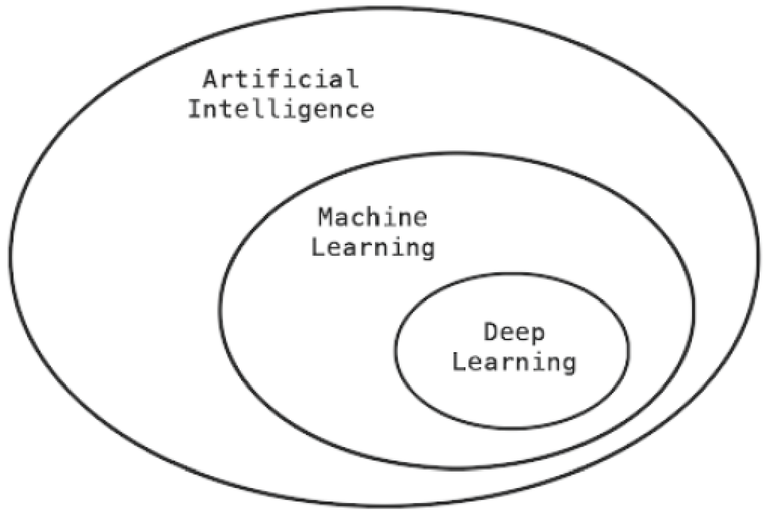
MACHINE LEARNING

Session 8 - Hertie‘s data society

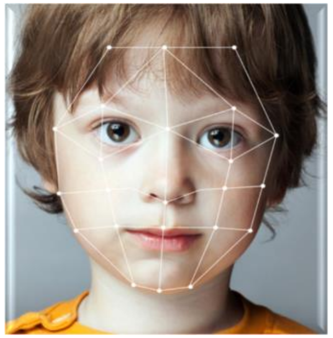
MACHINE LEARNING IS A SUBFIELD OF AI

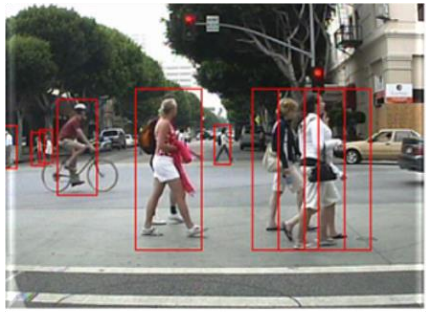
- **Subfield of AI** that aims at building models automatically with the help of training data
- **Learning** is the process where data is used to create a model that recognises patterns
- These learnt patterns can be used for analysing **unknown data**

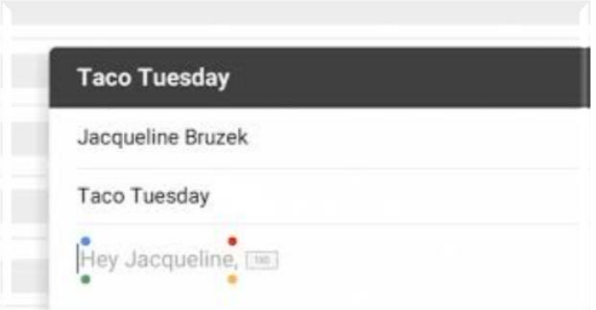


AREAS OF APPLICATION FOR MACHINE LEARNING

Today:







Tomorrow:

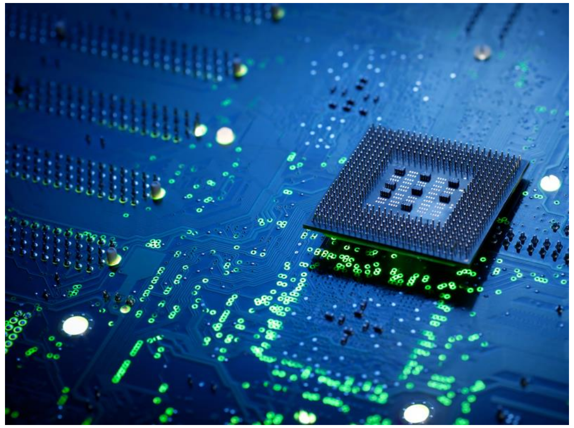
MORE DATA AND MORE COMPUTING POWER MAKE MACHINE LEARNING SUCCESSFUL TODAY

Big Data

St Peter‘s Place 2005

St Peter‘s Place 2013

Computing Power



Exponential growth: Doubling of computing power every ~18 months since the 1960s

SUPERVISED AND UNSUPERVISED LEARNING

Supervised learning:

- Data needs to be labelled
- Relationship between training inputs and training targets is mapped and you can measure how well it works
- E.g. dog and cat photos knowing on which photo is which type of animal



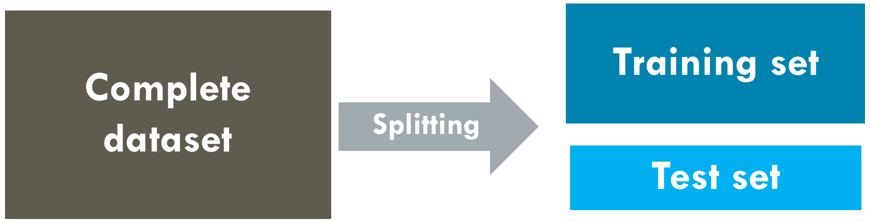

Unsupervised learning

- No need for labelled data
- Data is mapped e.g. according to a measurement likeness (clustering)
- E.g. photos of people not knowing who is on which photo

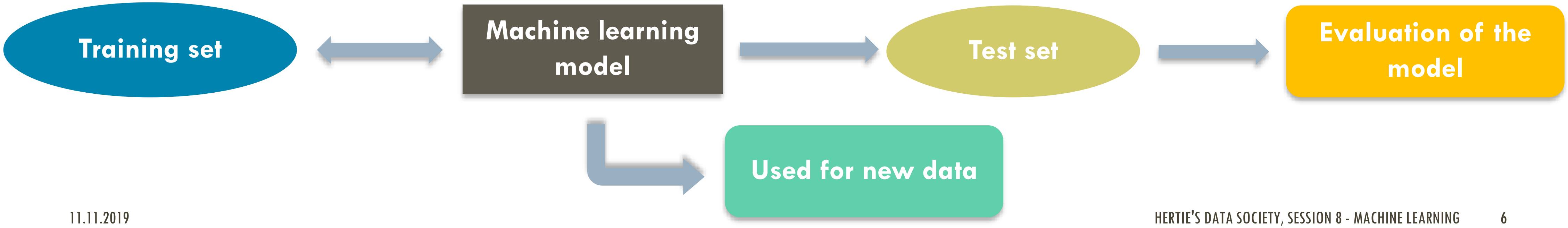
THE GOAL OF MACHINE LEARNING MODELS IS PREDICTION FOR UNKNOWN DATA

- **Machine learning models** focus on the quality of their predictions while causation and interpretability often are less important (e.g. „how accurate can I predict income on the basis of education data?“)

- For that, we split the data into a **training** and a **test dataset** (e.g. 80-20)



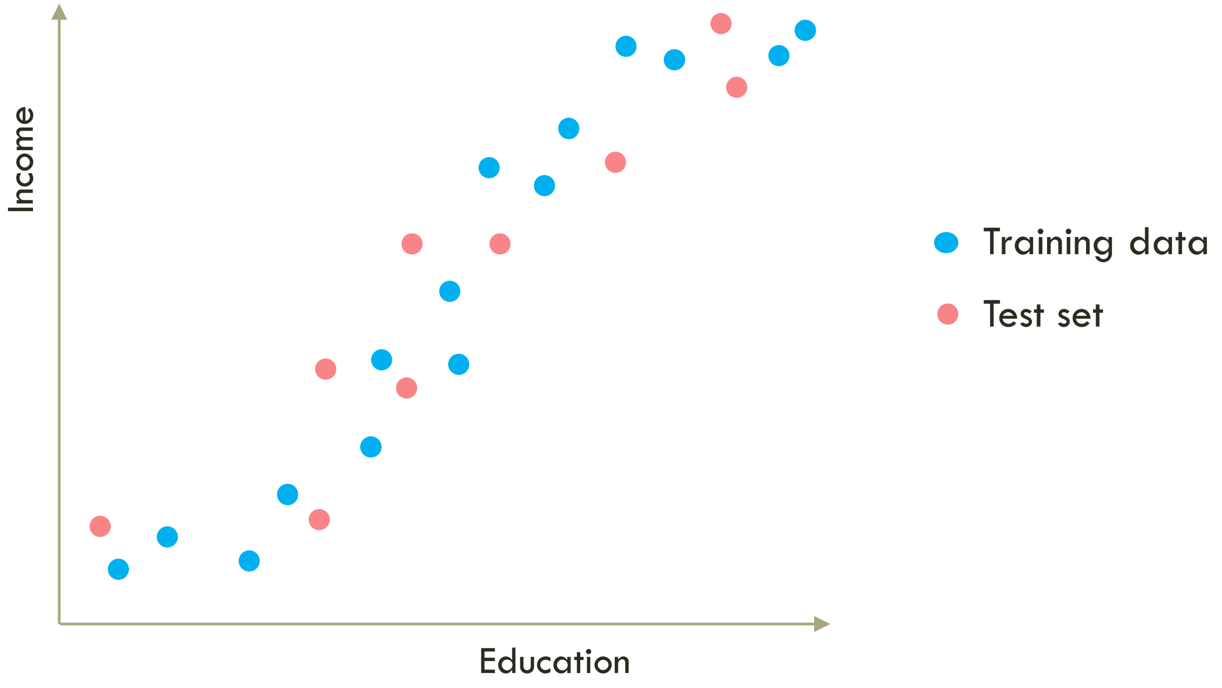
- The model is built on the training data („trained“) and then tested on the test set



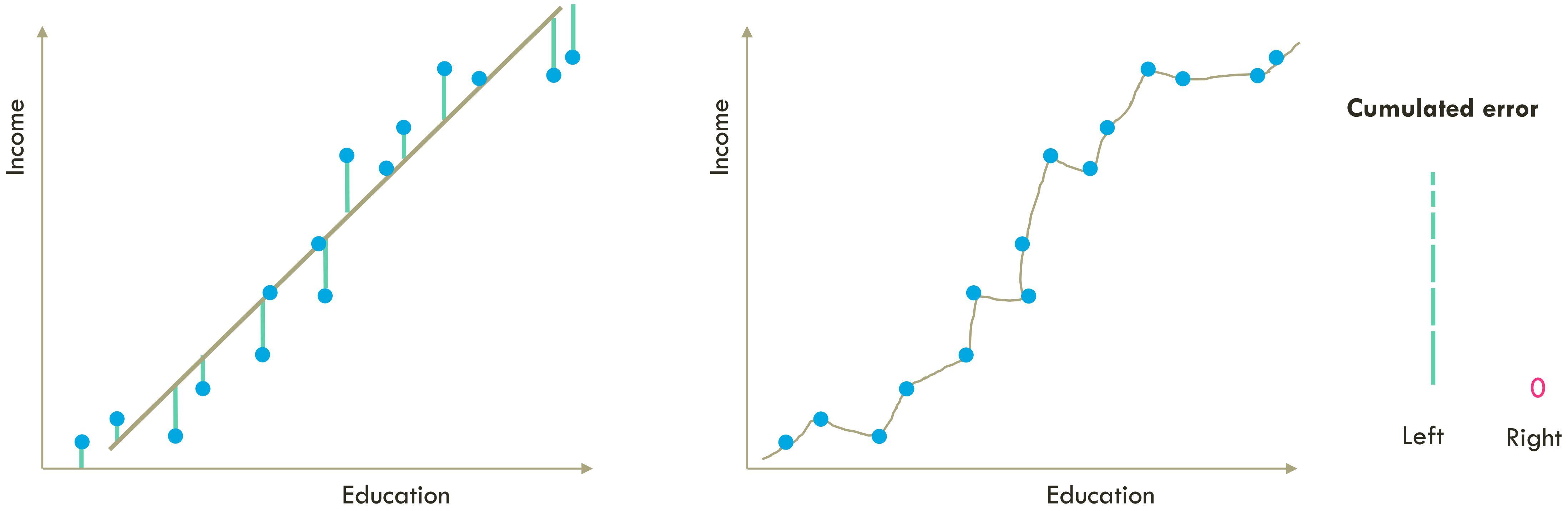
EXAMPLE DATA ON INCOME

Person [ID]	Income [in 1000 Euro]	Education [in years]	Work experience [in years]	Gender	Hair length [in cm]
1	100	22	14	Weiblich	20
2	93	18	15	Weiblich	25
3	35	12	13	Männlich	2
4	79	17	23	Weiblich	3
5	68	20	3	Weiblich	15
6	72	18	3	Weiblich	46
7	88	20	19	Weiblich	33
8	80	21	10	Weiblich	21
9	90	20	11	Weiblich	28
10	46	10	14	Männlich	10

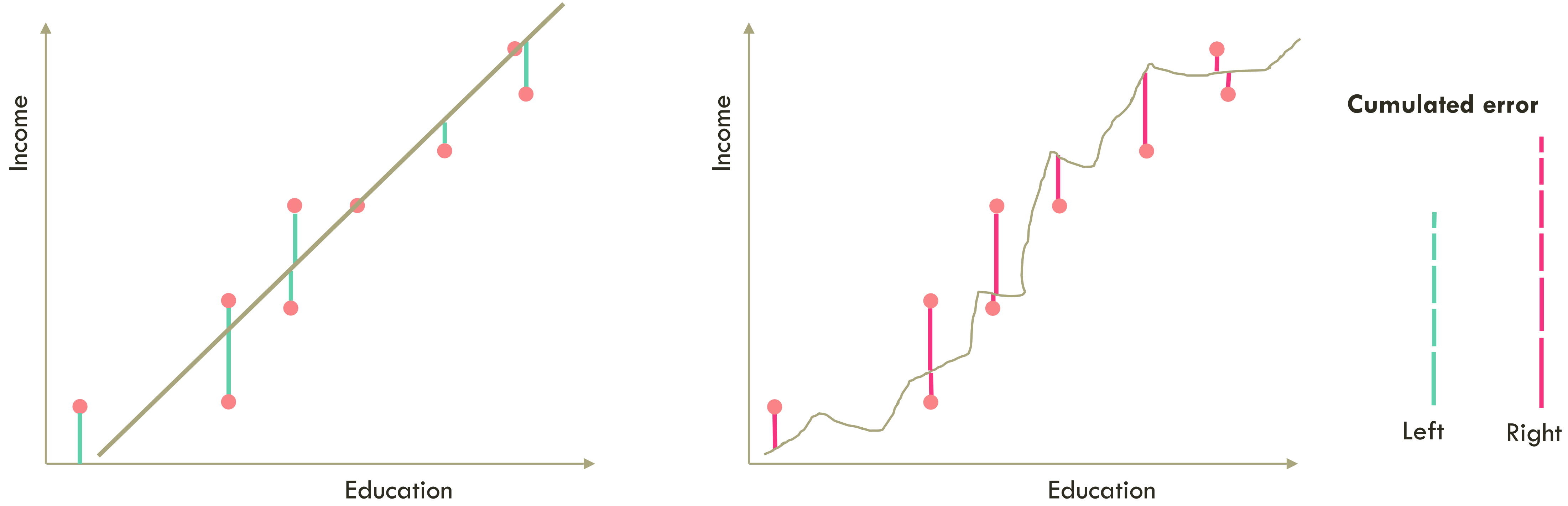
SPLIT THE DATASET INTO TRAINING AND TESTING DATA



BUILD THE MODELS USING THE TRAINING SET



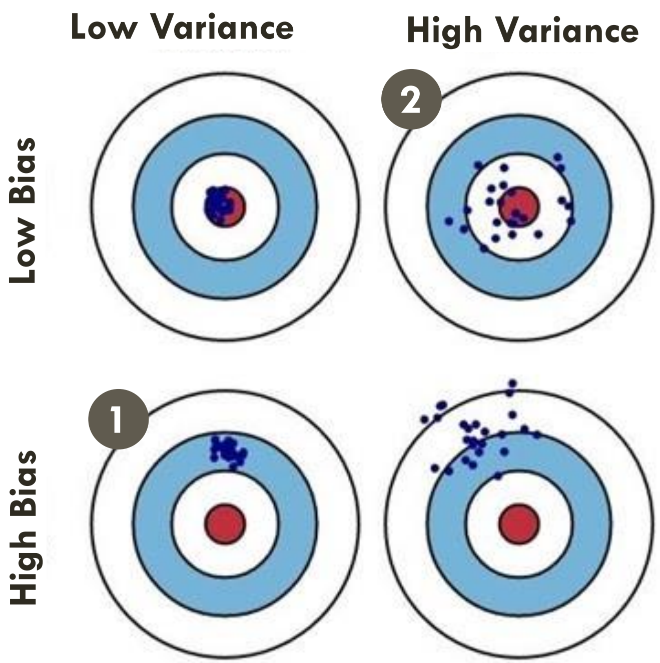
EVALUATE THE PERFORMANCE OF YOUR MODELS USING THE TEST SET



BIAS-VARIANCE TRADE-OFF

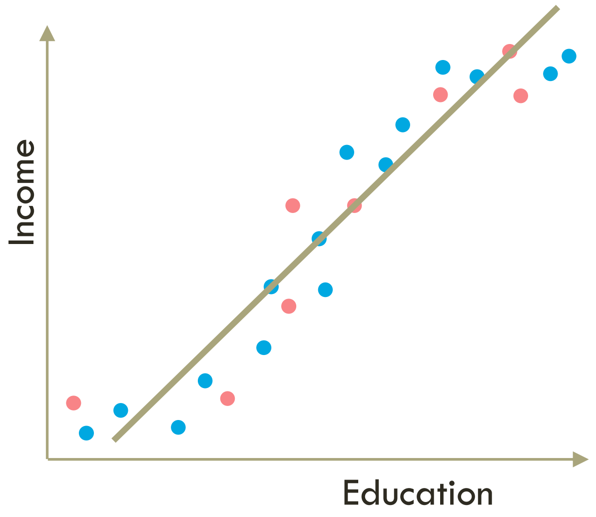
Bias: Difference between average prediction and correct value.

Variance: Variability of a model prediction for a given data point.



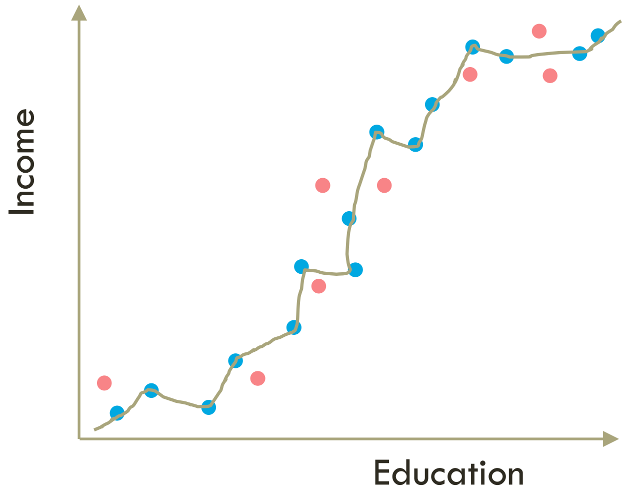
1 **Underspecification**

- Low Variance
- High Bias



2 **Overfitting**

- High Variance
- Low Bias



THE PREDICTION OF GROUP MEMBERSHIP IS CALLED CLASSIFICATION

- **Quantitative Target**= $f(x_1,x_2,\dots,x_k)$

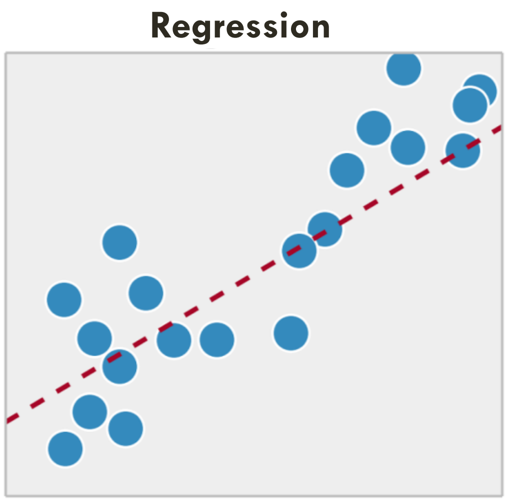
  e.g. Income

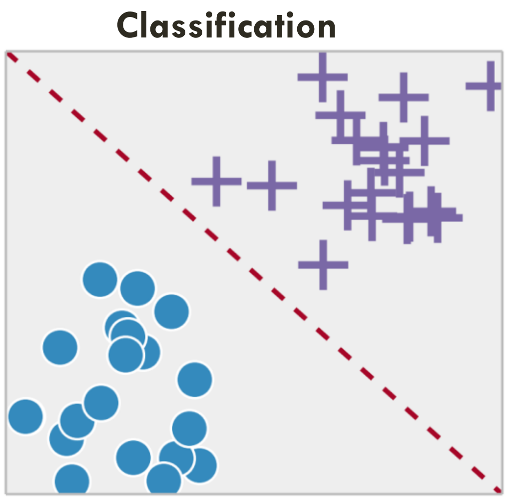
  30,000 €

  70,0035.01€

  …

  as f (Education, Sector, Work experience, …)





- **Qualitative Target** = $f(x_1,x_2,\dots,x_k)$

  e.g. Gender

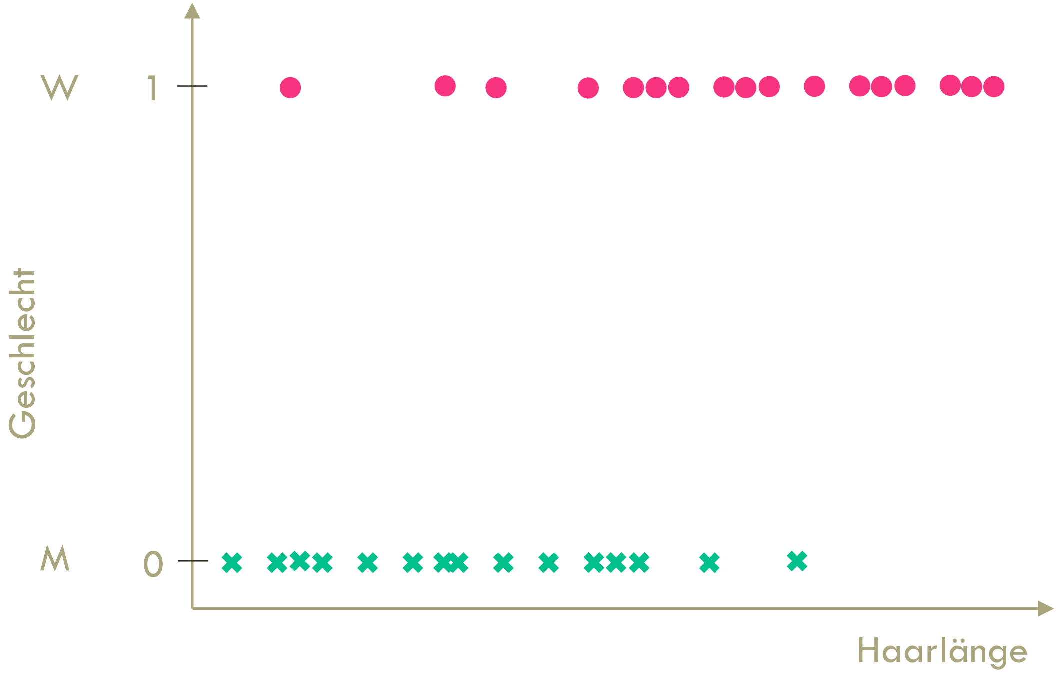
  1 if female

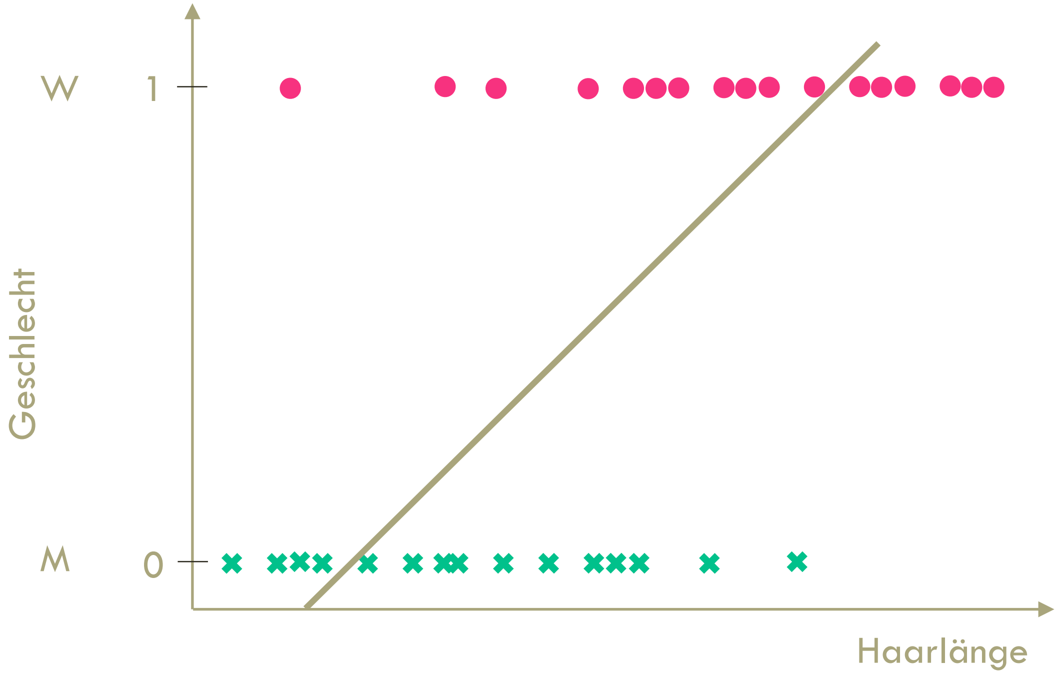
  0 if male

  as f (Hair length, Voice, …)

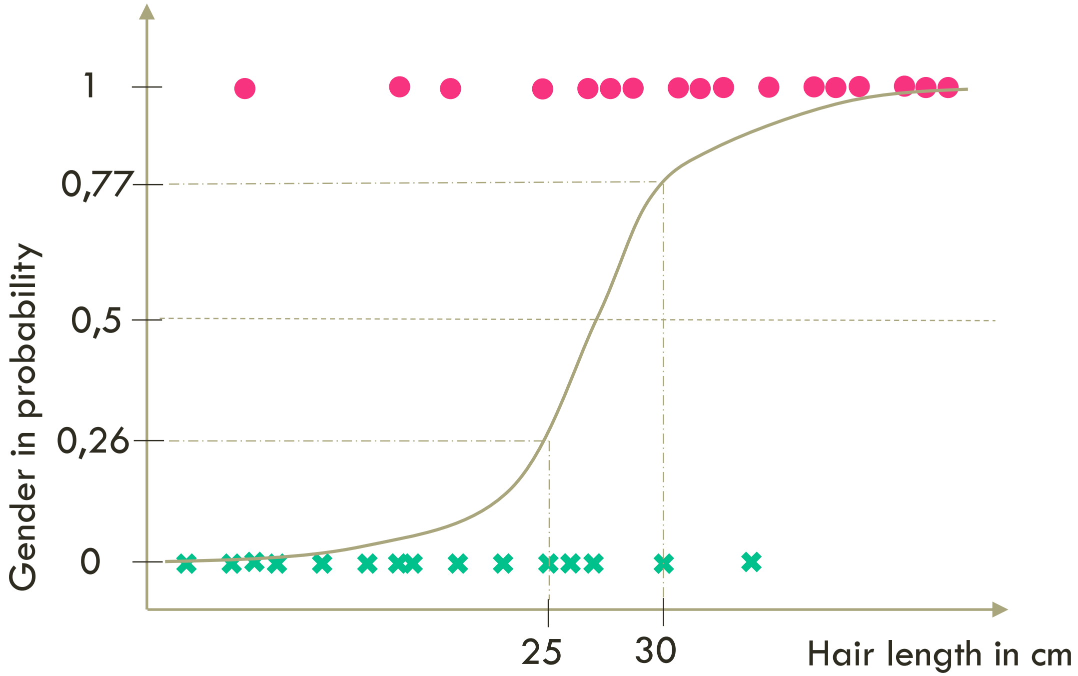
➢ Classification is also possible for more than two variables

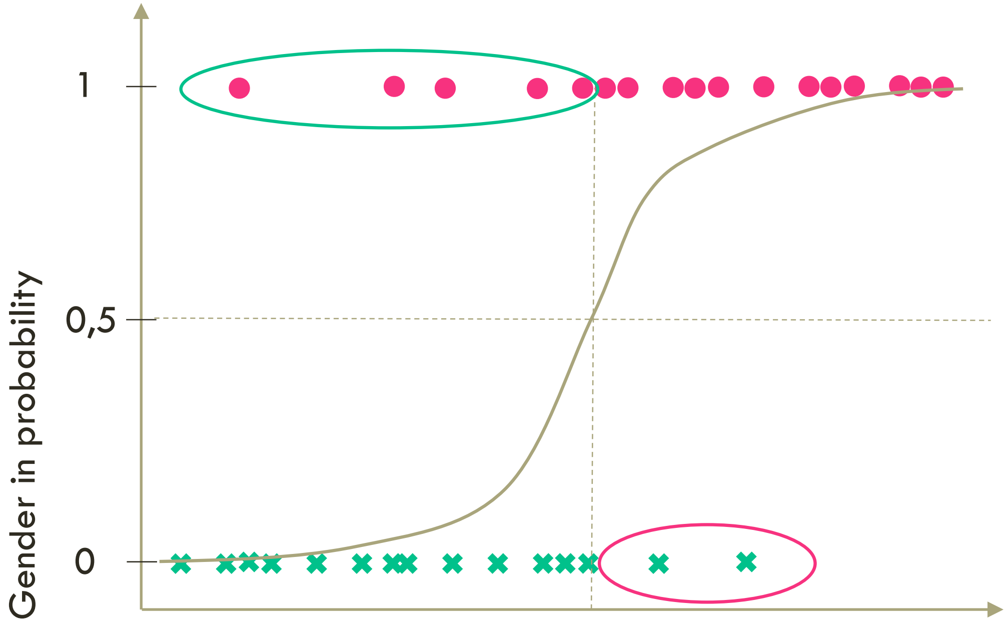
CLASSIFICATION NEEDS DIFFERENT MODELLING APPROACHES





LOGISTIC REGRESSION AS AN EXAMPLE MODEL FOR CLASSIFICATION





- Often the first model for data scientists to get a feel for the problem
- All observations having a probability above 0.5 are predicted as female, all below 0.5 as male

THE CONFUSION MATRIX IS USED TO MEASURE THE QUALITY OF FIT

True value

Predicted value

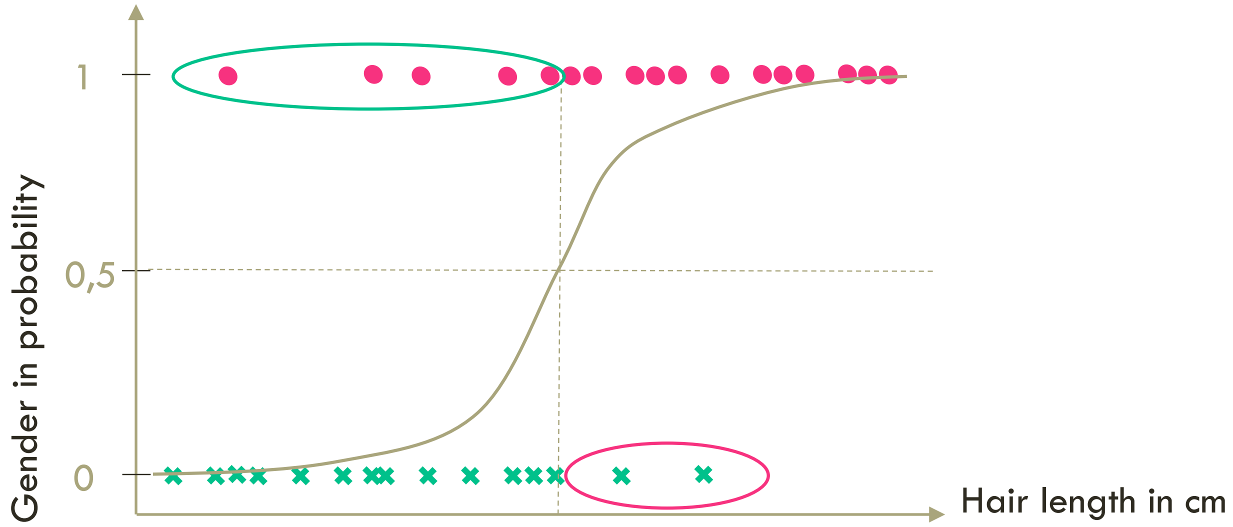
	Yes (Positive)	No (Negative)
Yes (Positive)	True Positives (TP)	False Positives (FP)
No (Negative)	False Negatives (FN)	True Negatives (TN)

Example: Hair length - Gender

True value

Predicted value

	1 (Female)	0 (Male)
1 (Female)	TP: 12	FP: 2
0 (Male)	FN: 5	TN: 13



<u>Beispiel Haarlänge - Geschlecht</u>

- TP: Als W vorhergesagt und tatsächlich W
- FP: Als W weiblich vorhergesagt, tatsächlich M
- FN: Als M vorhergesagt, tatsächlich W
- TN: Als M vorhergesagt und tatsächlich M

THE INDICATORS OF SUCCESS ARE DERIVED FROM THE CONFUSION MATRIX

True value

Predicted value

	Yes (Positive)	No (Negative)
Yes (Positive)	True Positives (TP)	False Positives (FP)
No (Negative)	False Negatives (FN)	True Negatives (TN)

Example: Hair length - Gender

True value

Predicted value

	1 (Female)	0 (Male)
1 (Female)	TP: 12	FP: 2
0 (Male)	FN: 5	TN: 13

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{12+13}{12+13+5+2} = 0.7813, so\ 78.13\%$$

- ➢ In what percentage of all cases was the model prediction correct?

$$\boldsymbol{Recall} = \frac{TP}{TP+FN} = \frac{12}{12+5} = 0.7059, so\ 70.59\%$$

- ➢ What percentage of the actual female was also predicted as a female?

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} = \frac{12}{12+2} = 0.8571, so\ 85.71\%$$

- ➢ What percentage of the female predicted is actually female?

OVERVIEW OF THE RESULTS

Example: Fraud

True Value

Predicted Value

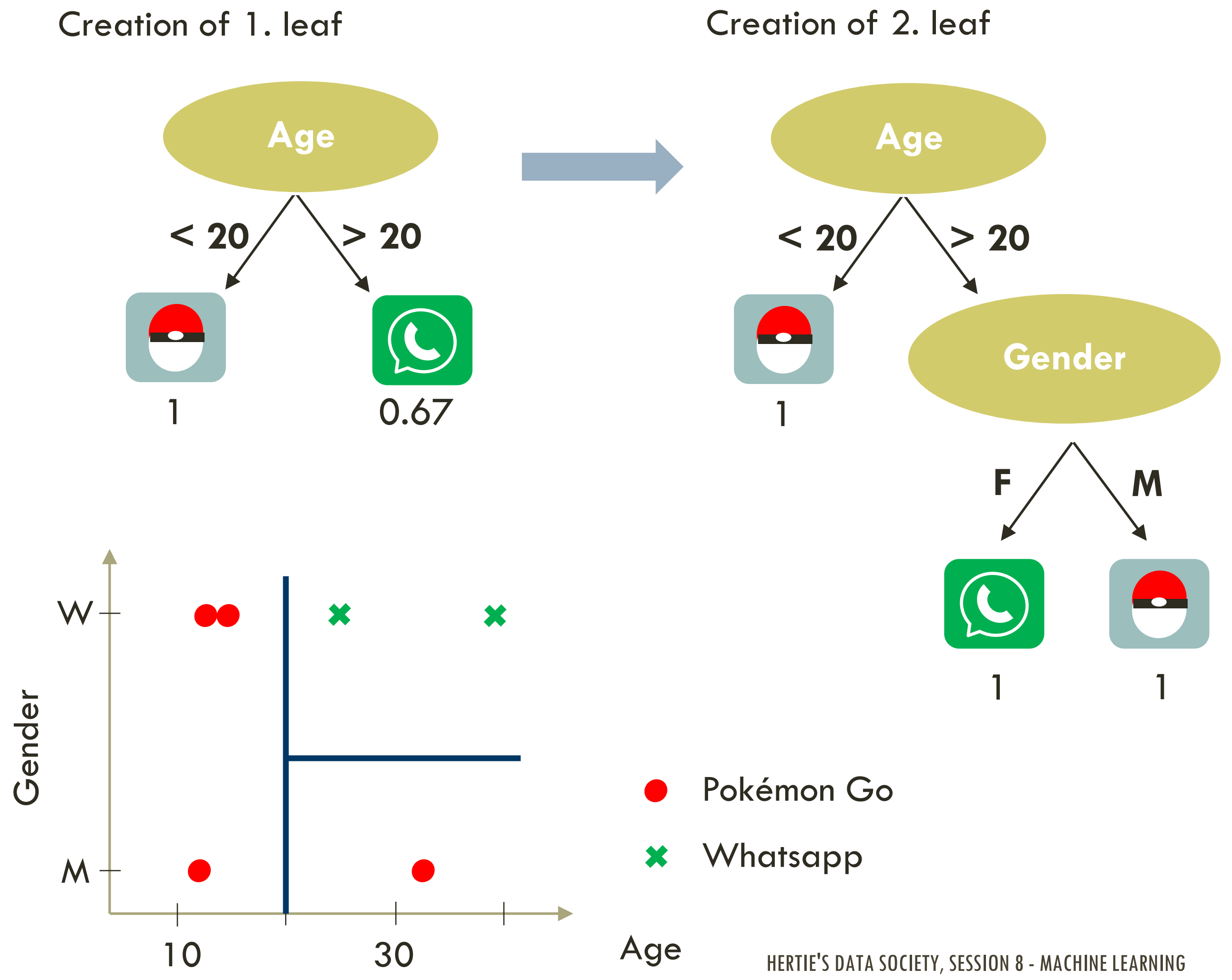
	1 (Fraud)	**0 (Not fraud)**
1 (Fraud)	TP: 777	FP: 116
0 (Not fraud)	FN: 858	TN: 552,331

- *Accuracy:* 99.82 %
- *Recall:* 47.52 %
- *Precision:* 87.01 %

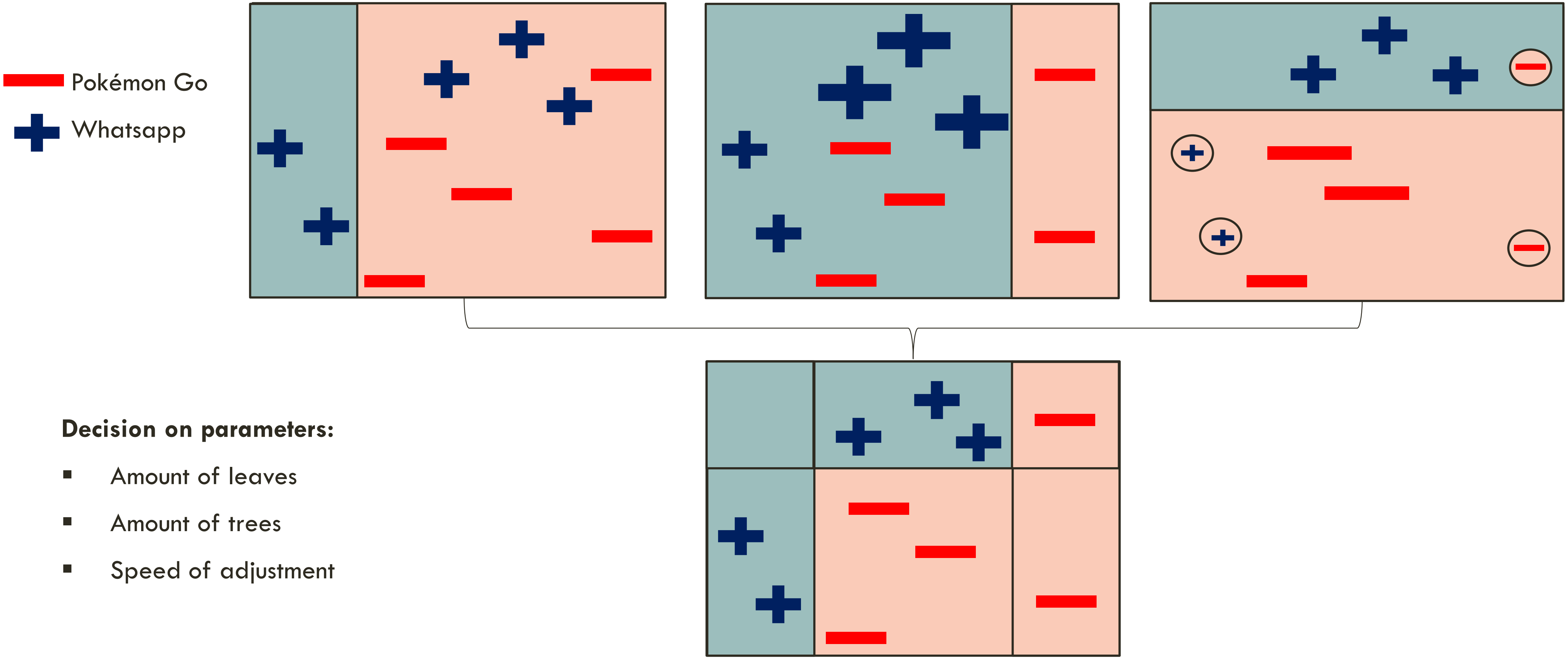
➢ *Better model for our dataset:* XGBoost (variant of Boosting)

DECISION TREES PART OBSERVATIONS INTO SUBGROUPS

Gender	Age	App-Download
F	15	
F	25	
M	32	
F	40	
M	12	
F	14	



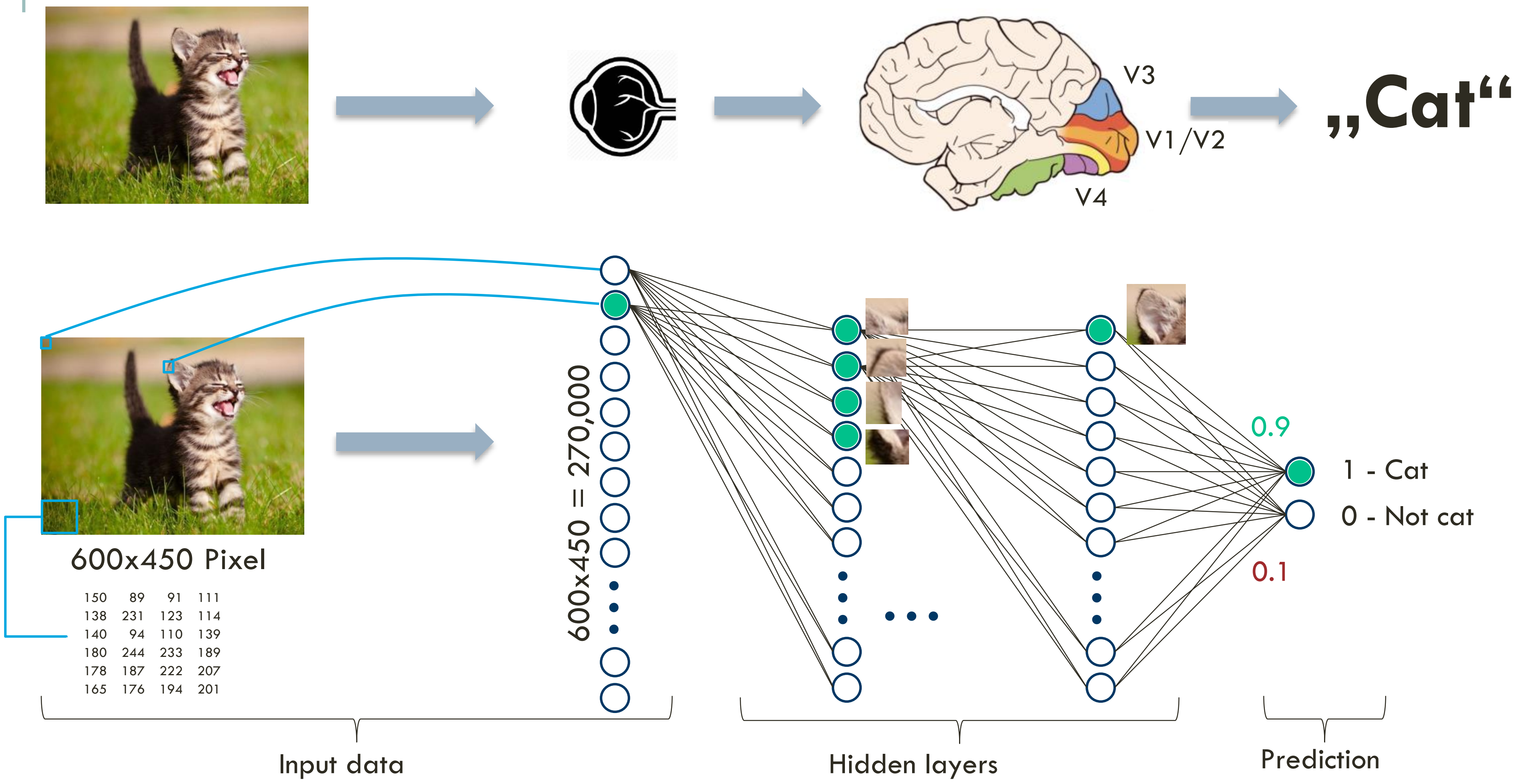
BOOSTING AS EXAMPLE OF IMPROVED DECISION TREES



Decision on parameters:

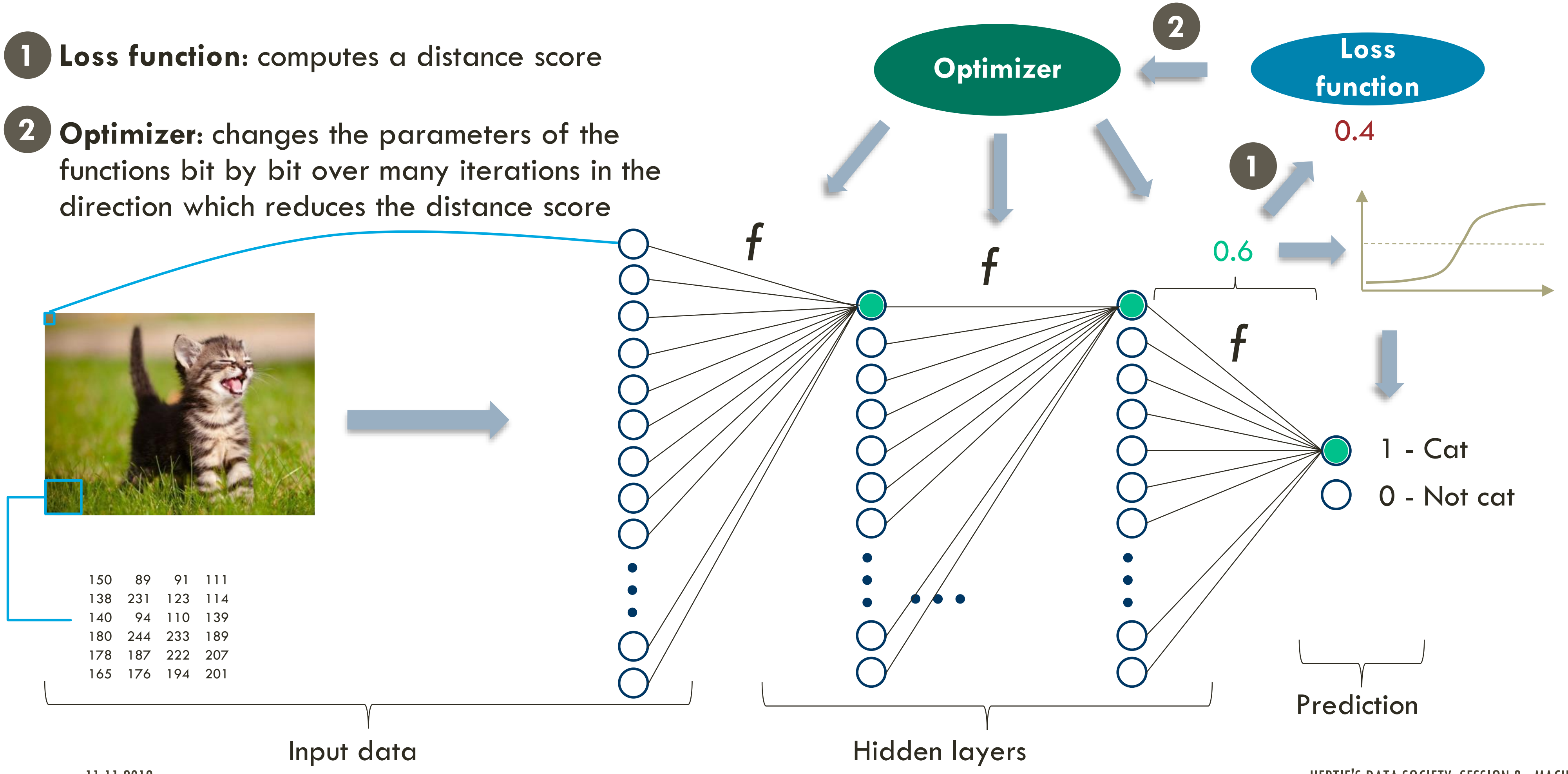
- Amount of leaves
- Amount of trees
- Speed of adjustment

NEURAL NETWORKS LEARN CONCEPTS



NEURAL NETWORKS CONSIST OF FUNCTIONS AND IMPROVE GRADUALLY

1. **Loss function:** computes a distance score
2. **Optimizer:** changes the parameters of the functions bit by bit over many iterations in the direction which reduces the distance score



THE K-MEANS CLUSTERING ALGORITHM IDENTIFIES GROUP MEMBERSHIPS