette analysis for KMeans clustering on sample data with n\_clusters = 2**

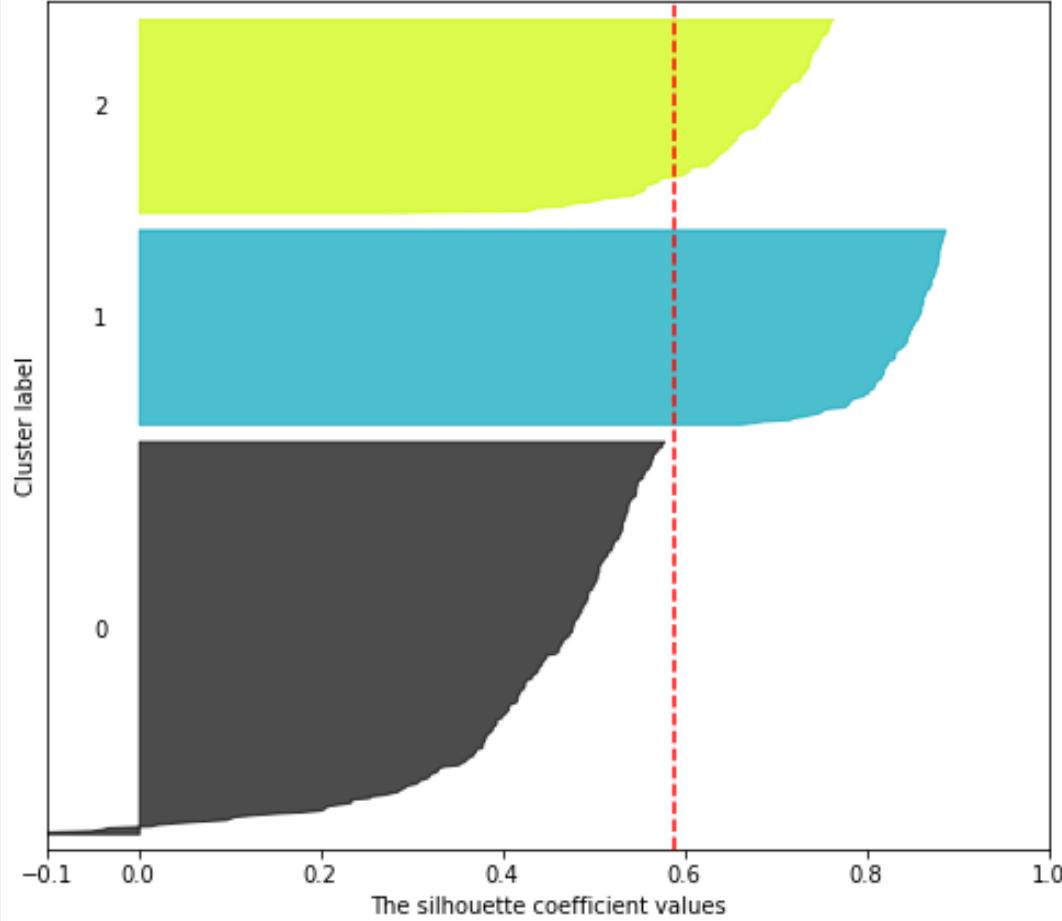


## 6.4 Compare Silhouette-Values with Python

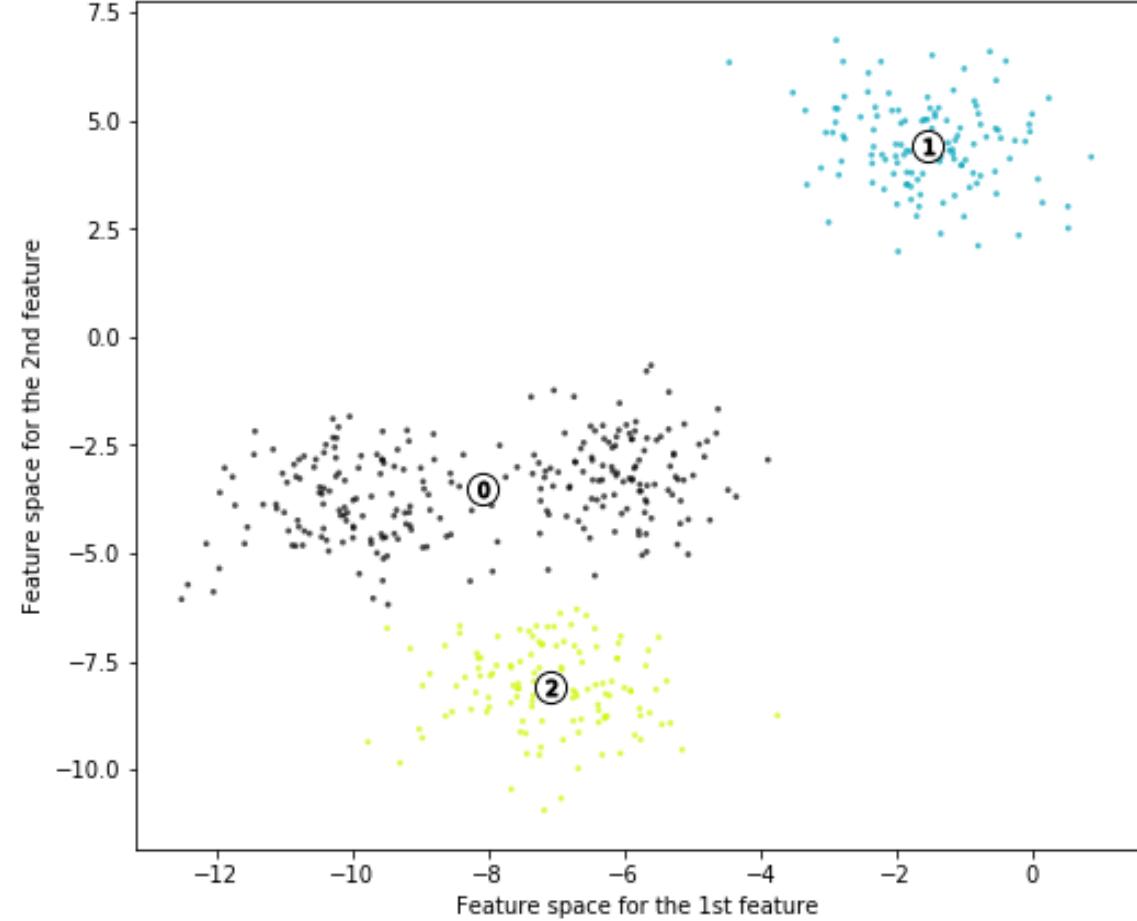
For n\_clusters = 3 The average silhouette\_score is : 0.5882004012129721

**Silhouette analysis for KMeans clustering on sample data with n\_clusters = 3**

The silhouette plot for the various clusters.



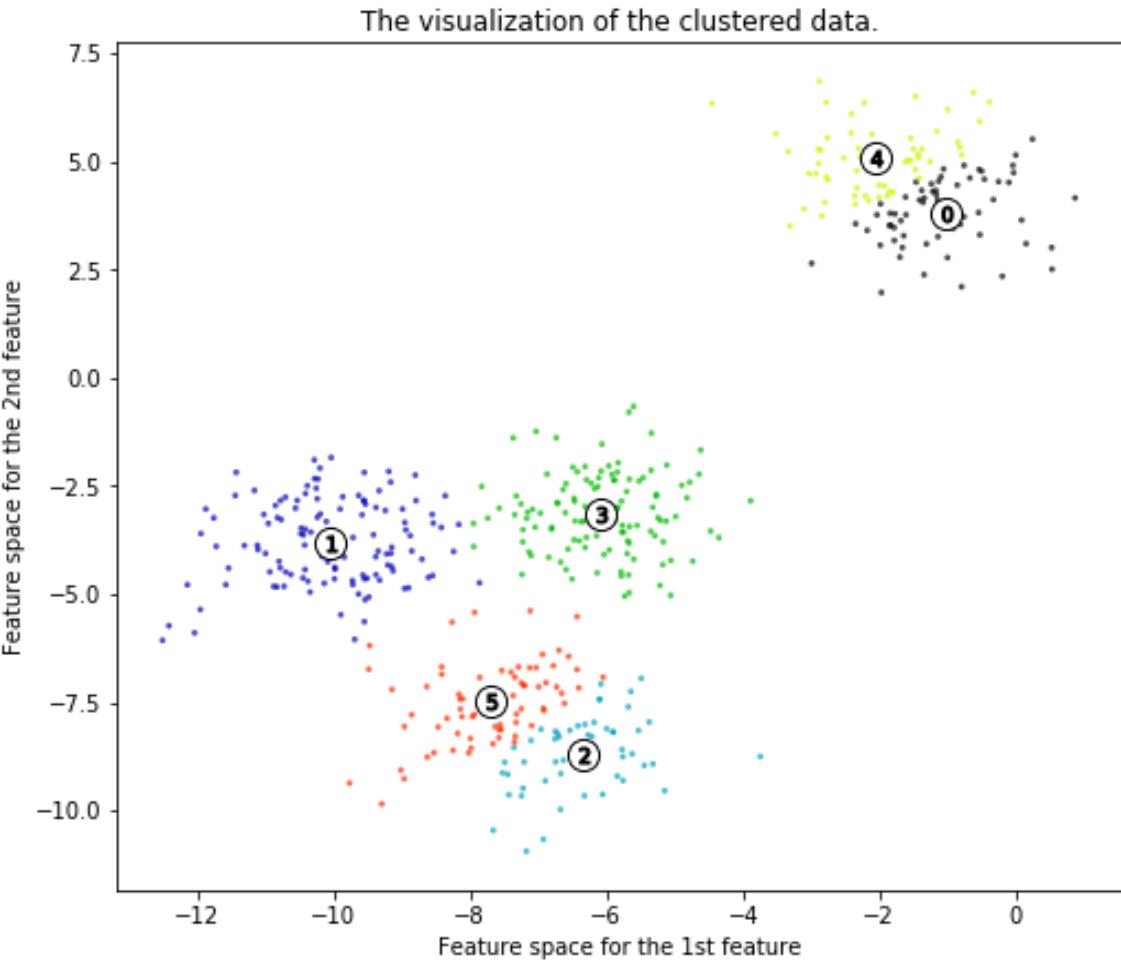
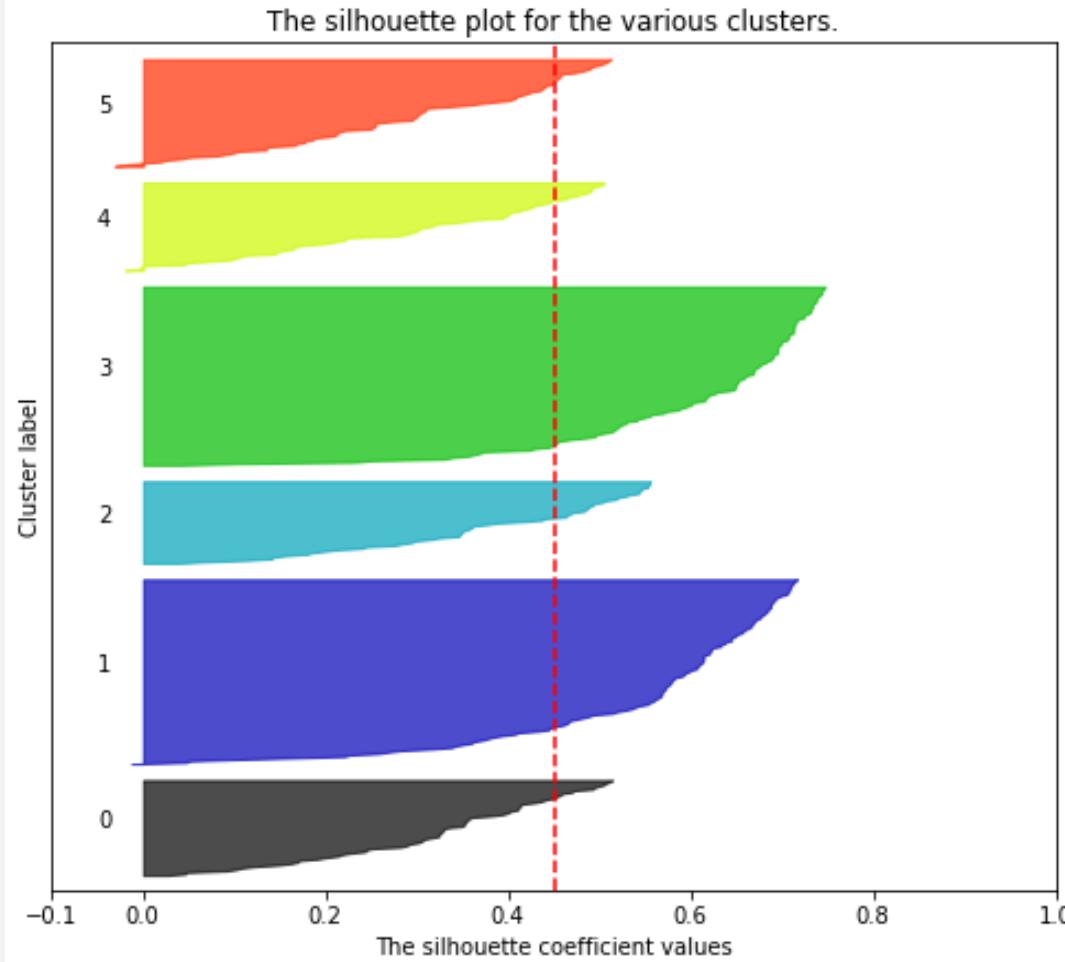
The visualization of the clustered data.



## 6.4 Compare Silhouette-Values with Python

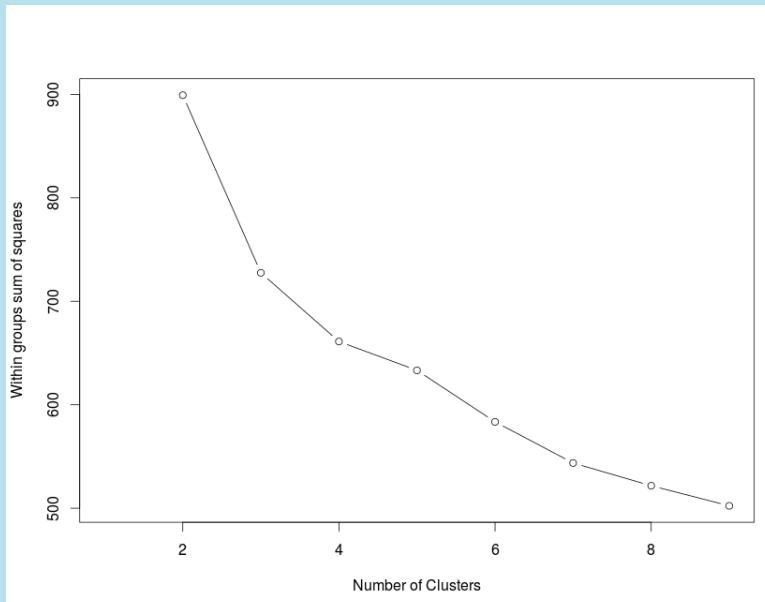
For n\_clusters = 5 The average silhouette\_score is : 0.56376469026194

**Silhouette analysis for KMeans clustering on sample data with n\_clusters = 6**



## 6.4 Evaluation of Clustering Models: Elbow criterion

- ▶ Choose a number of clusters so that adding another cluster doesn't give much better modeling of the data



What could be possible error measures to compute the quality of a parameterization?

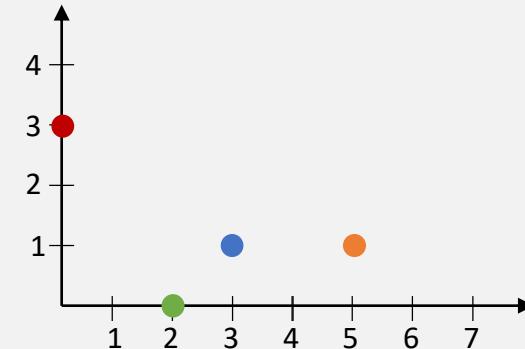
- Run k-means clustering on the dataset for a range of values of k (say, k from 1 to 10 in the examples above)
- For each value of k calculate the sum of squared errors (SSE)
- Plot a line chart of the SSE for each value of k
- If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best

# Your turn!

**Task**

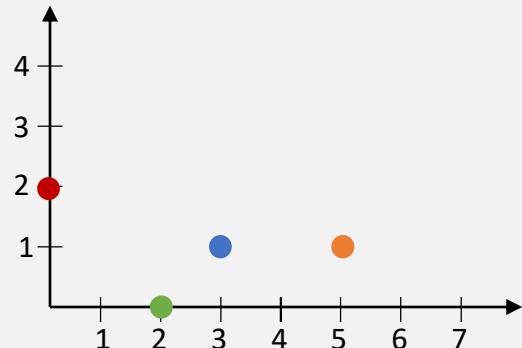
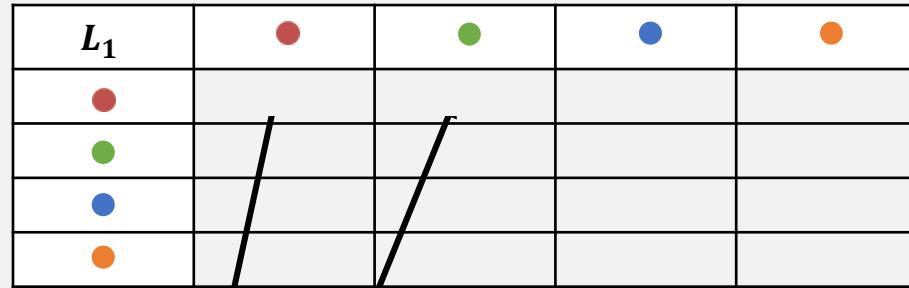
Please compute the Mankowski distances ( $L_1$ ,  $L_2$ , and  $L_3$ ) for the following dataset:

Observation	X	Y
●	0	2
●	2	0
●	3	1
●	5	1

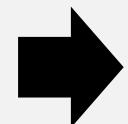


## 6.4 Classroom Task

Observation	X	Y
●	0	2
●	2	0
●	3	1
●	5	1



$$d(x, y) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{\frac{1}{r}}$$

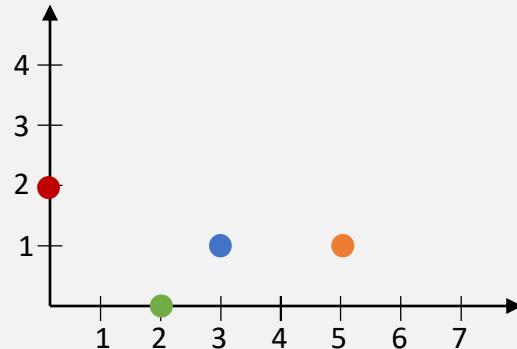


$$d(x, y) = ((|0 - 0|^1 + |2 - 2|^1))^{\frac{1}{1}}$$

$$d(x, y) = (|0 - 2| + |2 - 0|) = 2 + 2 = 4$$

## 6.4 Classroom Task

Observation	X	Y
●	0	2
●	2	0
●	3	1
●	5	1



$$d(x, y) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{\frac{1}{r}}$$

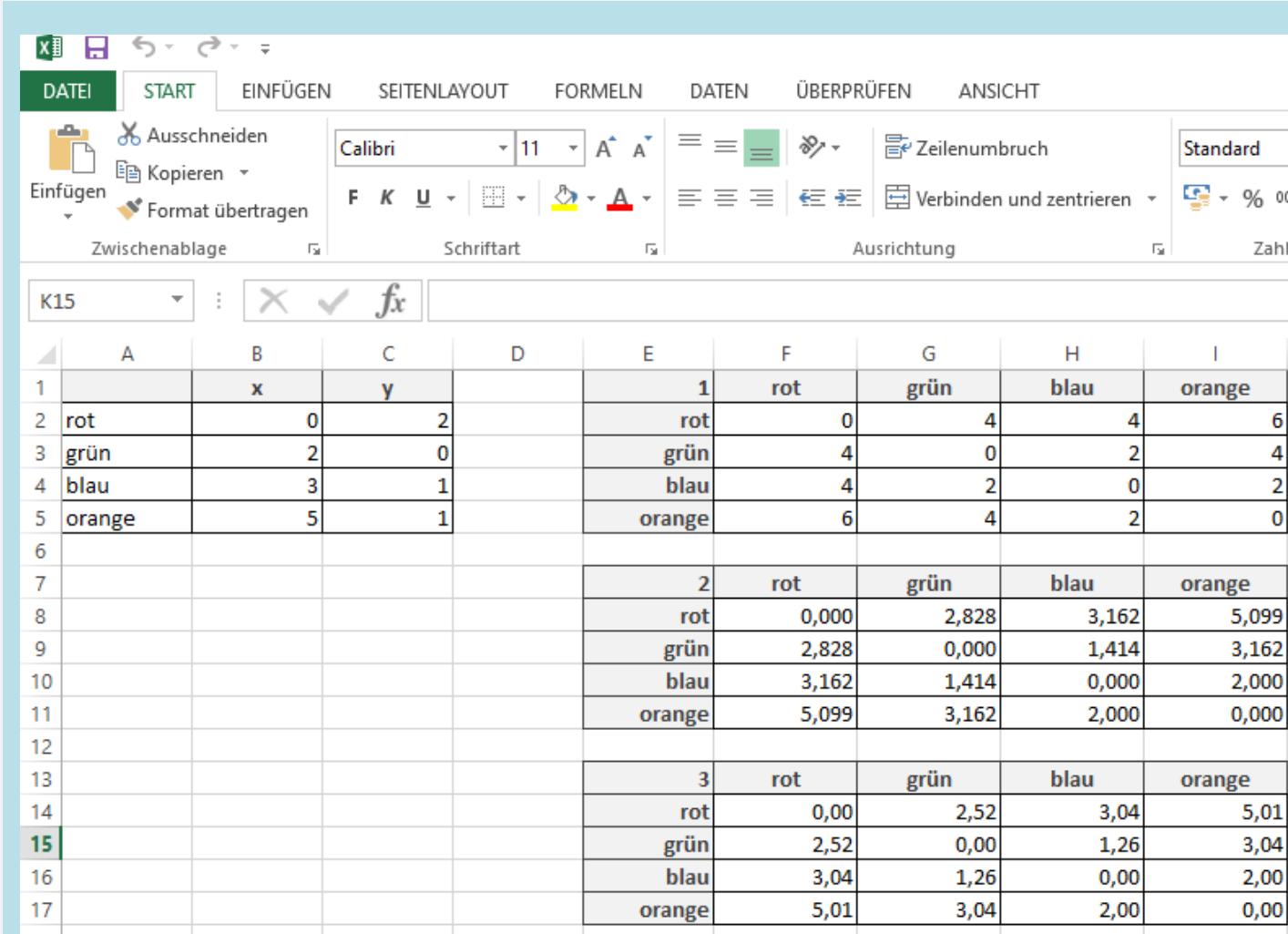
$L_1$	●	●	●	●
●	0	4	4	6
●	4	0	2	4
●	4	2	0	2
●	6	4	2	0

$L_2$	●	●	●	●
●	0	3	3	5
●	3	0	1	3
●	3	1	0	2
●	5	3	2	0

$L_3$	●	●	●	●
●	0	2	3	5
●	2	0	1	3
●	3	1	0	2
●	5	3	2	0

\*Results are rounded

## 6.4 Mankowski Distance Exercise Sheet



The screenshot shows an Excel spreadsheet with data and formulas for calculating Mankowski distances. The data is organized into three main sections:

- Section 1 (Rows 1-5):** Contains categorical data for four points (rot, grün, blau, orange) across two dimensions (x and y).

	A	B	C
1	x	y	
2	rot	0	2
3	grün	2	0
4	blau	3	1
5	orange	5	1
- Section 2 (Rows 6-11):** Contains the same data as Section 1, but with numerical values for each dimension.

	A	B	C	
1	rot	grün	blau	orange
2	rot	0	4	6
3	grün	4	0	4
4	blau	4	2	2
5	orange	6	4	0
- Section 3 (Rows 12-17):** Contains the calculated Mankowski distances between all pairs of points for various values of p (1, 2, 3).

	2	rot	grün	blau	orange
8	rot	0,000	2,828	3,162	5,099
9	grün	2,828	0,000	1,414	3,162
10	blau	3,162	1,414	0,000	2,000
11	orange	5,099	3,162	2,000	0,000

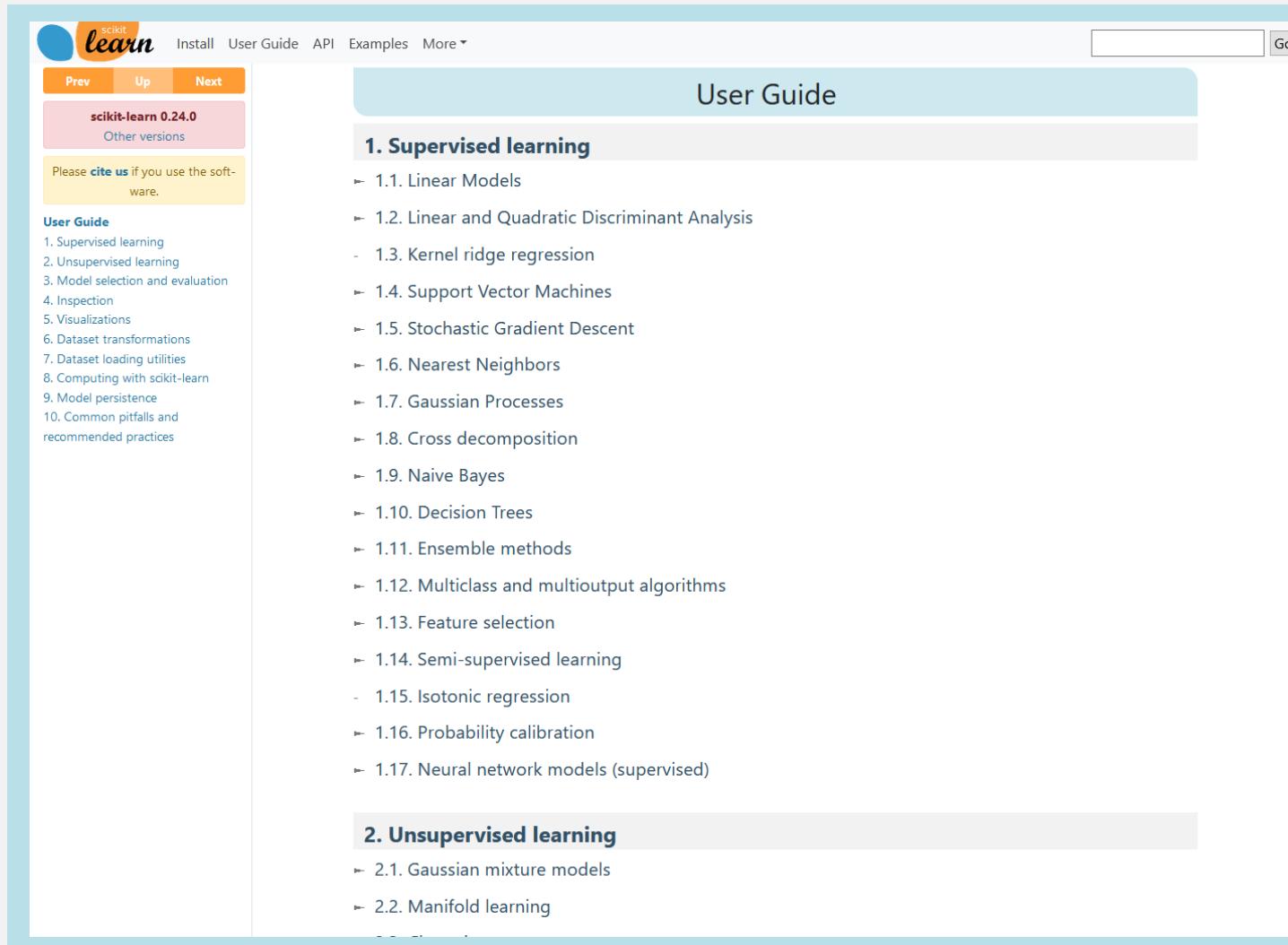
	3	rot	grün	blau	orange
14	rot	0,00	2,52	3,04	5,01
15	grün	2,52	0,00	1,26	3,04
16	blau	3,04	1,26	0,00	2,00
17	orange	5,01	3,04	2,00	0,00

- Still unsecure with the Mankowski distance computation? Check out the excel exercise sheet!



↗ Code/Lecture 6 - Compute Mankowski distances.xlsx

# 6. Scikit-Learn Userguide



The screenshot shows the scikit-learn User Guide website. The top navigation bar includes links for Install, User Guide, API, Examples, and More. A search bar with a Go button is also present. On the left, there's a sidebar with the scikit-learn logo, version information (scikit-learn 0.24.0), and a citation note. The main content area is titled "User Guide" and contains two sections: "1. Supervised learning" and "2. Unsupervised learning". The "Supervised learning" section lists 17 sub-topics from 1.1. Linear Models to 1.17. Neural network models (supervised). The "Unsupervised learning" section lists 2 sub-topics: 2.1. Gaussian mixture models and 2.2. Manifold learning.

Check out the scikit-learn user guide  
for more data and feature engineering  
algorithms (↗ [here](#)).

# 6. Exercises

## Workbook Exercises

- Please read the chapter 18.1-18.6, 18.8, 18.11, 19 and 21 from Russel, S., & Norvig, P. (2016) to understand the theory behind the concepts of this lecture. Then work through the exercises of each chapter. You can skip the parts about „neuronal networks“ and „probabilistic models“, we will handle these topics in the next chapter.
- Read the chapters 1, 3, 4, 6 and 7 of Géron, A. (2017) to recapitulate the algorithms and applications of machine learning. Solve the related exercises of each chapter.

## Coding Exercises

- Read the chapter 2 „End-to-End Machine Learning Project“ from Géron, A. (2017). Try to implement the project for yourself and solve the related coding exercises of this chapter.
- Read the userguide ([↗here](#)) from scikit-learn. Recapitulate the previous chapters and match them with the related chapters of the user guide.

## 6. References

### Literature

1. Banko, M., & Brill, E. (2001, July). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics* (pp. 26-33).
2. Beleites, C.; Baumgartner, R.; Bowman, C.; Somorjai, R.; Steiner, G.; Salzer, R. & Sowa, M. G. (2005): Variance reduction in estimating classification error using sparse datasets, Chemom Intell Lab Syst, 79, 91 - 100.
3. Bellman, R. E. (1957): Dynamic Programming. Princeton University Press.
4. Bruner, J. S., & Austin, G. A. (1986). *A study of thinking*. Transaction publishers.
5. Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Springer, Berlin, Heidelberg.
6. Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*.
7. Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), 8-12.
8. Kohavi, R. (1995): A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, Mellish, C. S. (ed.) In *Artificial Intelligence Proceedings 14 th International Joint Conference*, 20 -- 25. August 1995, Montréal, Québec, Canada, Morgan Kaufmann, USA, , 1137 - 1145.
9. Kim, J.-H. (2009): Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap , Computational Statistics & Data Analysis , 53, 3735 - 3745.

## 6. References

### Literature

10. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22.
11. Mitchell, T. M. (1997). Machine learning. McGraw Hill.
12. Russell, S., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Global Edition.

### News articles

1. Wired (2017): Artificial Intelligence Is About to Conquer Poker, But Not Without Human Help. Online available at <https://www.wired.com/2017/01/ai-conquer-poker-not-without-human-help/>.

### Images

All images that were not marked other ways are made by myself, or licensed ↗CC0 from ↗Pixabay.

## 6. Glossary

<b>Clustering</b>	<i>Grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters)</i>
<b>Confusion Matrix</b>	<i>Specific table layout to visualize errors in classification tasks</i>
<b>Error due to Bias</b>	<i>Difference between the expected prediction and the true value</i>
<b>Error due to Variance</b>	<i>Variability of a model prediction for a given data point</i>
<b>Hypothesis Space</b>	<i>The space of conjunctive hypotheses represented by a vector of <math>n</math> constraints</i>
<b>Machine Learning</b>	<i>A computer program is said to learn from experience ‘E’, with respect to some class of tasks ‘T’ and performance measure ‘P’ if its performance at tasks in ‘T’ as measured by ‘P’ improves with experience ‘E’. (Mitchell, 1997)</i>
<b>Metric</b>	<i>Function that defines the distance between each pair of points of a set</i>
<b>Overfitting</b>	<i>When the model has low bias and high variance, i.e. is too specific (high total error)</i>
<b>Reinforcement Learning</b>	<i>Learning by using a system of reward and punishment</i>

## 6. Glossary

**Supervised Learning** *Learning by mapping an input to an output based on example input-output pairs*

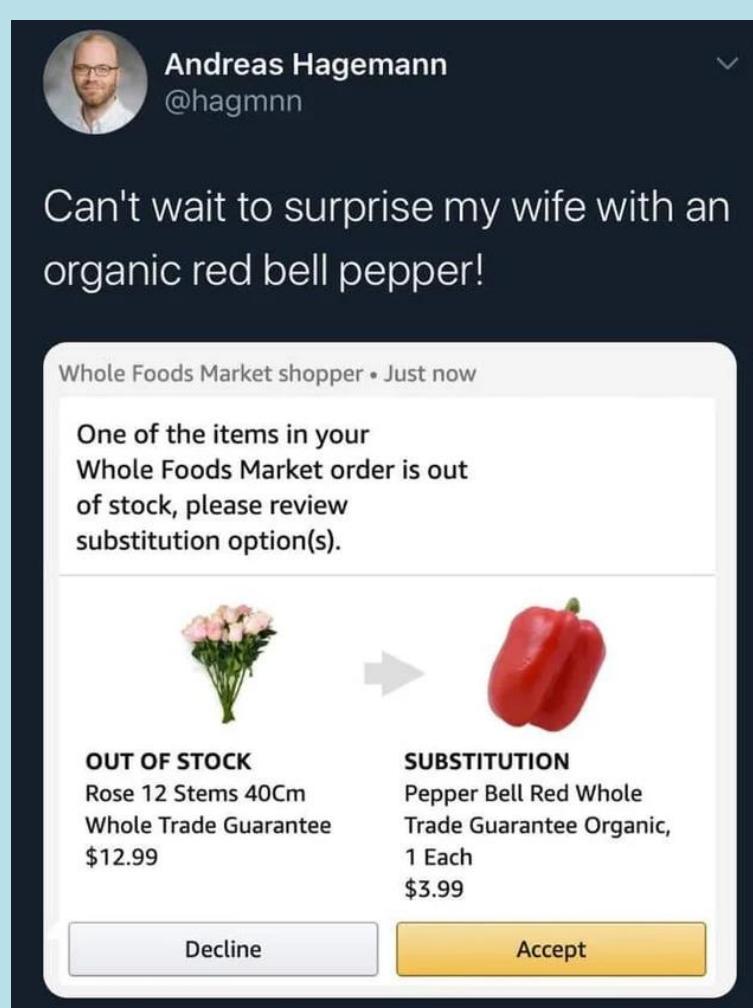
**Test set** *Data set to test the performance of the learning model*

**Train set** *Data set to train the learning model*

**Underfitting** *When the model has high bias and low variance, i.e. is too general (high total error)*

**Unsupervised Learning** *Learning undetected patterns in a data set with no pre-existing labels*

## 6. When Machine Learning Fails...



# Information Gain

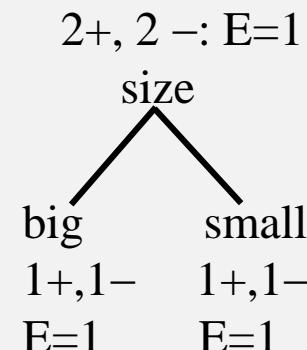
- The information gain of a feature  $F$  is the expected reduction in entropy resulting from splitting on this feature.

$$Gain(S, F) = Entropy(S) - \sum_{v \in Values(F)} \frac{|S_v|}{|S|} Entropy(S_v)$$

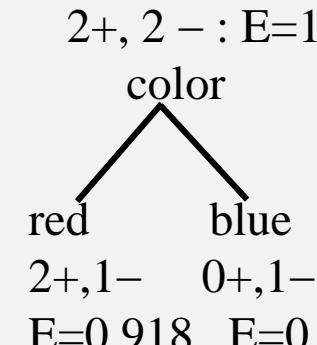
where  $S_v$  is the subset of  $S$  having value  $v$  for feature  $F$ .

- Entropy of each resulting subset weighted by its relative size.
- Example:

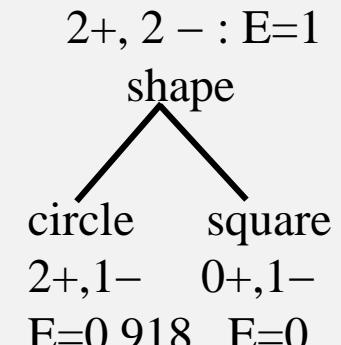
- <big, red, circle>: +      <small, red, circle>: +
- <small, red, square>: -      <big, blue, circle>: -



$$Gain = 1 - (0.5 \cdot 1 + 0.5 \cdot 1) = 0$$



$$Gain = 1 - (0.75 \cdot 0.918 + 0.25 \cdot 0) = 0.311$$



$$Gain = 1 - (0.75 \cdot 0.918 + 0.25 \cdot 0) = 0.311$$