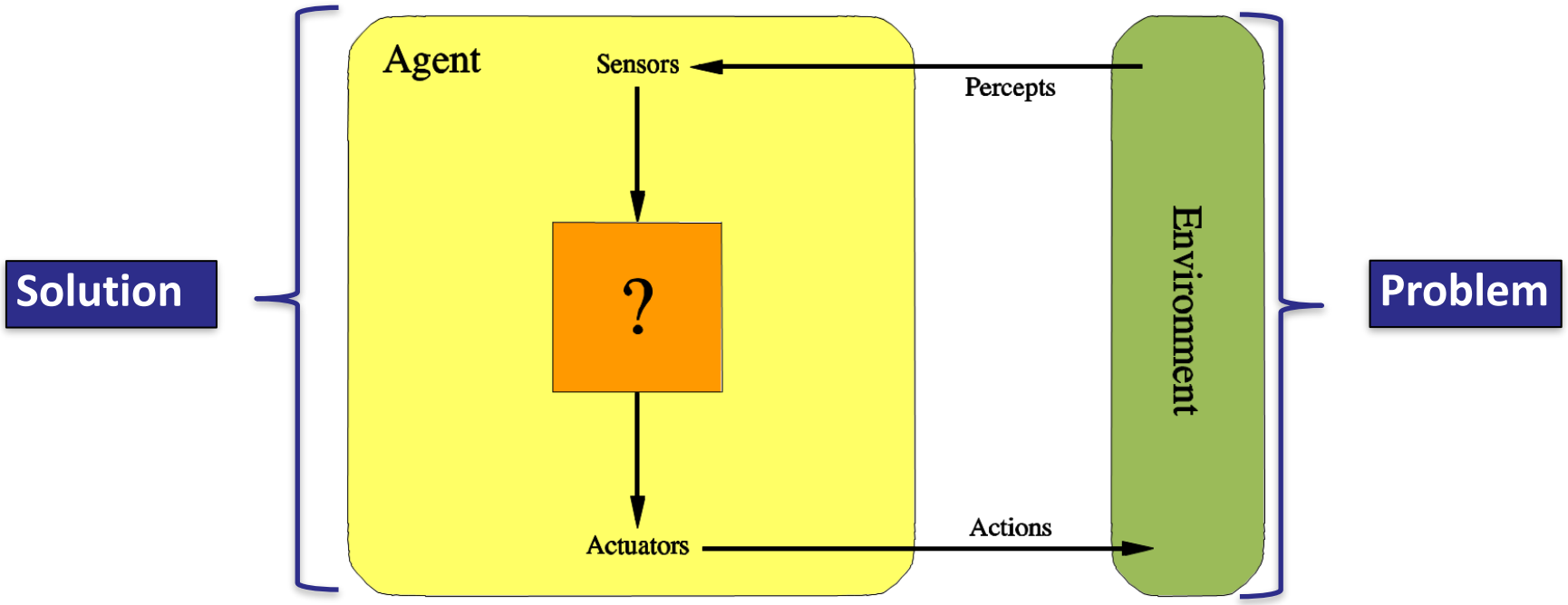
Course
Knowledge-based Systems

Lecture 10 – Basics of Machine Learning for Knowledge Representation

Dr. Mohsen Mesgar

Universität Duisburg-Essen

Recall


Solution
Agent
Sensors
Percepts
?
Actuators
Actions
Environment
Problem

Any other open questions?

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Machine Learning (ML)

ML

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- An agent should ideally behave like humans. We as humans can acquire knowledge to conduct a task using some supervision
- Since we can do it, so should AI agents
- In ML, we aim for algorithms that learn to gather the knowledge required for solving a problem

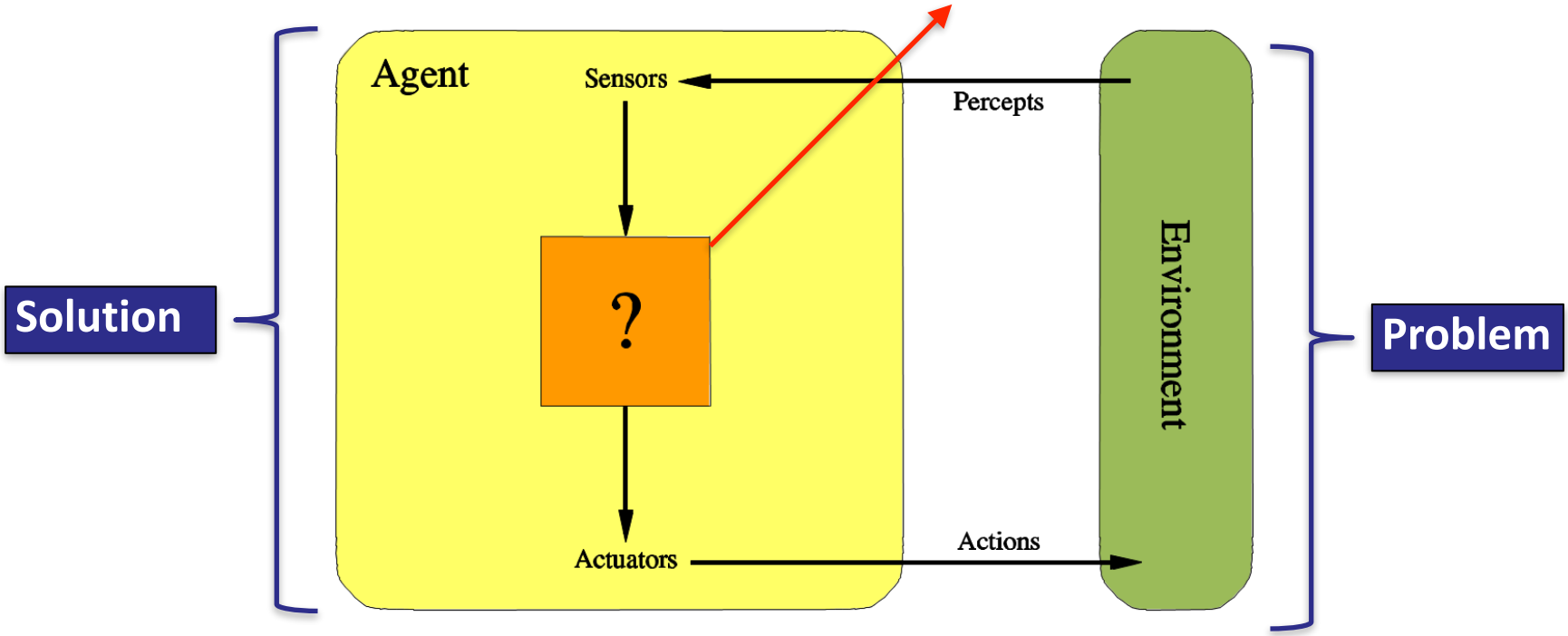
ML

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- ML lets agents **learn to optimize their knowledge** to achieve the predefined goals.

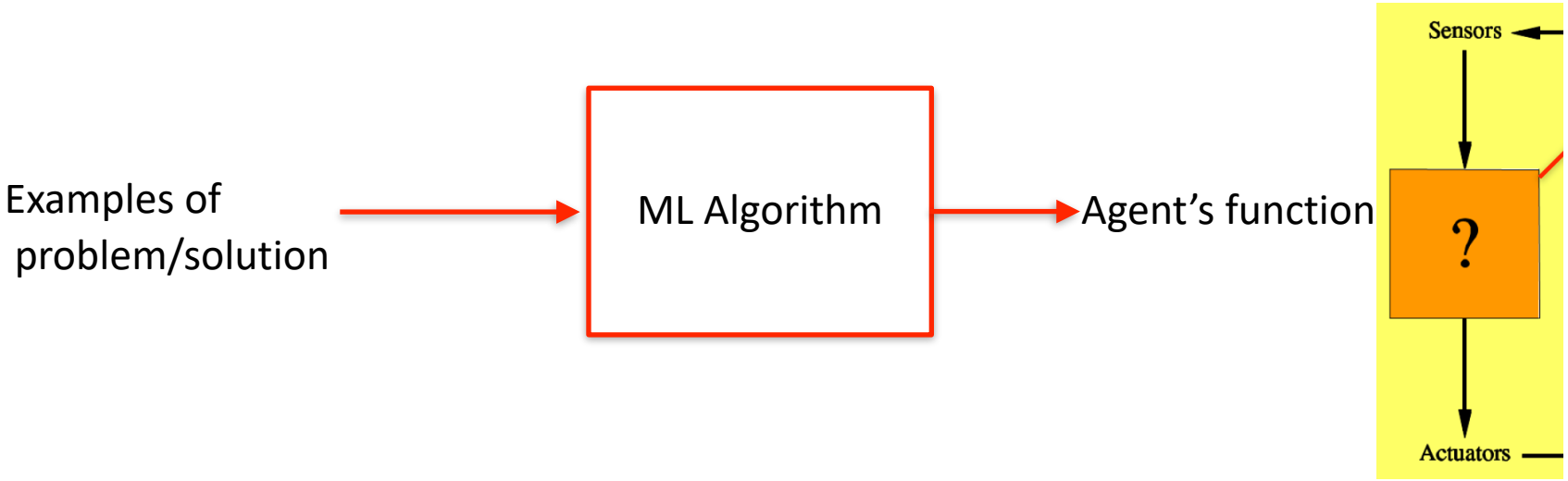
ML algorithms find this function that solves the problem



ML


Examples of problem/solution
ML Algorithm
Agent's function
?

ML

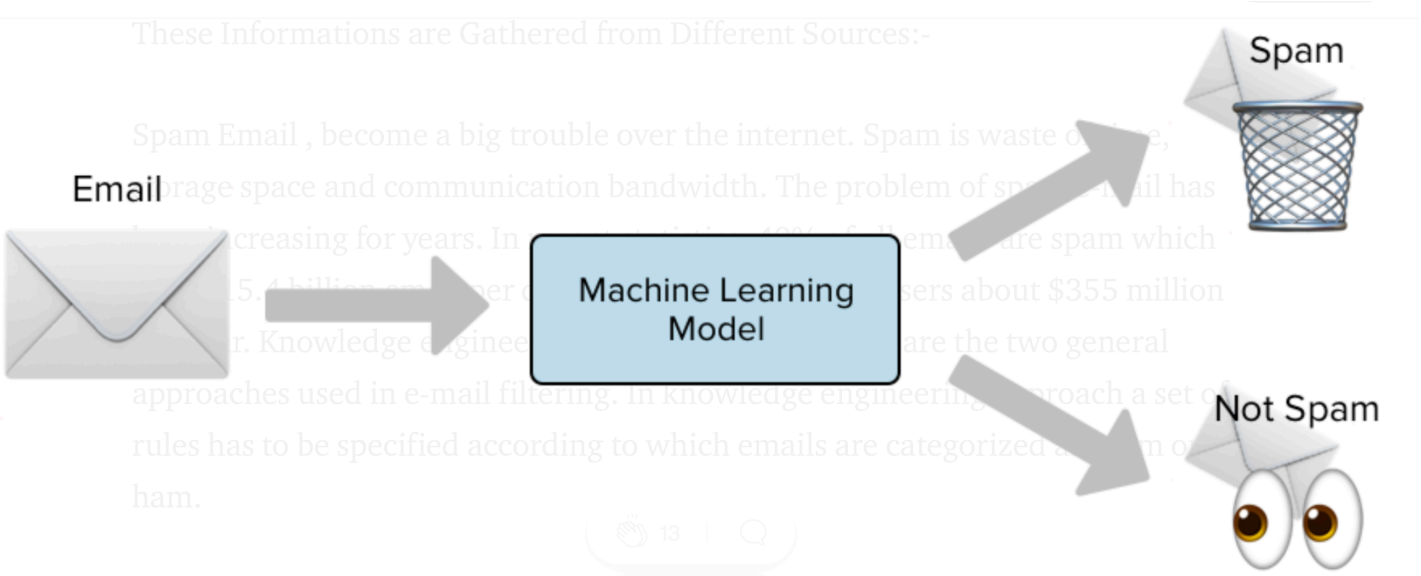

Examples of problem/solution
ML Algorithm
Agent's function
Sensors
?
Actuators

Example: Spam email filtering

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Why is this task important? Spam prevents the user from making full and good use of time, storage capacity and network bandwidth



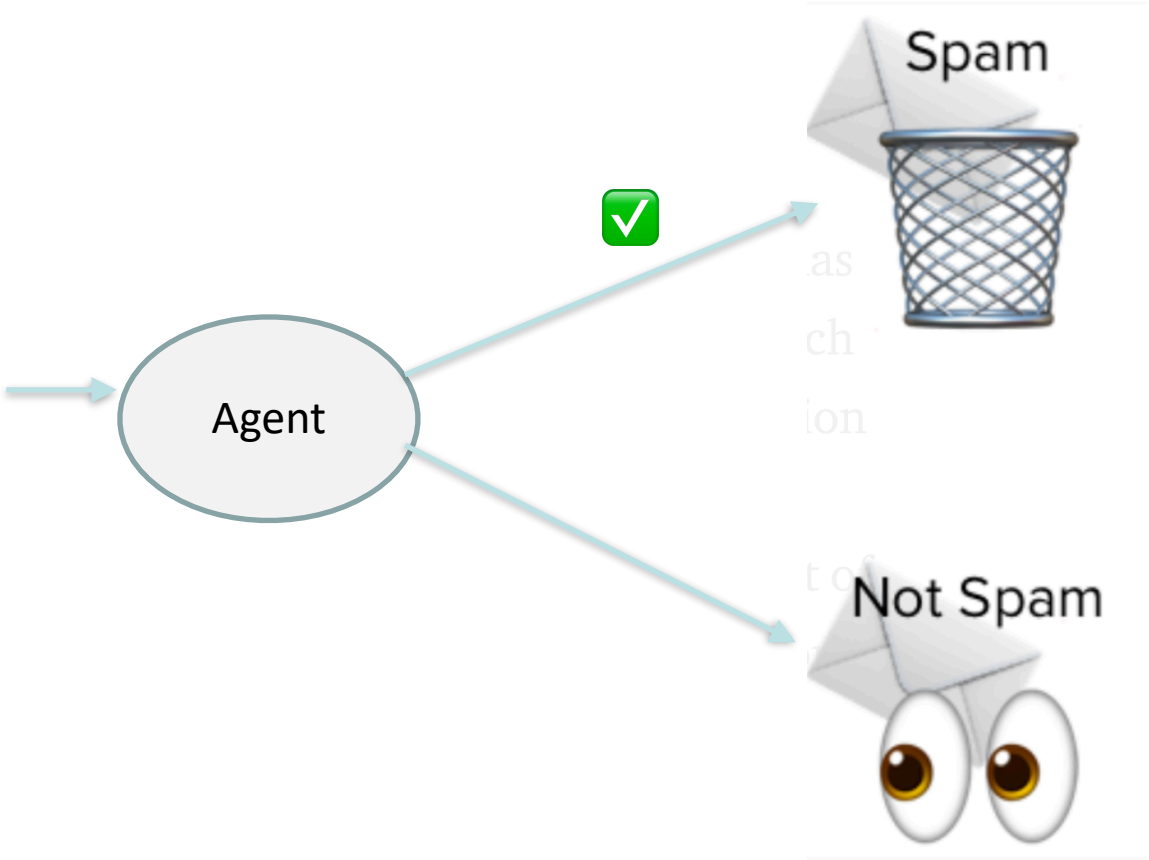
https://medium.com/analytics-vidhya/email-spam-classifier-using-naive-bayes-a51b8c6290d4

Example: Spam email filtering

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Hello,
Do you want free printer cartridges? Why pay more when you can get them ABSOLUTELY FREE!
Just...

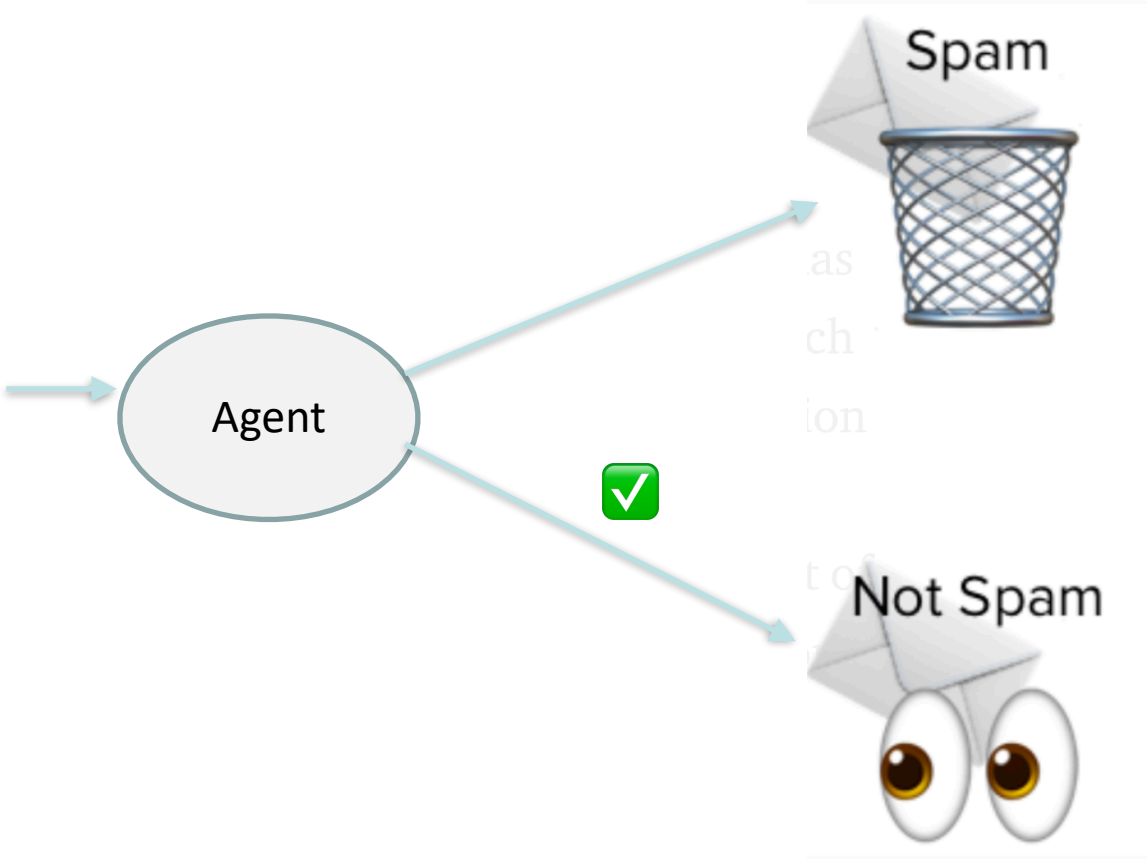


Example: Spam email filtering

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Hi Anna,
how is it going in with the new apartment? I just wanted to invite you and John for my birthday party next weekend...

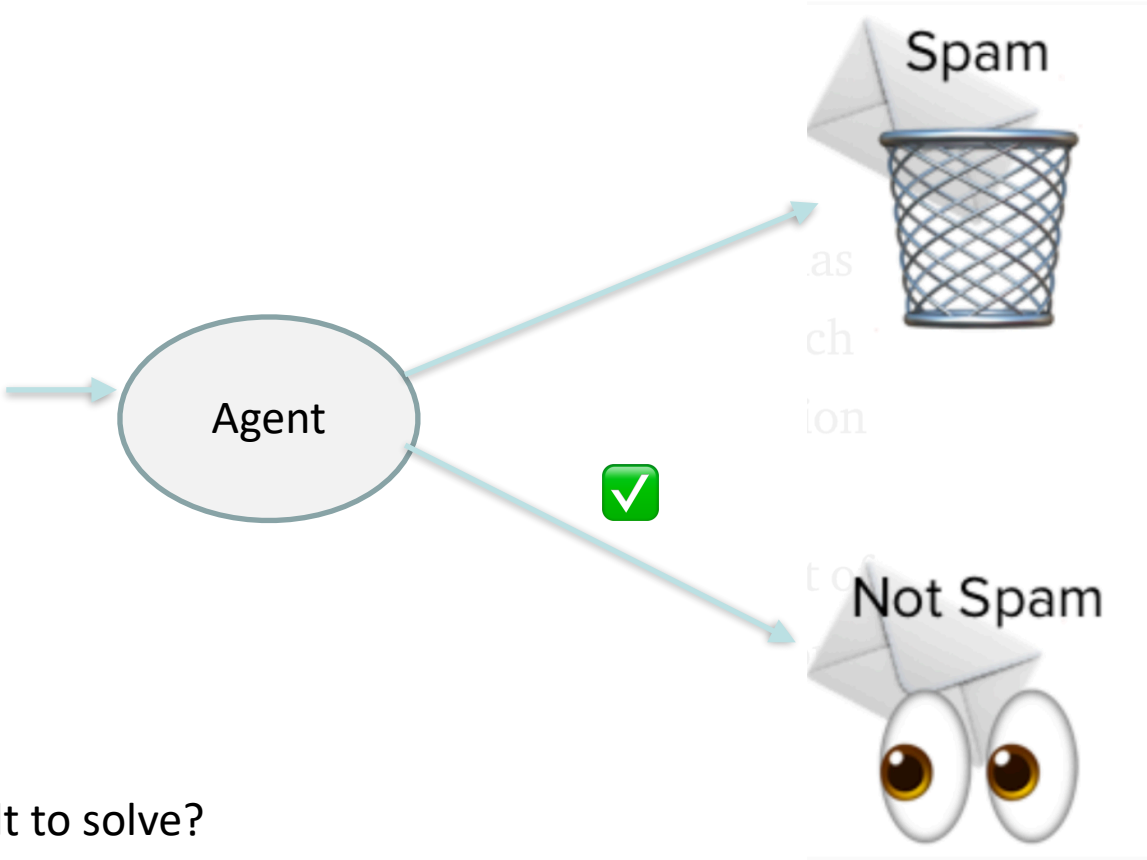


Example: Spam email filtering

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Hi Anna,
how is it going in with the new apartment? I just wanted to invite you and John for my birthday party next weekend...



Why is this problem difficult to solve?

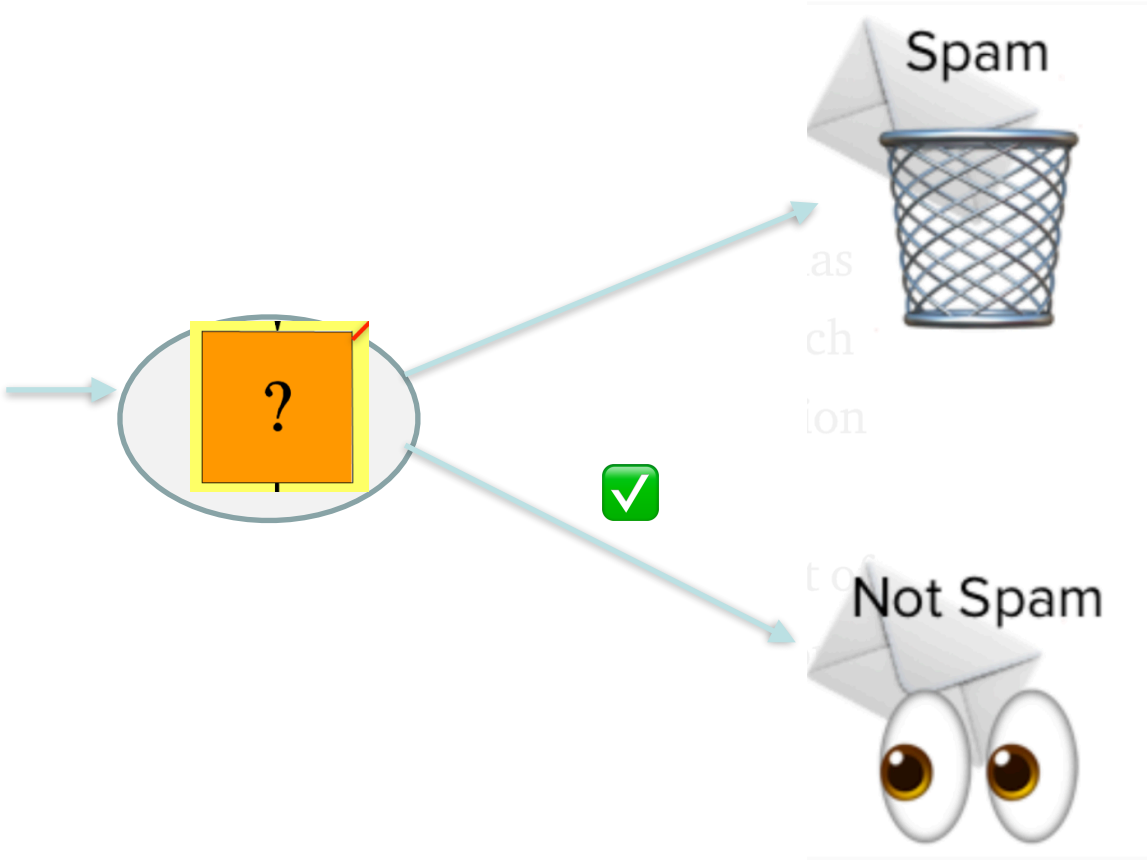
What knowledge should the agent learn?

Spam email filtering: agent

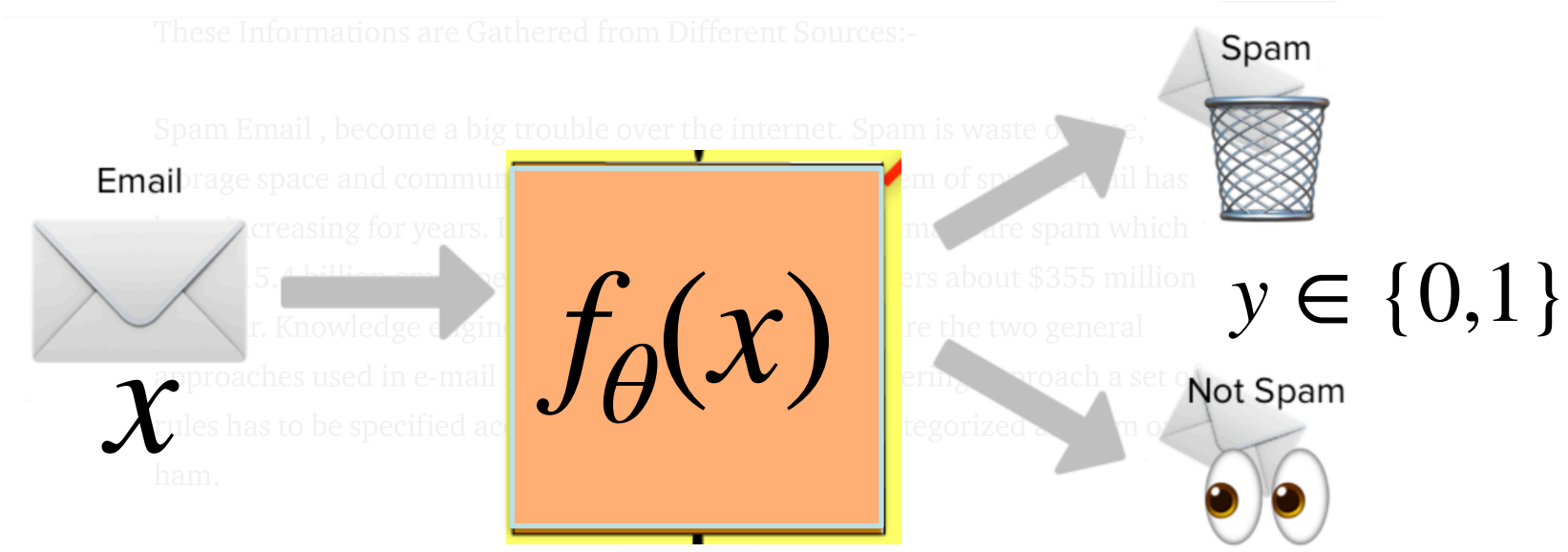
UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Hi Anna,
how is it going in with the new apartment? I just wanted to invite you and John for my birthday party next weekend...



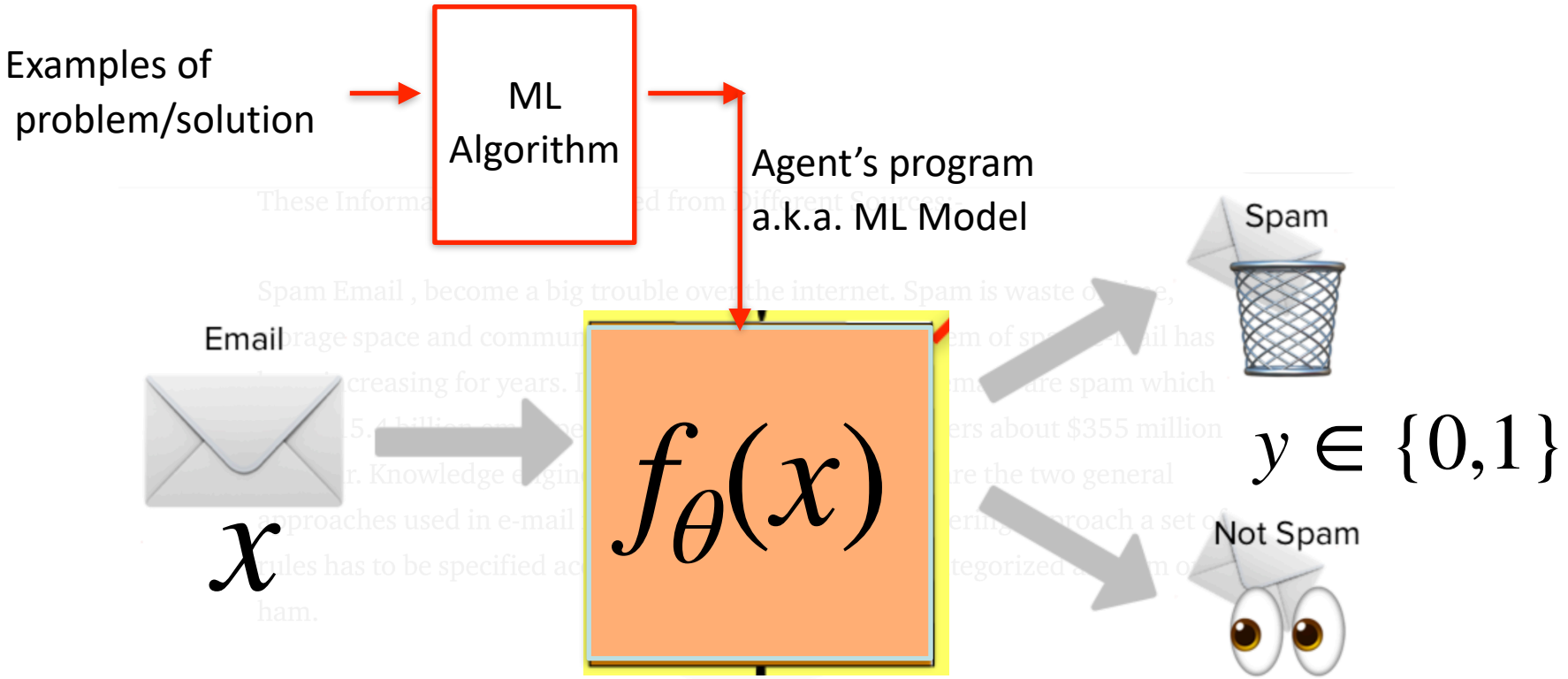
Spam email filtering: Inference phase



X, and y are variables (they can have different values).

θ indicates a set of parameters that define function f

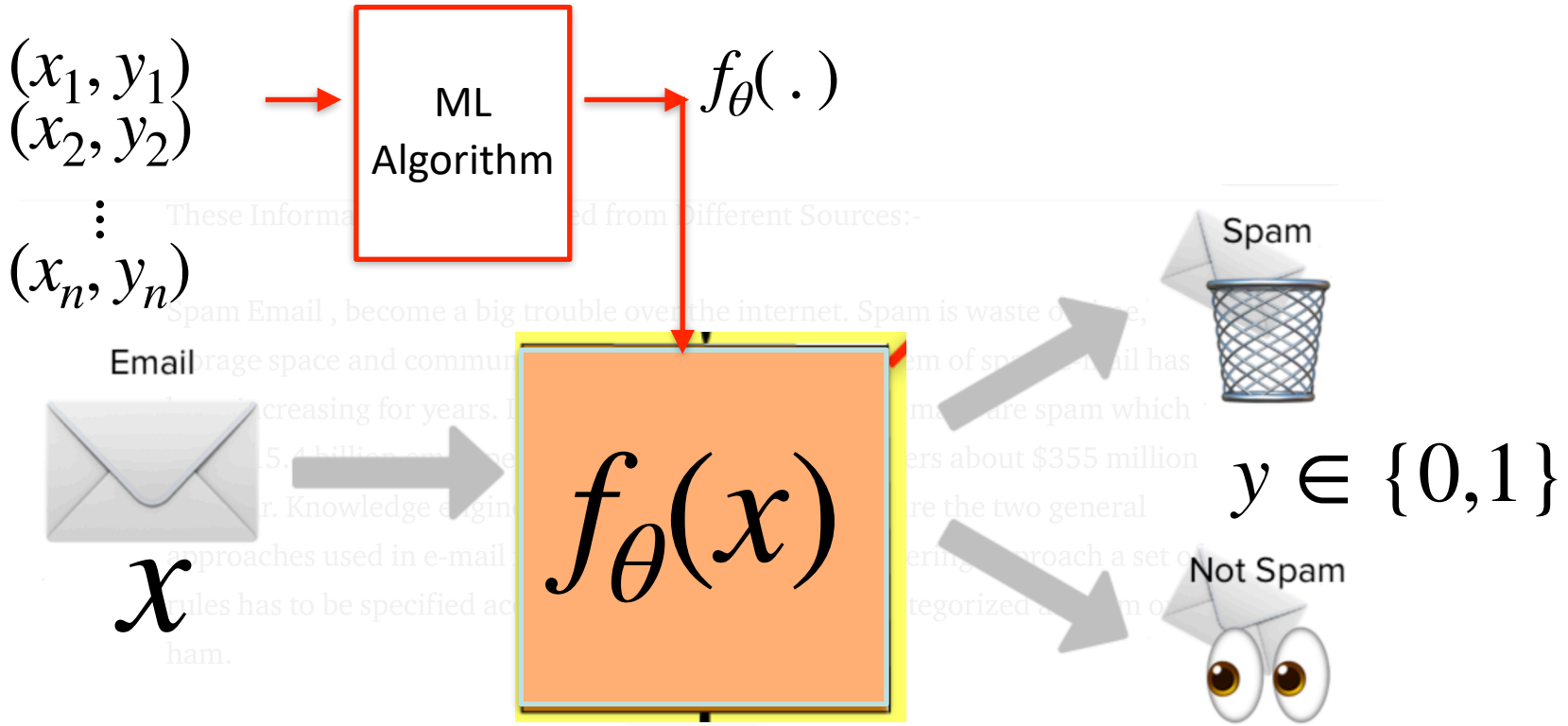
Spam email filtering: Learning phase



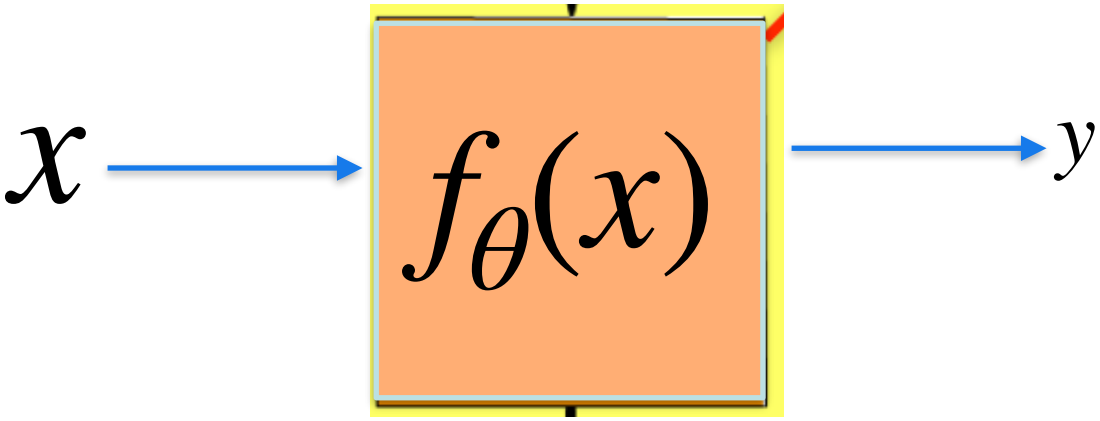
Spam email filtering: Learning phase

UNIVERSITÄT DUISBURG ESSEN

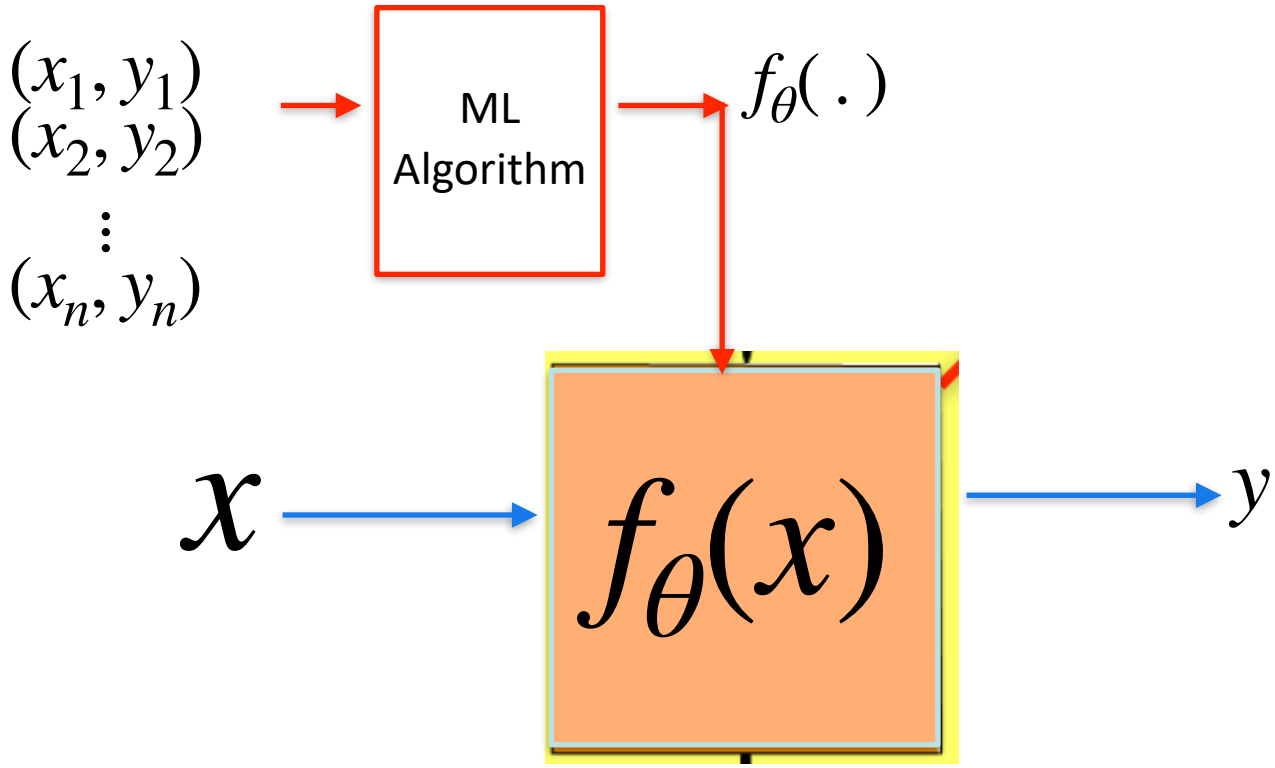
Offen im Denken



Inference



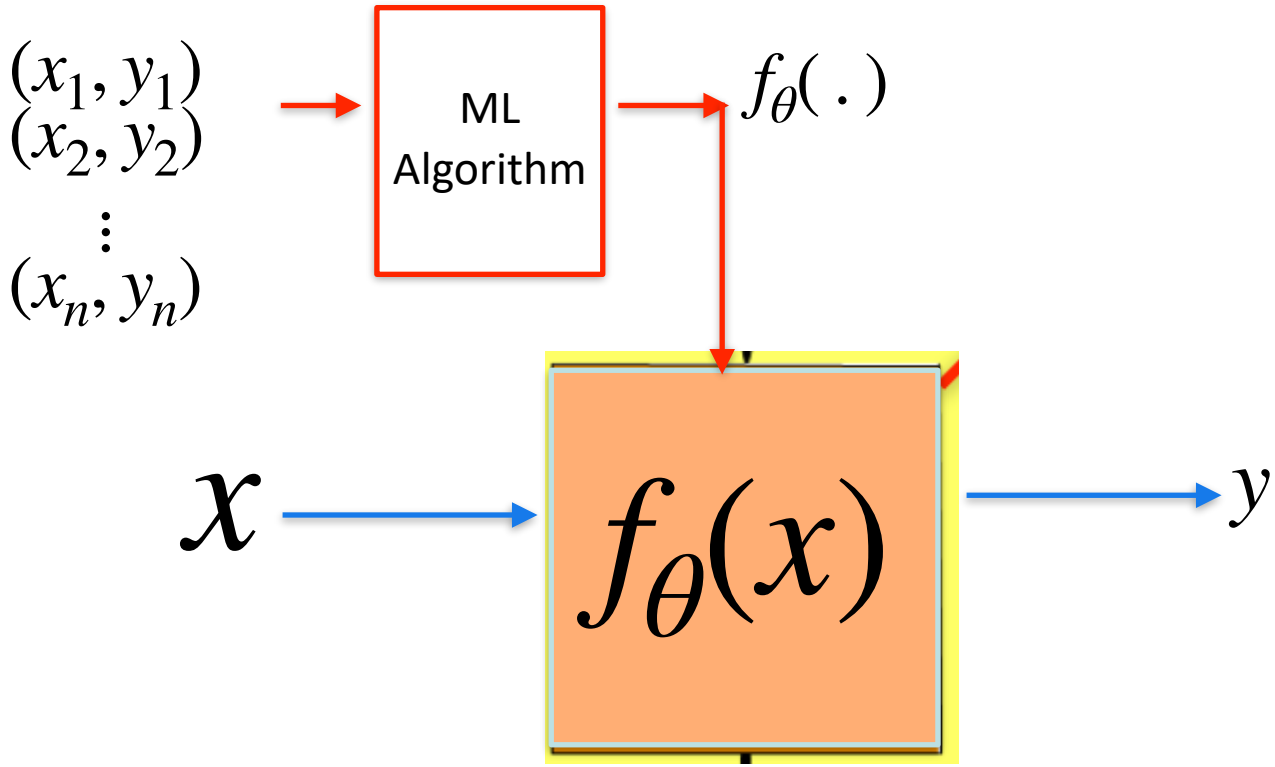
Learning from examples



Learning from examples

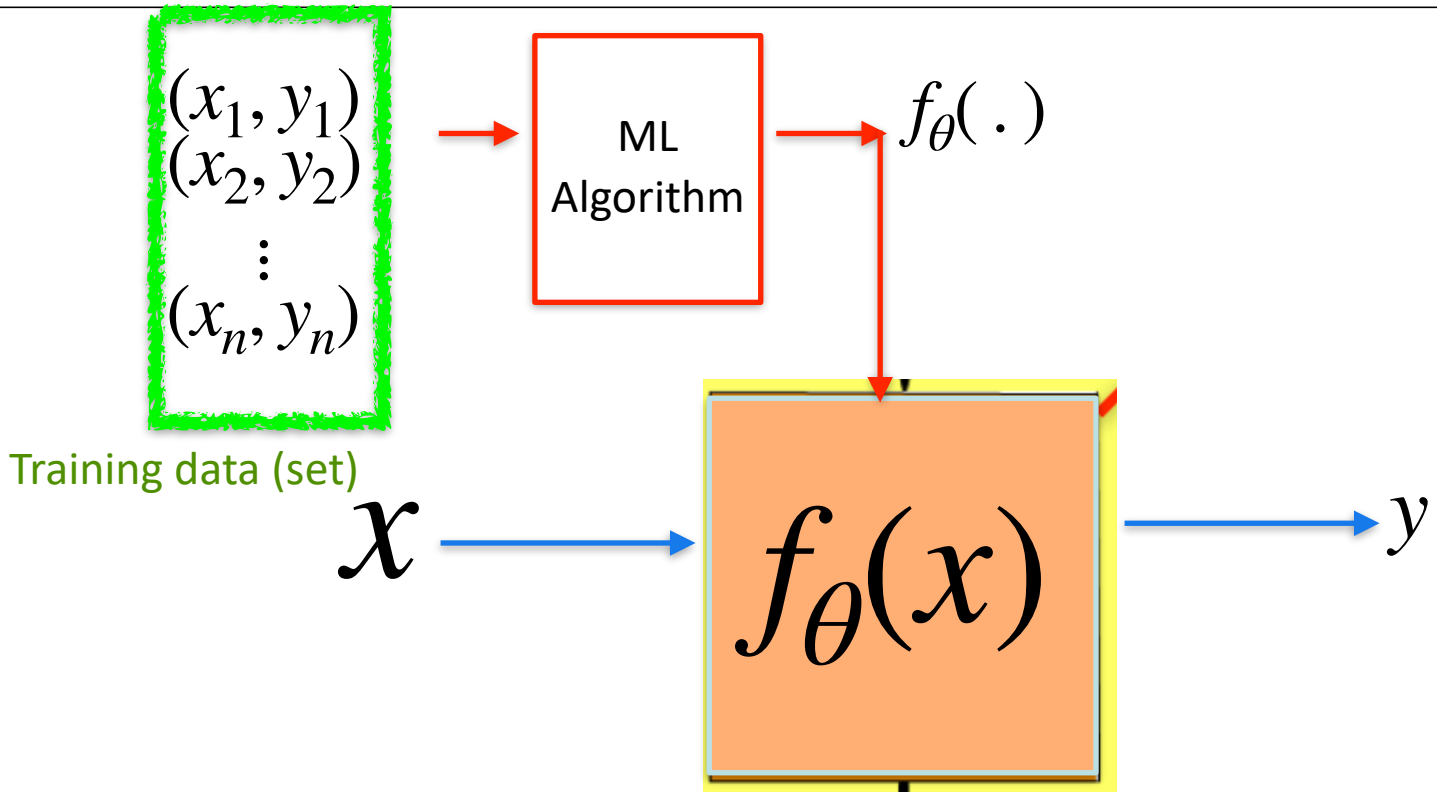
UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

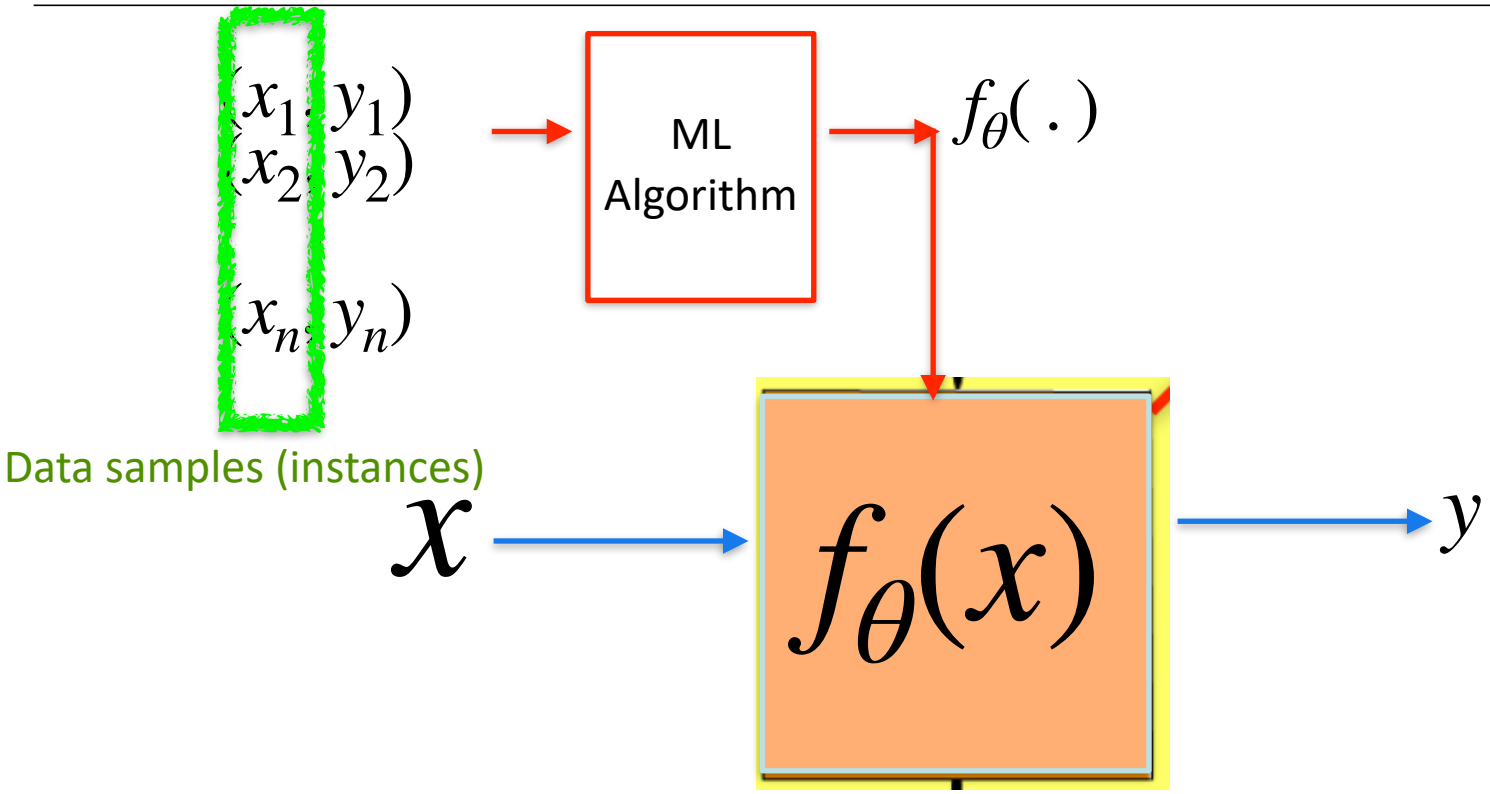


In ML, we aim to optimize parameters θ so that they capture as much as knowledge required to map input x to output y.

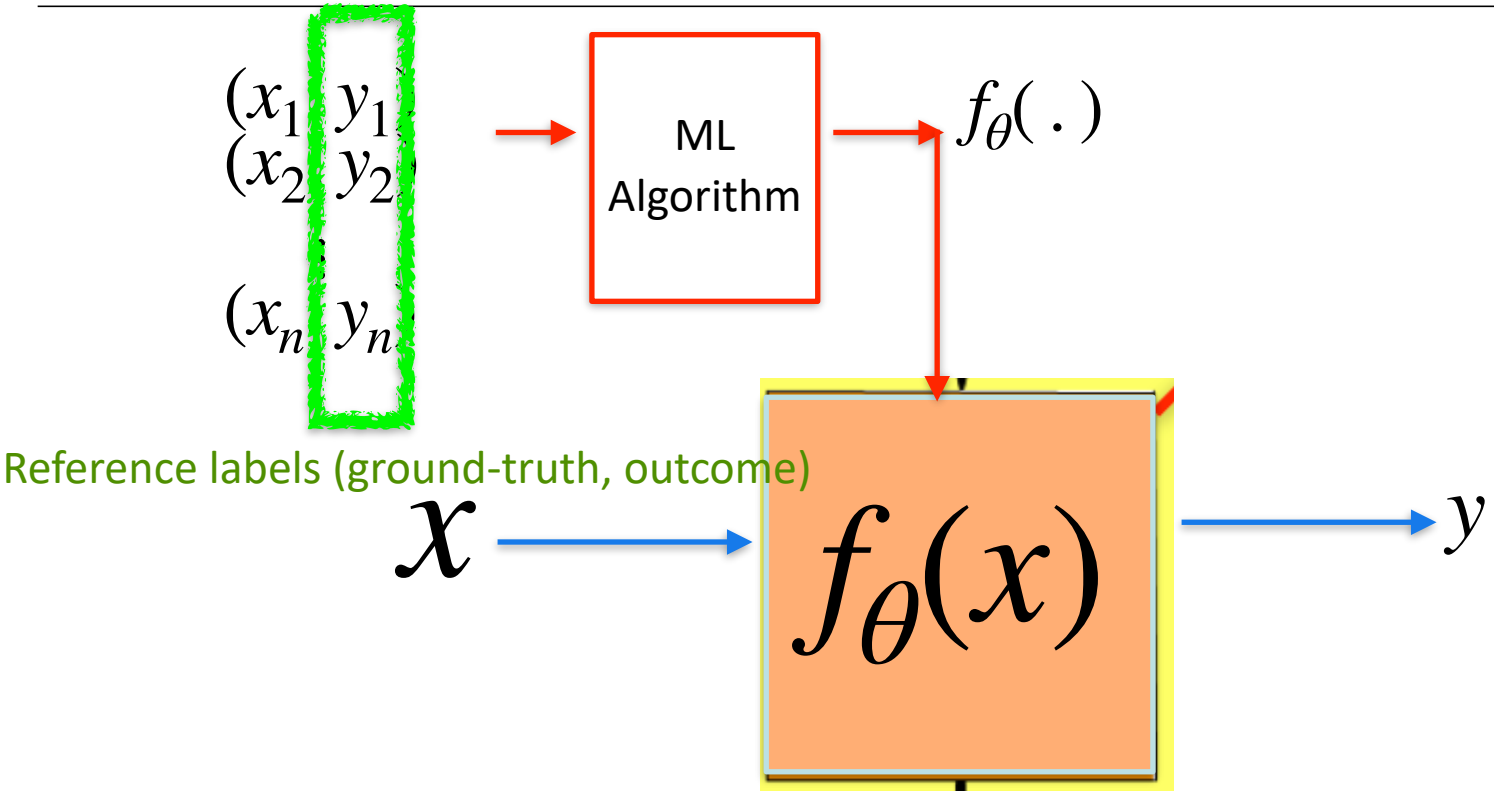
Terminology



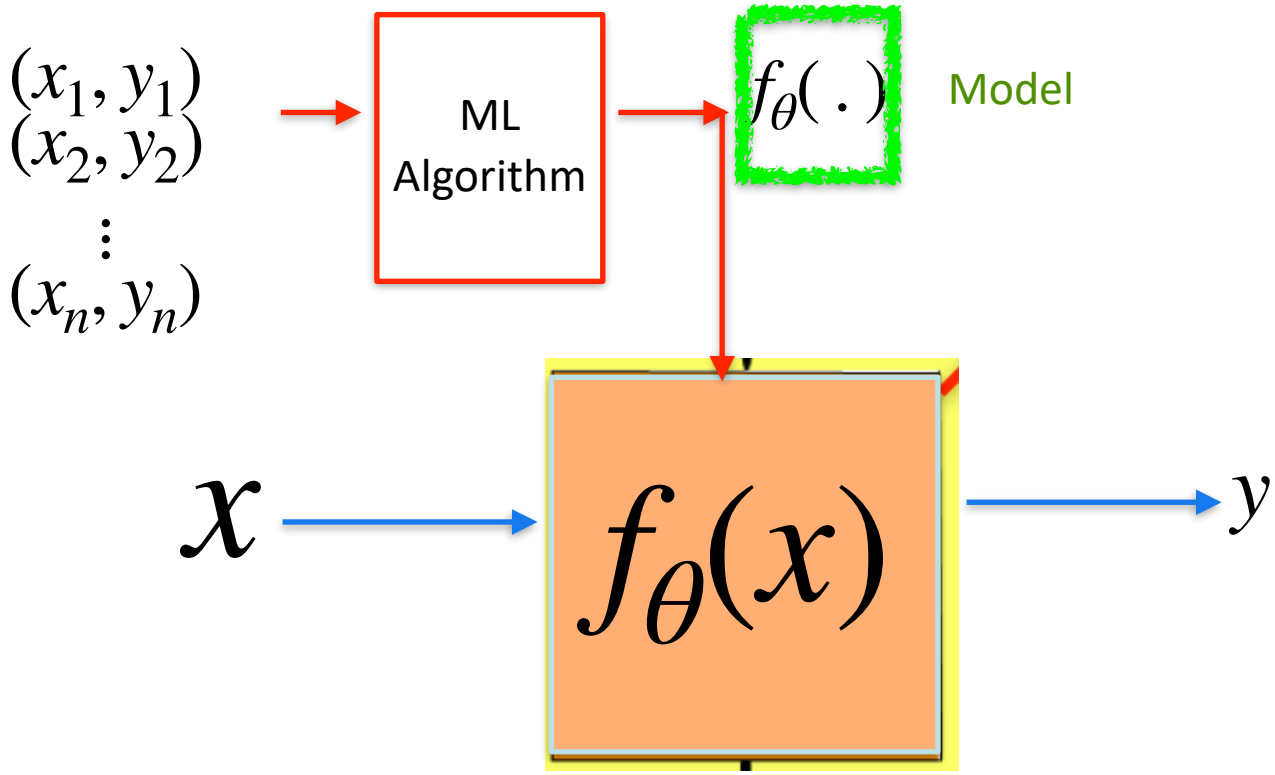
Terminology



Terminology



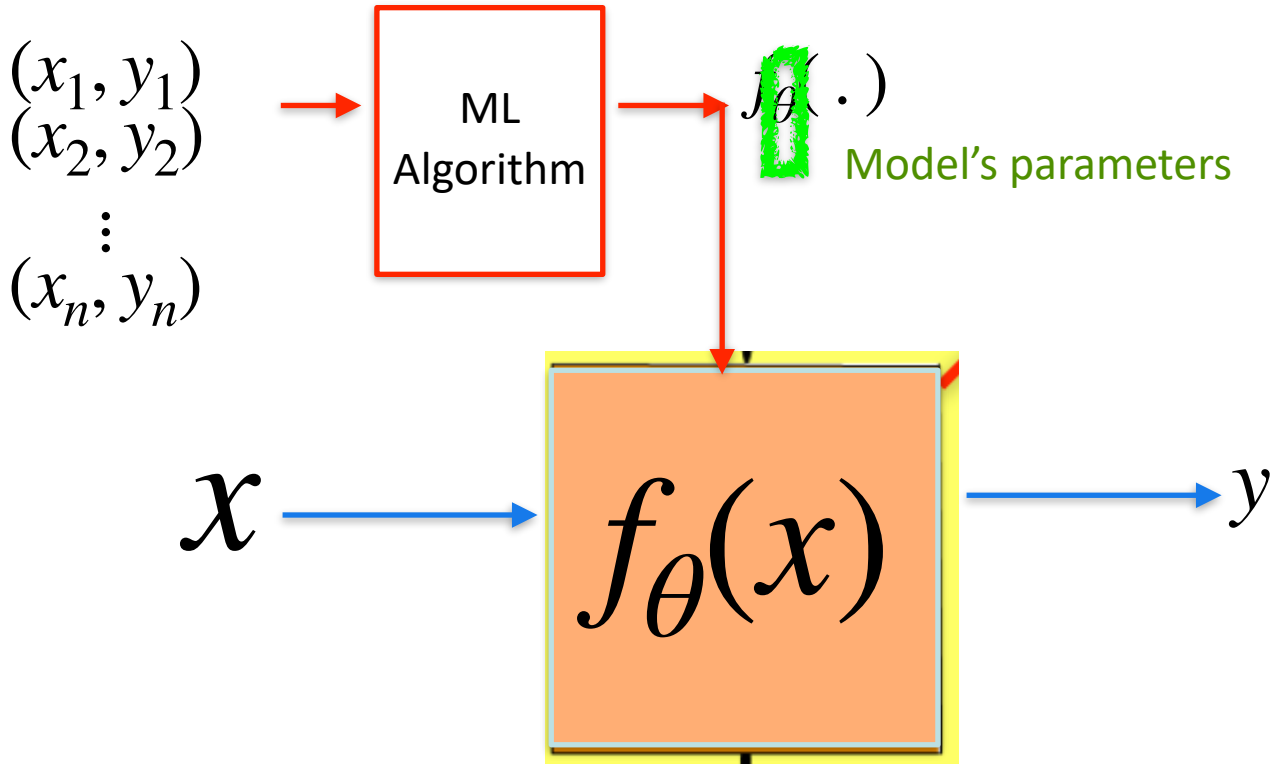
Terminology



Terminology

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



Feature

Feature

UNIVERSITÄT DUISBURG ESSEN

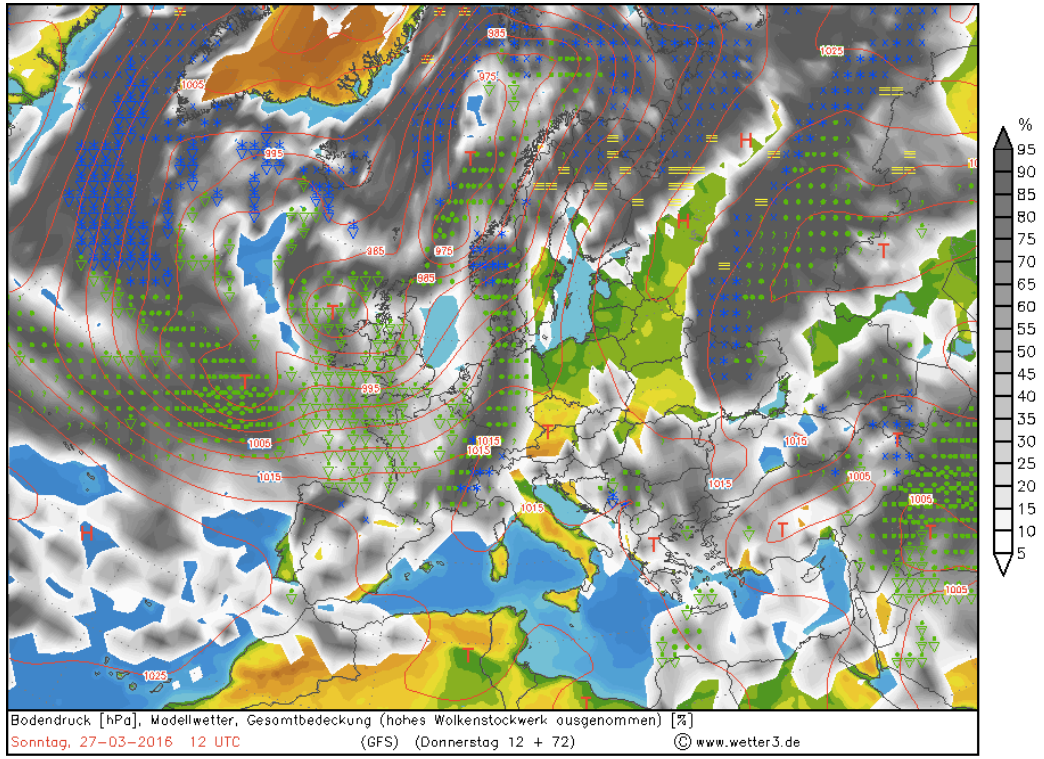
Offen im Denken

- Models need to be implemented on computers
- However, computers only understand binary language, e.g., 0110111
- How can we **encode (a.k.a represent)** instances such that computers can process them?
  - **Features**
- Features are measurable properties that can describe instances in one experiment.
- Features are better to be **informative**, **discriminating** and **independent**.

Example: Weather forecast

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



http://www.orniwetter.info/wp-content/uploads/2016/03/20160327_Modellwetter.gif

Possible Features

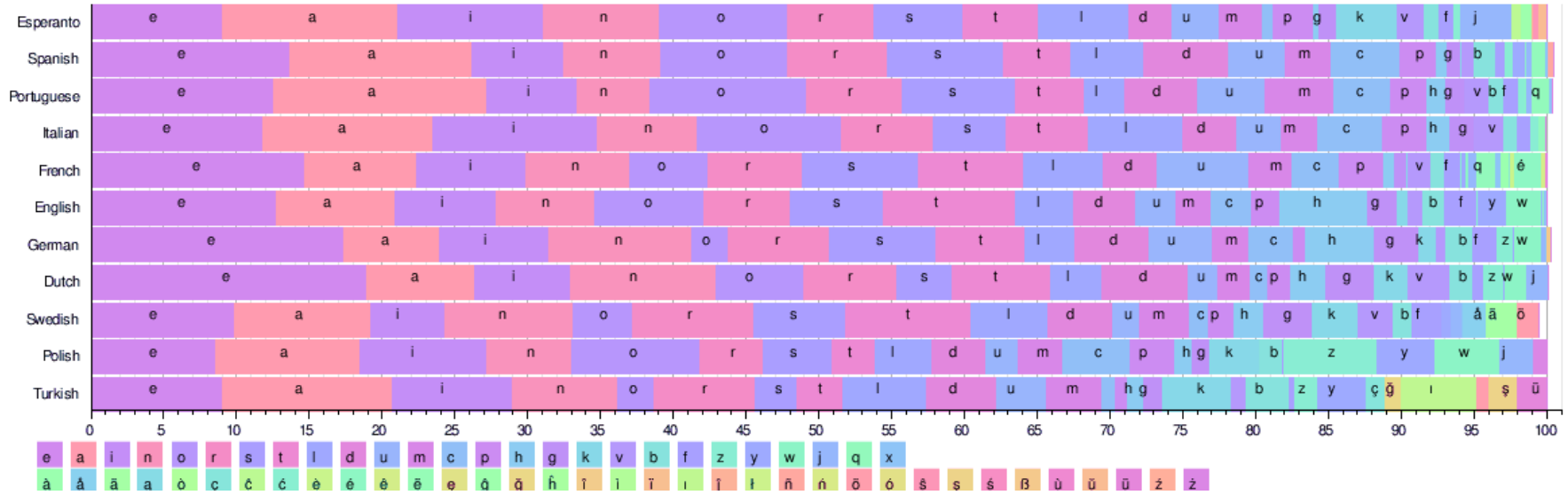
- Yesterday's weather
- Current temperature
- Current humidity
- Air pressure
- …

Possible features for language identification?

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Character set
- Special words
- Length of tokens
- Distribution of characters / character bigrams



Feature vectors

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Features are variables and can take values.
- We can encode the knowledge about each data sample by values of features that we defined for an experiment.
- This process is also known as **feature extraction or encoding.**
- The module that does the process is named **feature extractor** or **encoder.**

Example: Weather forecast

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Possible Features

- F1: Yesterday's highest temperature
- F2: Current temperature
- F3: Did it rain yesterday?
- F4: Air pressure
- …

x_i : weather in day i

ID	F1	F2	F3
1	19	21	Yes
2	21	23	No
3	23	27	No
4	27	26	No

Example: Weather forecast

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Possible Features

- F1: Yesterday's highest temperature
- F2: Current temperature
- F3: Did it rain yesterday?
- …

x_i : weather in day i

Feature vector that Represent x_1

ID	F1	F2	F3
1	19	21	Yes
2	21	23	No
3	23	27	No
4	27	26	No

Missing values

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Missing values for features are a critical issue for many algorithms

→ Replace missing values with mean or median of feature.

→ Apply smoothing algorithm. Many different smoothing algorithms exist. Most simple one: Add-one-smoothing (every feature occurs at least once).

Feature types

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Input (x)	feature vector	y
Hello, Do you want free printr cartriges? Why pay more when you can get them ABSOLUTELY FREE! Just	# "free": 2 YOUR_NAME: 0 # Misspelled: 2 FromFriend?: 0 ...	Spam or NotSpam
	Pixel-7,12: 1 Pixel-7,13: 0 ... # Loops: 0 ...	2

Feature types

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Common feature types:

- **nominal (categorical):**
  - "mode of transportation" —> {car, bus, train, tram, bicycle}
  - there is no agreed way to order these from highest to lowest.
- **binary:**
  - "isEngineer" —> {yes, no}
  - A specific type of categorical features
- **ordinal:**
  - Similar to categorical but there is a clear ordering of the categories
  - "educational experience" —> {elementary school graduate, high school graduate, some college and college graduate}
  - There is an order between values
  - distance between values is identical
- **Numerical**
  - The value is a number in a valid interval
  - "weight" $\in [0,150]$

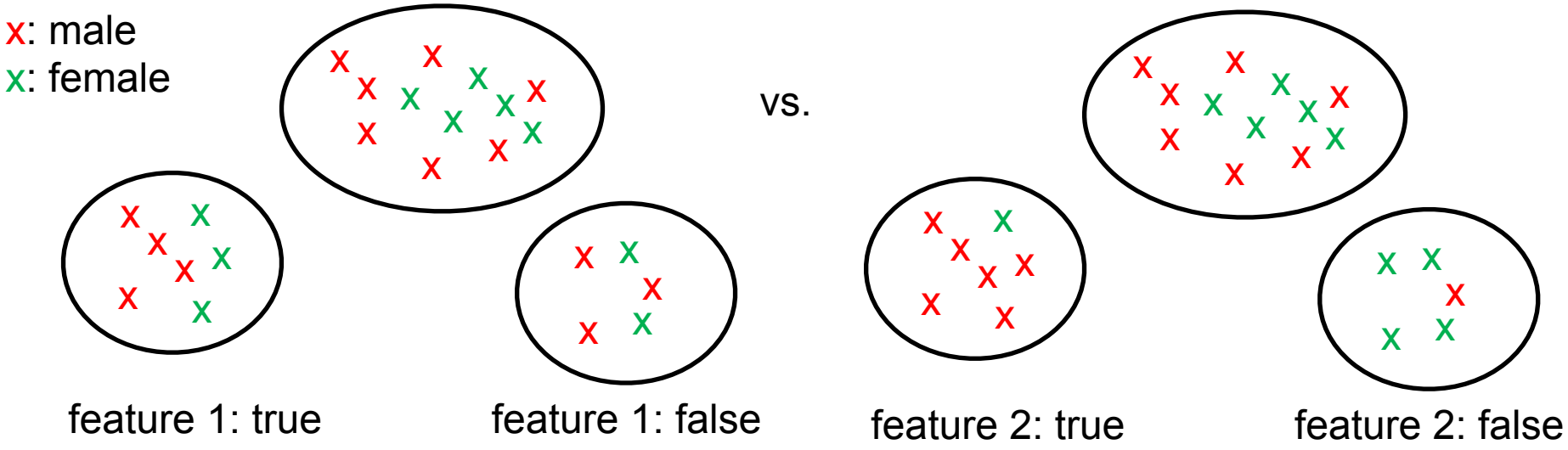
Feature Selection

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- A good feature is able to split the instances into two (or more) parts so that each part is as pure as possible
  → ideally, each part contains only examples of a single class

Entropy: measure of impurity (also: unpredictability, unorderedness, average information content, surprise)



Entropy

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Entropy (H): is the lack of predictability= disorder = diversity = impurity
  - We calculate entropy **over a set of examples** for **the target labels**
  - $$H(y) = -\sum_{y_i \in C} p(y_i) \log p(y_i)$$
  - y indicates the output variable
  - D indicates the given set of examples
  - $p(y_i)$ shows the probability of having label y_i for the output variable (y = y_i)
  - Log is the logarithmic function on base 2

Entropy

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $H(y) = -\sum_{y_i \in C} p(y_i) \log p(y_i)$
- $H(y) = -p(y=0)\log p(y=0) - p(y=1) log p(y=1)$
- $H(y) = -\frac{7}{10} log(\frac{7}{10}) - \frac{3}{10} \log p(\frac{3}{10})$
- $H(y) = -0.7(-0.51) - 0.3(-1.74)$
- $H(y) = 0.36 + 0.52 = 0.88$
- $H(y) = 0.88$

X = has-long-hair	Y
Yes	Female (0)
Yes	Female (0)
Yes	Male (1)
Yes	Female (0)
Yes	Female (0)
Yes	Female (0)
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

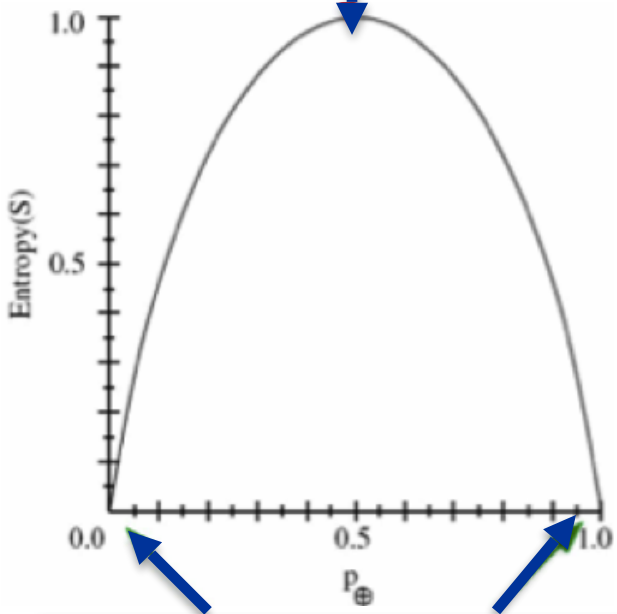
- High H(y) shows high diversity between labels
- Low H(y) shows low diversity

Entropy: interpretation

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Maximal value if labels are equally distributed (high diversity) in the dataset



Minimal value if only one label is in the dataset (low diversity)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = ?$

X = has-long-hair	Y
Yes	Female (0)
Yes	Female (0)
Yes	Male (1)
Yes	Female (0)
Yes	Female (0)
Yes	Female (0)
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = ?$

X = has-long-hair	Y
Yes	Female (0)
Yes	Female (0)
Yes	Male (1)
Yes	Female (0)
Yes	Female (0)
Yes	Female (0)
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = ?$

X = has-long-hair	Y
Yes	Female (0)
Yes	Female (0)
Yes	Male (1)
Yes	Female (0)
Yes	Female (0)
Yes	Female (0)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = -p(y = 0|x = yes)\log p(y = 0|x = yes) - p(y = 1|x = yes)\log p(y = 1|x = yes)$

X = has-long-hair	Y
Yes	Female (0)
Yes	Female (0)
Yes	Male (1)
Yes	Female (0)
Yes	Female (0)
Yes	Female (0)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = -p(y = 0|x = yes)\log p(y = 0|x = yes) - p(y = 1|x = yes)\log p(y = 1|x = yes)$
- $H(y|x = yes) = -\frac{5}{6}\log(\frac{5}{6}) - \frac{1}{6}\log(\frac{1}{6}) = 0.65$

X = has-long-hair	Y
Yes	Female (0)
Yes	Female (0)
Yes	Male (1)
Yes	Female (0)
Yes	Female (0)
Yes	Female (0)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = -p(y = 0|x = yes)\log p(y = 0|x = yes) - p(y = 1|x = yes)\log p(y = 1|x = yes)$
- $H(y|x = yes) = -\frac{5}{6}\log(\frac{5}{6}) - \frac{1}{6}\log(\frac{1}{6}) = 0.65$
- $p(x = no) = ?$

X = has-long-hair	Y
Yes	Female (0)
Yes	Female (0)
Yes	Male (1)
Yes	Female (0)
Yes	Female (0)
Yes	Female (0)
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = -p(y = 0|x = yes)\log p(y = 0|x = yes) - p(y = 1|x = yes)\log p(y = 1|x = yes)$
- $H(y|x = yes) = -\frac{5}{6}\log(\frac{5}{6}) - \frac{1}{6}\log(\frac{1}{6}) = 0.20$
- $p(x = no) = \frac{4}{10} = 0.4$

X = has-long-hair	Y
Yes	Female (0)
Yes	Female (0)
Yes	Male (1)
Yes	Female (0)
Yes	Female (0)
Yes	Female (0)
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = -p(y = 0|x = yes)\log p(y = 0|x = yes) - p(y = 1|x = yes)\log p(y = 1|x = yes)$
- $H(y|x = yes) = -\frac{5}{6}\log(\frac{5}{6}) - \frac{1}{6}\log(\frac{1}{6}) = 0.20$
- $p(x = no) = \frac{4}{10} = 0.4$
- $H(y|x = no) = ?$

X = has-long-hair	Y
Yes	Female (0)
Yes	Female (0)
Yes	Male (1)
Yes	Female (0)
Yes	Female (0)
Yes	Female (0)
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = -p(y = 0|x = yes)\log p(y = 0|x = yes) - p(y = 1|x = yes)\log p(y = 1|x = yes)$
- $H(y|x = yes) = -\frac{5}{6}\log(\frac{5}{6}) - \frac{1}{6}\log(\frac{1}{6}) = 0.20$
- $p(x = no) = \frac{4}{10} = 0.4$
- $H(y|x = no) = ?$

X = has-long-hair	Y
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = -p(y = 0|x = yes)\log p(y = 0|x = yes) - p(y = 1|x = yes)\log p(y = 1|x = yes)$
- $H(y|x = yes) = -\frac{5}{6}\log(\frac{5}{6}) - \frac{1}{6}\log(\frac{1}{6}) = 0.20$
- $p(x = no) = \frac{4}{10} = 0.4$
- $H(y|x = no) = -p(y = 0|x = no)\log p(y = 0|x = no) - p(y = 1|x = no)\log p(y = 1|x = no)$

X = has-long-hair	Y
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = -p(y = 0|x = yes)\log p(y = 0|x = yes) - p(y = 1|x = yes)\log p(y = 1|x = yes)$
- $H(y|x = yes) = -\frac{5}{6}\log(\frac{5}{6}) - \frac{1}{6}\log(\frac{1}{6}) = 0.20$
- $p(x = no) = \frac{4}{10} = 0.4$
- $H(y|x = no) = -p(y = 0|x = no)\log p(y = 0|x = no) - p(y = 1|x = no)\log p(y = 1|x = no)$
- $H(y|x = no) = -\frac{2}{4}\log(\frac{2}{4}) - \frac{2}{4}\log(\frac{2}{4}) = 1.0$

X = has-long-hair	Y
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $p(x = yes) = \frac{6}{10} = 0.6$
- $H(y|x = yes) = -p(y = 0|x = yes)\log p(y = 0|x = yes) - p(y = 1|x = yes)\log p(y = 1|x = yes)$
- $H(y|x = yes) = -\frac{5}{6}\log(\frac{5}{6}) - \frac{1}{6}\log(\frac{1}{6}) = 0.20$
- $p(x = no) = \frac{4}{10} = 0.4$
- $H(y|x = no) = -p(y = 0|x = no)\log p(y = 0|x = no) - p(y = 1|x = no)\log p(y = 1|x = no)$
- $H(y|x = no) = -\frac{2}{4}\log(\frac{2}{4}) - \frac{2}{4}\log(\frac{2}{4}) = 1.0$
- $I(x) = 0.6(0.2) + 0.4 * (1.0) = 0.12 + 0.4 = 0.52$

X = has-long-hair	Y
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

Information Gain When We Consider a Feature x

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- $IG(x) = H(y) - I(x)$
- $H(y)$: is the entropy of the given dataset without considering feature x
- $I(x)$: is the **average entropy** of the given set when it is split with respect to **values of feature x**
- $I(x) = \sum p(x)H(y|x)$
- Let's compute I(x)
- $I(x) = p(x = yes)H(y|x = yes) + p(x = no)H(y|x = no)$
- $I(x) = 0.52$
- $H(y) = 0.88$
- IG(x) = H(y) - I(x) = 0.88 - 0.52 = 0.36

- Features with higher information gain are better representative of data samples.

X = has-long-hair	Y
Yes	Female (0)
Yes	Female (0)
Yes	Male (1)
Yes	Female (0)
Yes	Female (0)
Yes	Female (0)
No	Male (1)
No	Female (0)
No	Female (0)
No	Male (1)

Information gain for feature selection

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- The information gain of a feature is the **change in entropy** when applying this feature.
- It can be **used to automatically reduce the feature space** to the most **predictive features**.

Filter-based feature selection:

- apply a filter (e.g. information gain) on features
- select only n-best
- train new machine learning model
- evaluate

Data

Getting data

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- We should **get some data samples that represent the task** as it may happen in **inference phase**
- Our data have to contain instances to represent all possible outcome
- **First** try to **find an existing and standard benchmark dataset** (often not a perfect fit)
- We can **collect data from scratch (very time-consuming)**

Where to get the ground truth labels?

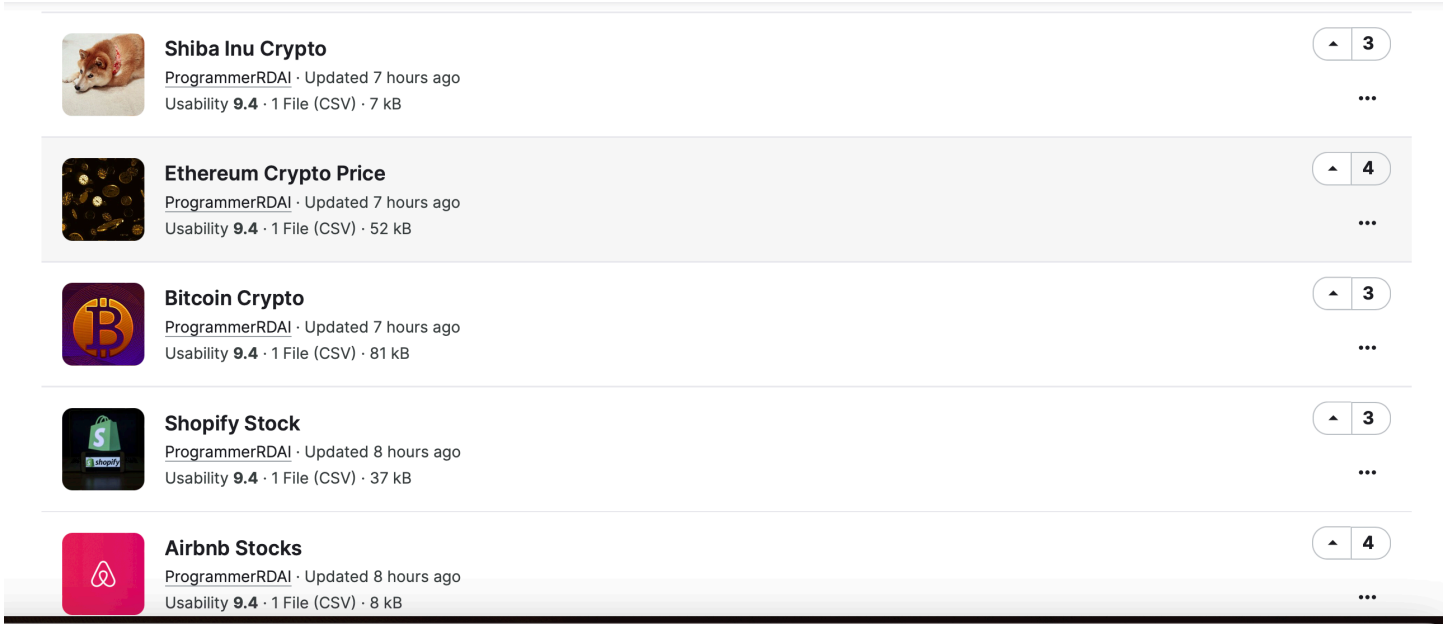
- Often manually annotated by experts *(Is this email spam or not?)*
- Most ML algorithms need a considerable amount of annotated data
- How much exactly depends on the problem

Example sources to get datasets

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Huggingface: (https://huggingface.co/datasets)
  - SMS_SPAM (https://huggingface.co/datasets/sms_spam/viewer/plain_text/train)
- Kaggle: (https://www.kaggle.com/datasets)



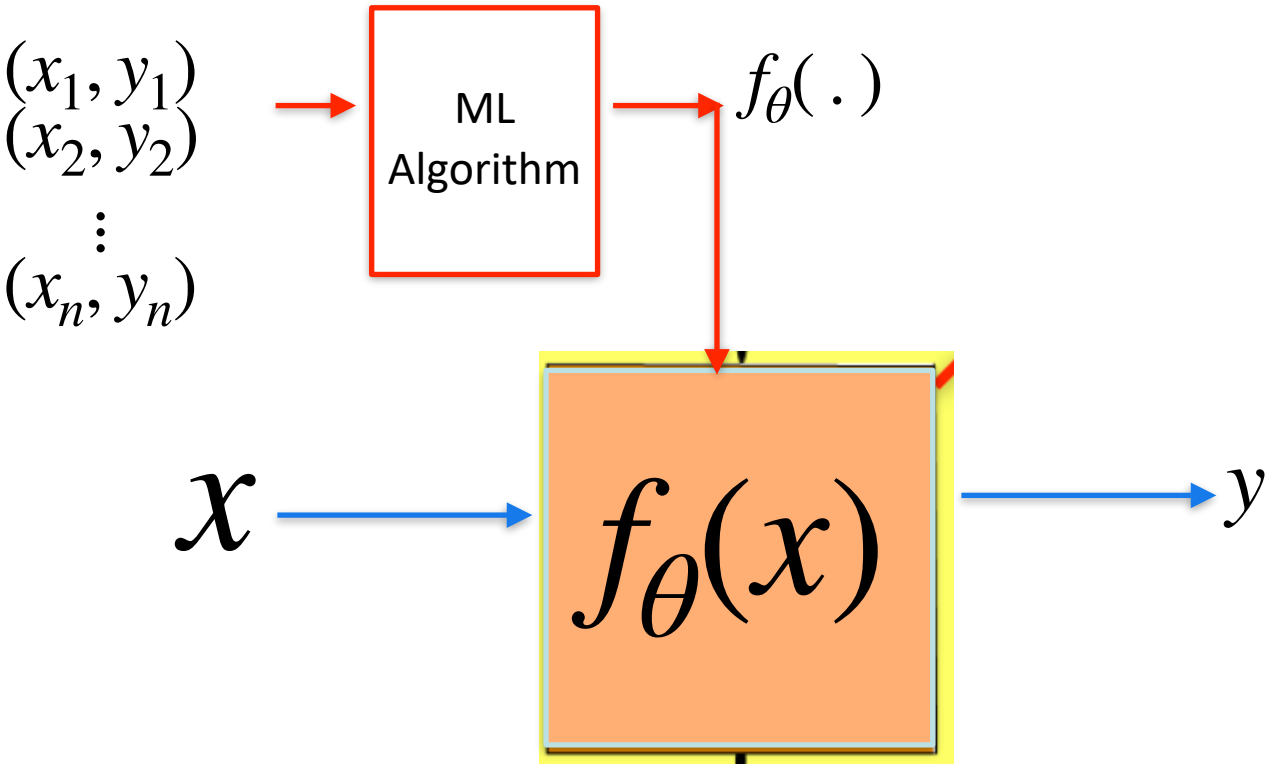
Model Training

ML

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

How to use data samples in a dataset to learn function $f_\theta(\,.\,)$?



Supervised Learning

SL

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Training data samples x_i with reference labels y_i
- SL algorithms try to learn function $f_\theta(x)$ that approximates the data

$$x_1, y_1, \hat{y}_1 = f_\theta(x_1)$$
$$x_2, y_2, \hat{y}_2 = f_\theta(x_2)$$
$$\vdots$$
$$x_n, y_n, \hat{y}_n = f_\theta(x_n)$$

such that

The sum of the errors between $f_\theta(x_i)$ and y_i is minimum

The learned function predict y for x which has not been seen in the training data.

SL

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Training data samples x_i with reference labels y_i
- SL algorithms try to learn function $f_\theta(x)$ that approximates the data

data samples

$$\begin{aligned} x_1,\ & y_1, \hat{y}_1 = f_\theta(x_1) \\ x_2,\ & y_2, \hat{y}_2 = f_\theta(x_2) \\ & \vdots \\ x_n,\ & y_n, \hat{y}_n = f_\theta(x_n) \end{aligned}$$

such that

The sum of the errors between $f_\theta(x_i)$ and y_i is minimum

The learned function predict y for x which has not been seen in the training data.

SL

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Training data samples x_i with reference labels y_i
- SL algorithms try to learn function $f_\theta(x)$ that approximates the data

Reference outcome

$$x_1, y_1, \hat{y}_1 = f_\theta(x_1)$$
$$x_2, y_2, \hat{y}_2 = f_\theta(x_2)$$
$$\vdots$$
$$x_n, y_n, \hat{y}_n = f_\theta(x_n)$$

such that

The sum of the errors between $f_\theta(x_i)$ and y_i is minimum

The learned function predict y for x which has not been seen in the training data.

SL

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Training data samples x_i with reference labels y_i
- SL algorithms try to learn function $f_\theta(x)$ that approximates the data

$$
\begin{aligned}
x_1, y_1, &\hat{y}_1 = f_\theta(x_1) \\
x_2, y_2, &\hat{y}_2 = f_\theta(x_2) \\
&\vdots \\
x_n, y_n, &\hat{y}_n = f_\theta(x_n)
\end{aligned}
$$

Model's predictions

such that

The sum of the errors between $f_\theta(x_i)$ and y_i is minimum

The learned function predict y for x which has not been seen in the training data.

SL application

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- SL has applications in building models that deal with two major categories of problems
  - **Classification**
  - **Regression**

SL application

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

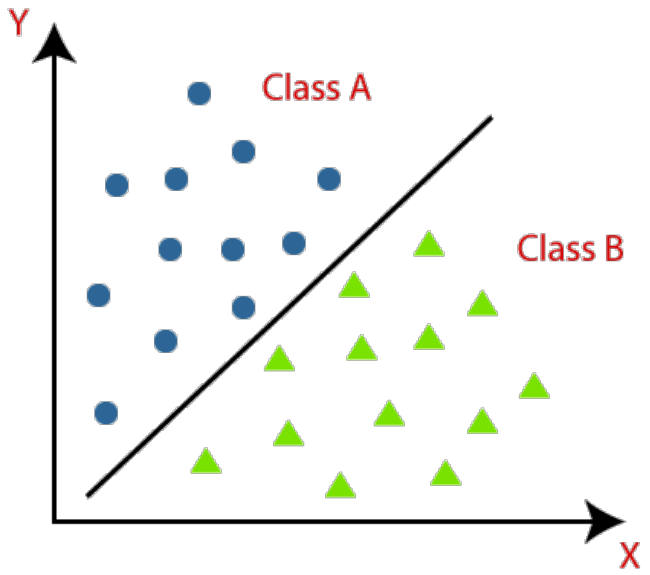
- SL has applications in building models that deal with two major categories of problems
  - **Classification**
  - Regression

Classification

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **Goal:** Assigning an input data sample x into a category
  - y is a categorical variable (or feature)



Classification

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **Goal:** Assigning an input data sample x into a category
- In these problems, y is a categorical variable (or feature)

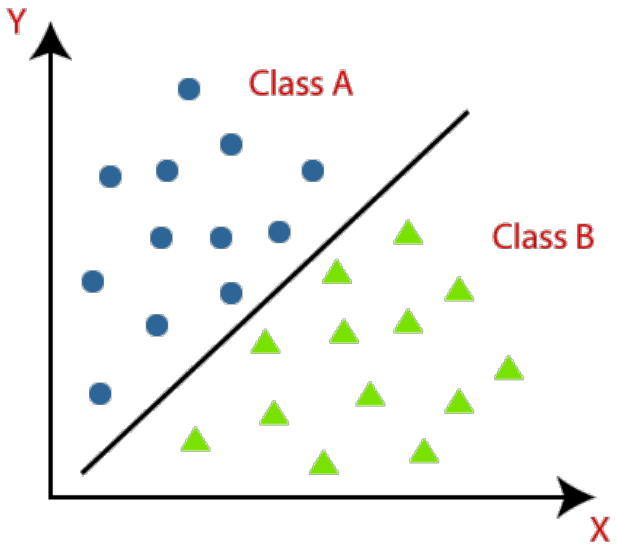
- Different types of classification problems:
  - **Binary classification:** predict one of two classes
  - **Multi-class classification:** predict one of more than two classes
  - **Multi-label classification:** predict one or more classes for each example
  - **Imbalanced classification:** classification tasks where the distribution of examples across the class labels is not equal.

Binary classification

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- In these problems, y is a binary variable (or feature)



- Examples:
  - Is this image a cat or dog?
  - Is this movie review positive or negative?

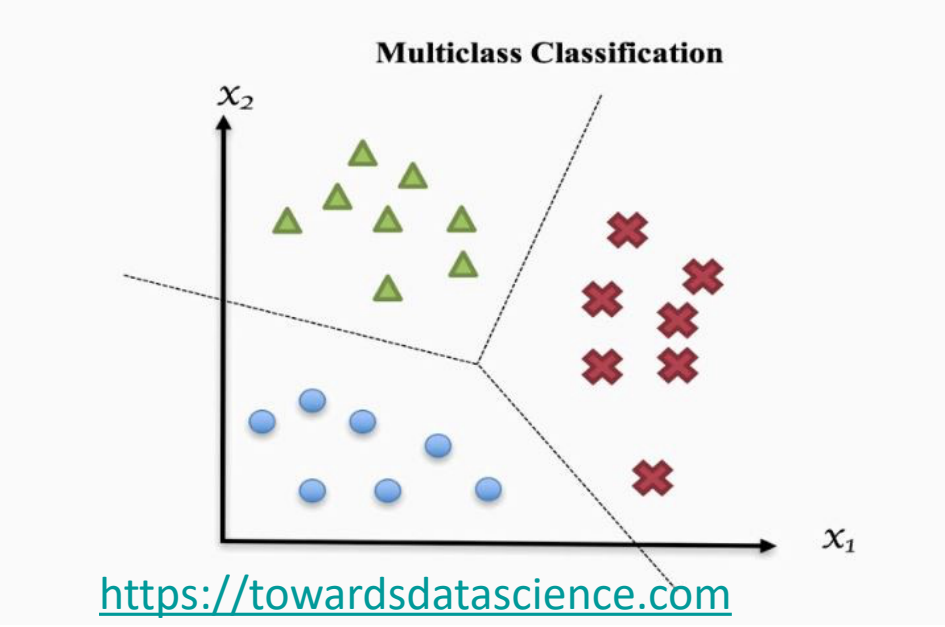
Multi-Class classification

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- In these problems, the number of possible labels is more than two

- **Examples**:
  - Is this image a cat or dog or panda?
  - Is this movie review very positive, positive, neutral, negative or very negative?



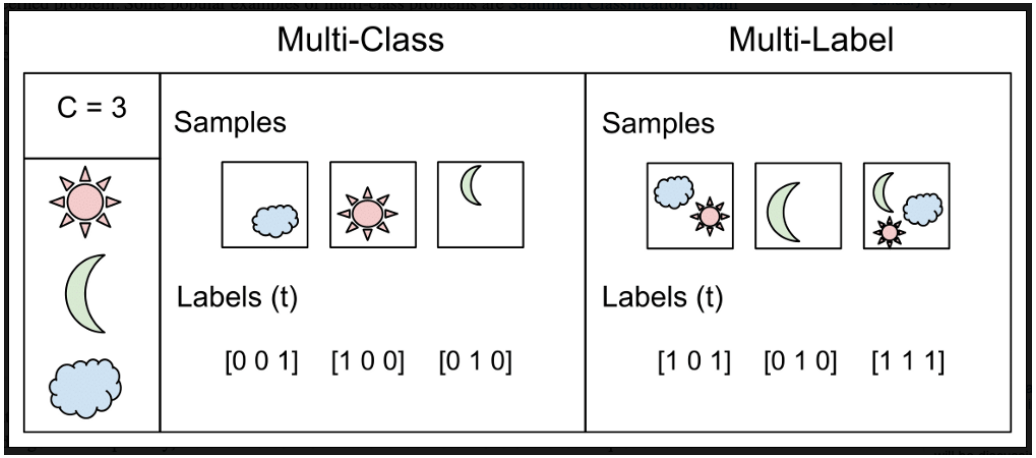
https://towardsdatascience.com

Multi-label classification

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- In these problems, each sample can be assigned to multiple class labels



https://medium.com

- **Examples**:
  - What entities do exist in an image? (cat, dog, pandas, …)
  - What is this movie review about? (author, director, Costume Designer, …)

SL application

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

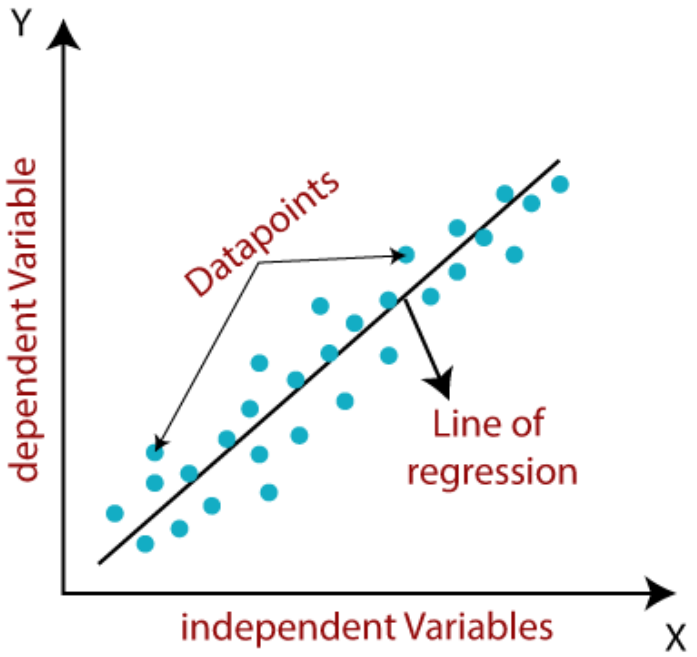
- SL has applications in building models that deal with two major categories of problems
  - Classification
  - **Regression**

Regression

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- An algorithm to model the relationship between dependent and independent variables
- Some well-known algorithms
  - Linear regression
  - Polynomial regression



https://www.javatpoint.com/

Unsupervised Learning

Unsupervised learning

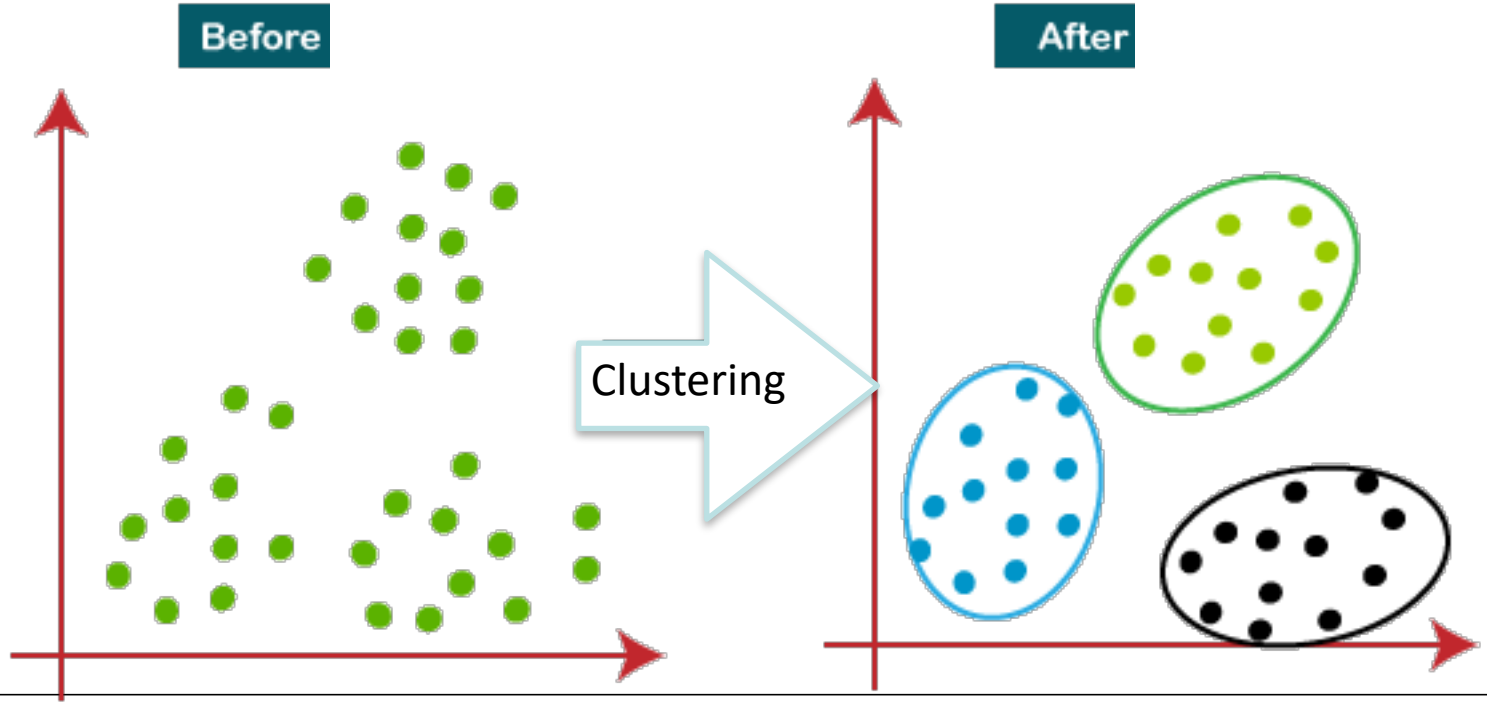
- Unsupervised learning algorithms find a function that analyzes and clusters unlabeled datasets.
- These algorithms discover hidden patterns or data groupings without the need for human intervention.
- Useful for tasks that need
  - Clustering
  - Dimensionality reduction

Clustering

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Clustering algorithms group unlabeled data based on their similarities or differences
- Clustering algorithms process raw, unclassified data objects into groups represented by structures or patterns in the information.

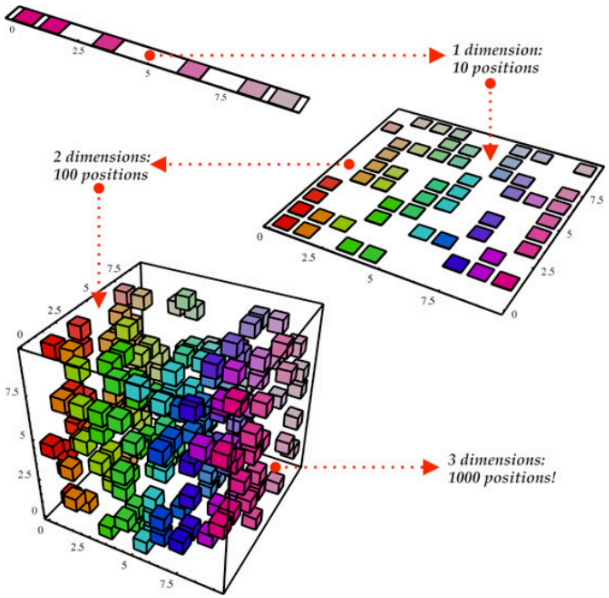


Dimensionality reduction

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Dimensionality reduction algorithms help to reduce the number of data inputs to a manageable size while also preserving the integrity of the dataset as much as possible.
- These algorithms are used when the number of features, or dimensions, in a given dataset is too high.



www.medium.com

Evaluation

How Good is my Model?

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Test the model with new data.

- Provide an instance with its features, but without the label.
- *Dear model, what do you do now?*

→ The model uses the knowledge captured in its parameters to return an outcome.

Evaluate the performance of the model on new data using some examples.

What if performance not great? Improve!

- Change features
- Change classifier

Cross-validation

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- k -fold cross-validation: method to evaluate the model on the training set

1. Partition the data set into k parts of equal size.
2. Train the model on all k-1 parts and test it on the k^{th} part.
   → fold 1
3. Repeat this so that each part forms part of the test set once.
   → k folds
4. Collect the predictions for all folds and calculate the evaluation score.

- n = # of instances in training data
- If $k = n$ → leave-one-out cross-validation

5-fold cross-validation animation

UNIVERSITÄT DUISBURG ESSEN

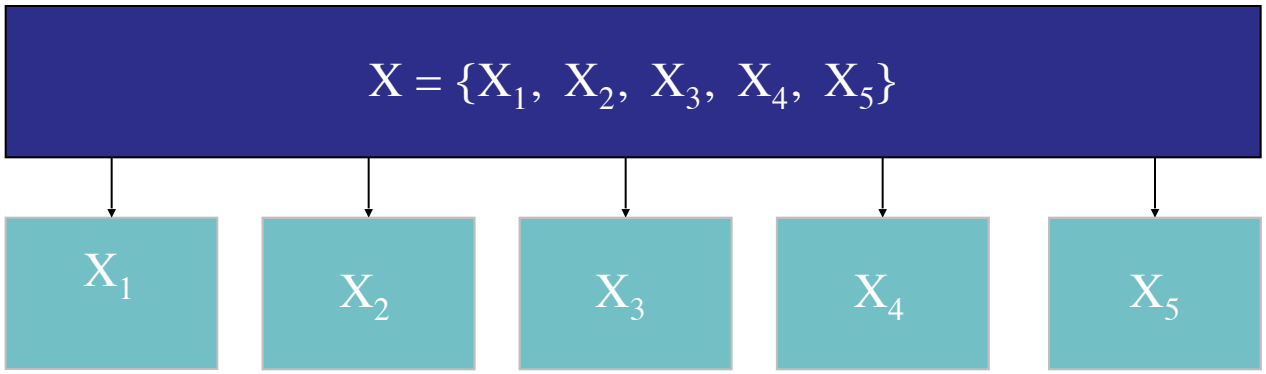
Offen im Denken

$$X = \{X_1,\ X_2,\ X_3,\ X_4,\ X_5\}$$

5-fold cross-validation animation

UNIVERSITÄT DUISBURG ESSEN

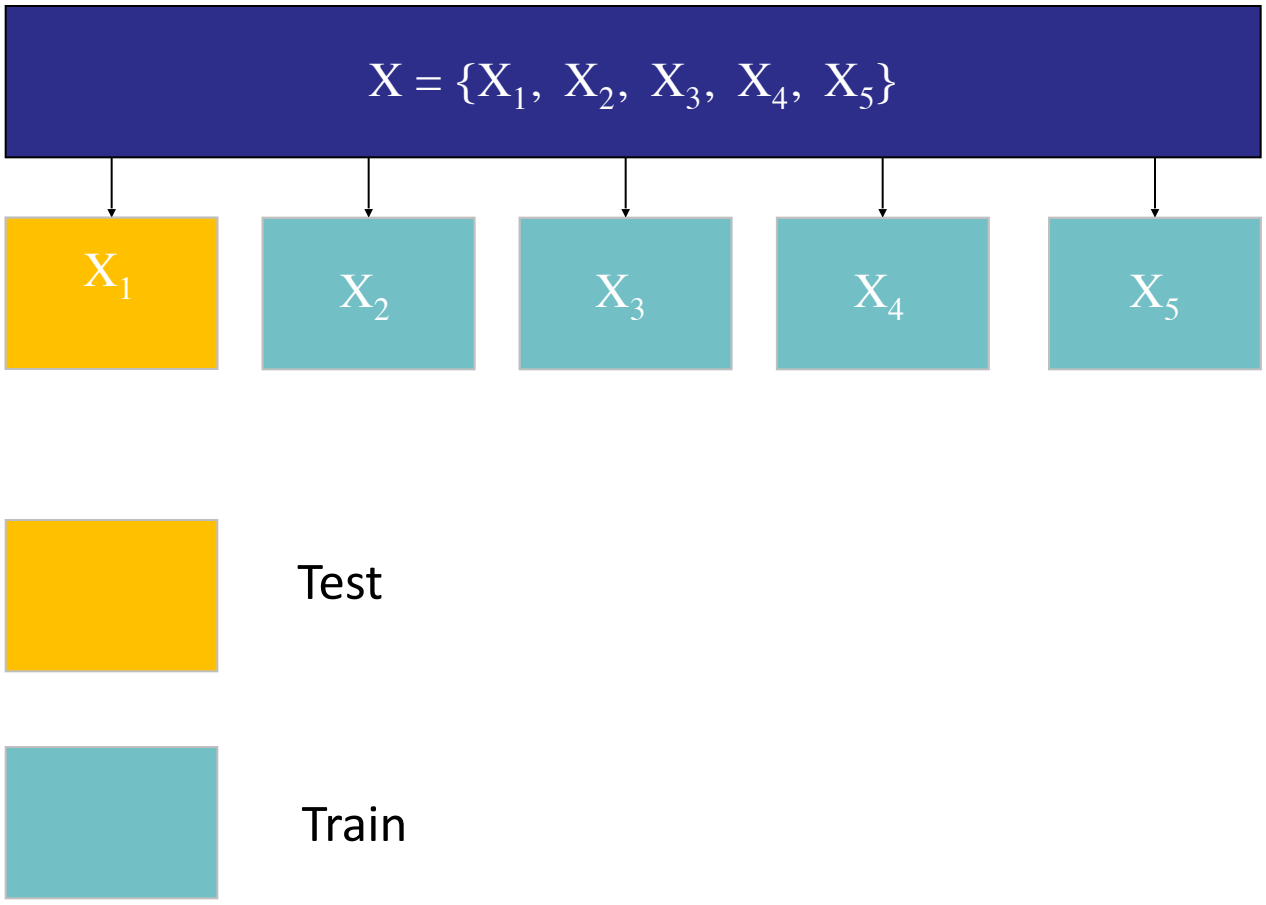
Offen im Denken



5-fold cross-validation animation

UNIVERSITÄT DUISBURG ESSEN

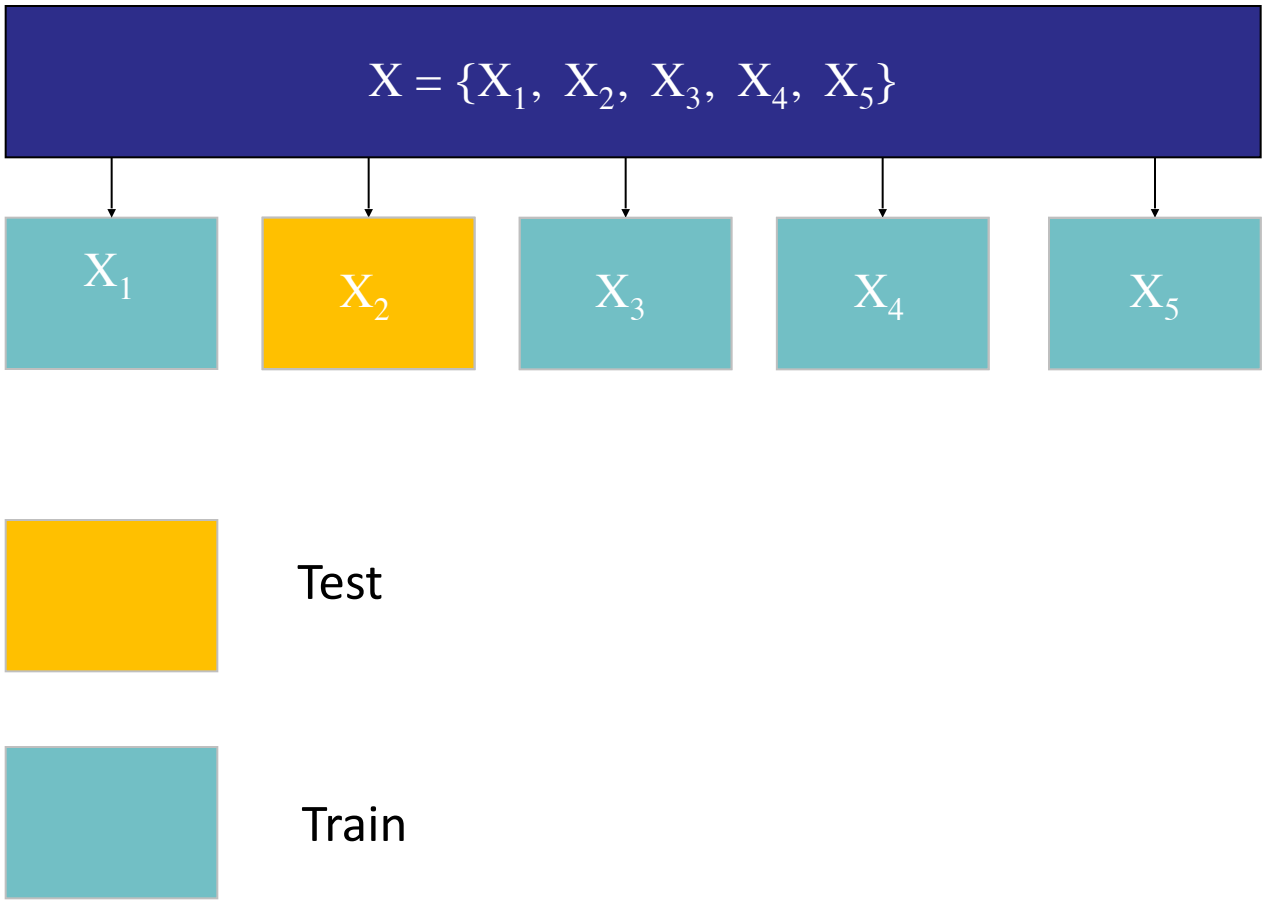
Offen im Denken



5-fold cross-validation animation

UNIVERSITÄT DUISBURG ESSEN

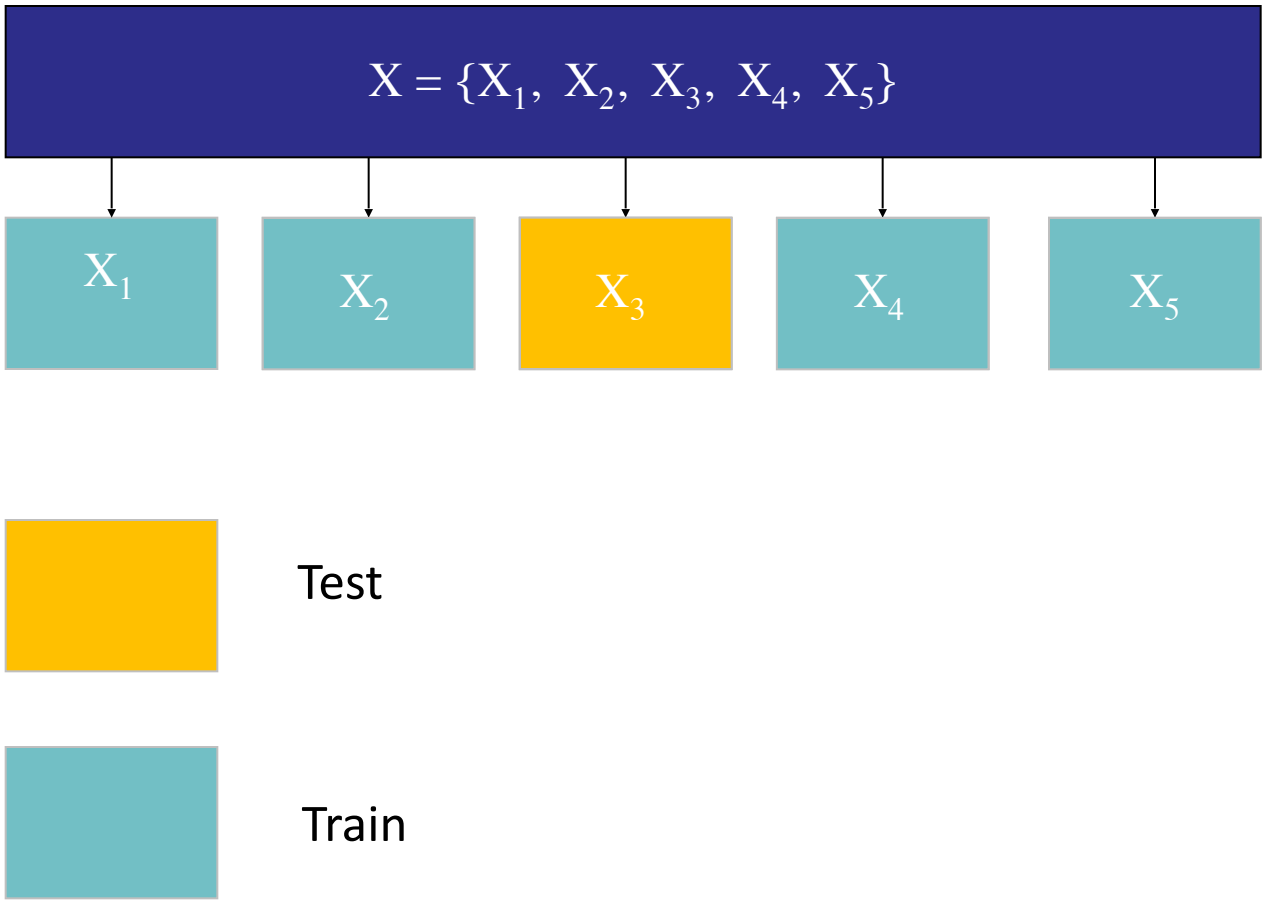
Offen im Denken



5-fold cross-validation animation

UNIVERSITÄT DUISBURG ESSEN

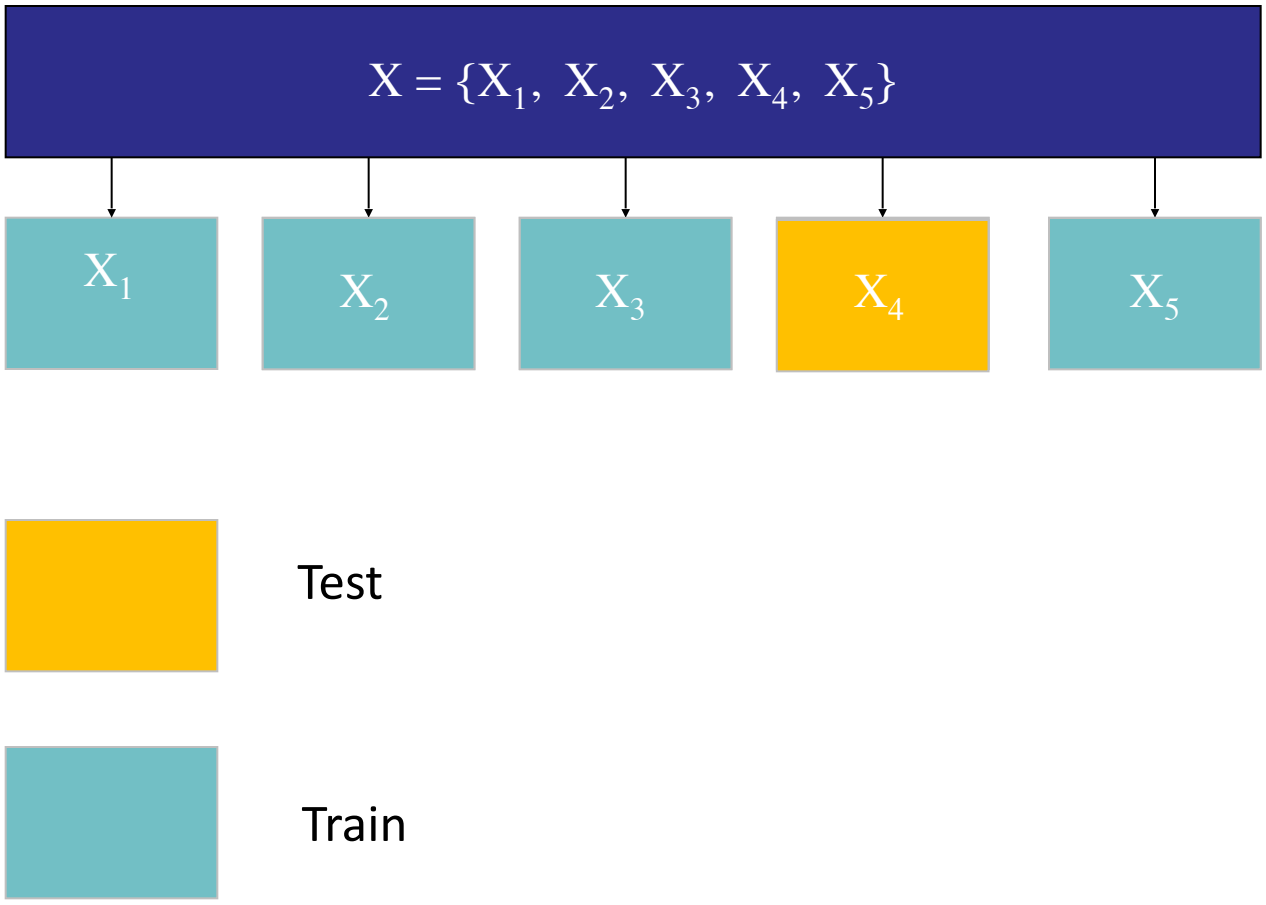
Offen im Denken



5-fold cross-validation animation

UNIVERSITÄT DUISBURG ESSEN

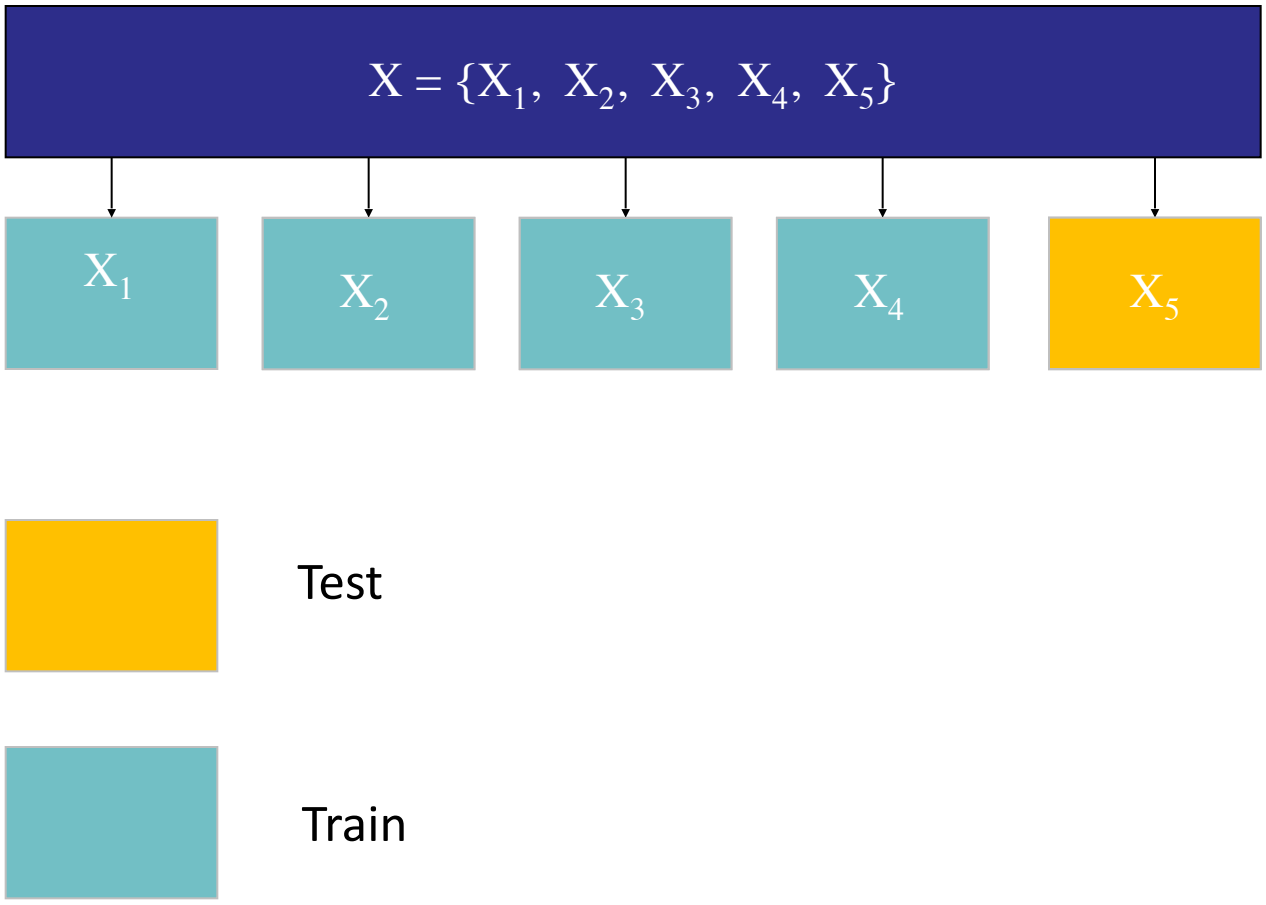
Offen im Denken



5-fold cross-validation animation

UNIVERSITÄT DUISBURG ESSEN

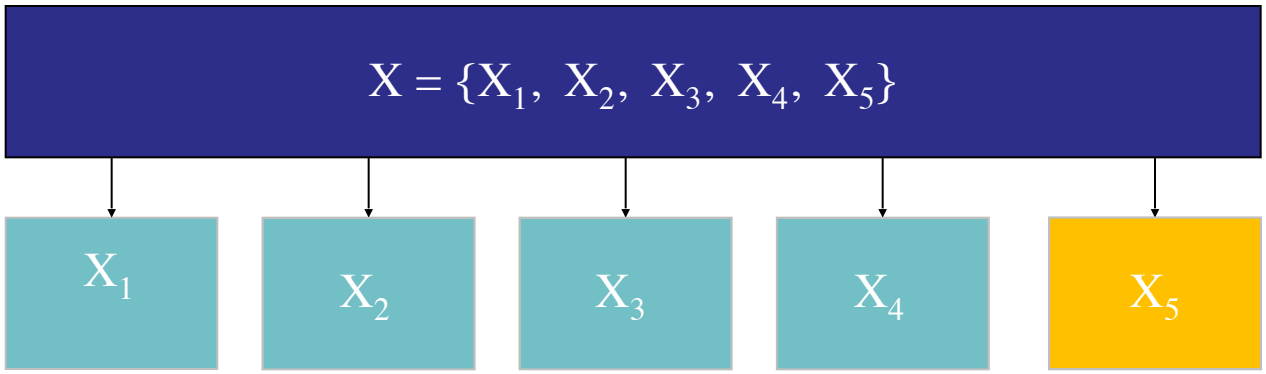
Offen im Denken



5-fold cross-validation animation

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



Test

Train

Result: combine the predictions from each run and calculate the evaluation score once.

Test set

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- You can evaluate your model on totally unseen set of samples
- This set is called test set.

- A test set should be representative of the task
  - Comparable class distribution

- The test set needs to be big enough.
  - Commonly: 10% of whole amount of the data given
  - 80% can be used for training

Accuracy Metric

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Accuracy = $\frac{\text{correctly classified instances}}{\text{all instances}} = \frac{TP + TN}{TP + FP + TN + FN}$

- Confusion matrix:

		Predicted class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

- In most tasks, the classes are imbalanced.
  - Spam classification: only 5% of mails are spam.
  - Classify everything as "not spam" → spam detector with 95% accuracy
  - The other way around would be even worse.

http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Class-imbalance Problem

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Example: Language Identification

Scenario A	Scenario B
99 out of 100 texts are English. Accuracy of 99% when always outputting “English”.	2 out of 100 texts are English. Accuracy of 2% when always outputting “English”.

Precision, Recall, F_1-measure

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

		Predicted class	
		Class = Yes	Class = No
Actual Class	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

- Precision $P = \frac{TP}{TP + FP}$ → How correct are the decisions?
- Recall $R = \frac{TP}{TP + FN}$ → How many of the interesting cases do we catch?
- Higher recall usually leads to lower precision.
- F_1-measure $F1 = \frac{2 * P * R}{P + R}$ → harmonic mean of precision and recall

http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Example: Confusion Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Instances with ground truth (true) labels

Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	0 0	0
	Negative	0	0

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix

Instances with ground truth (true) labels

Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	0 0	0
	Negative	0	0

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix

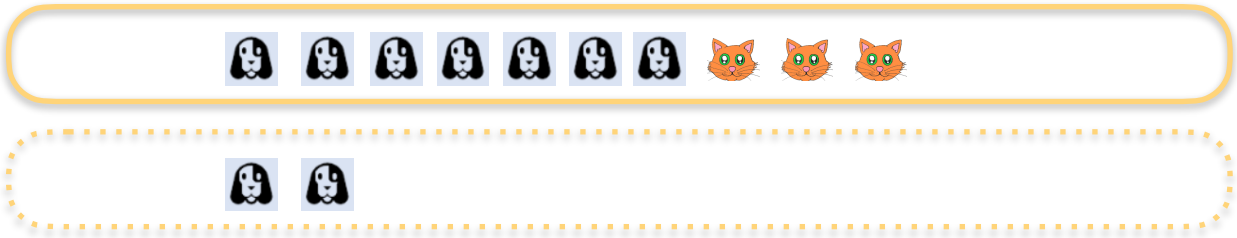
Instances with ground truth (true) labels

Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	0 1	0
	Negative	0	0

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix



Instances with ground truth (true) labels

Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	0 2	0
	Negative	0	0

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Instances with ground truth (true) labels

Predictions

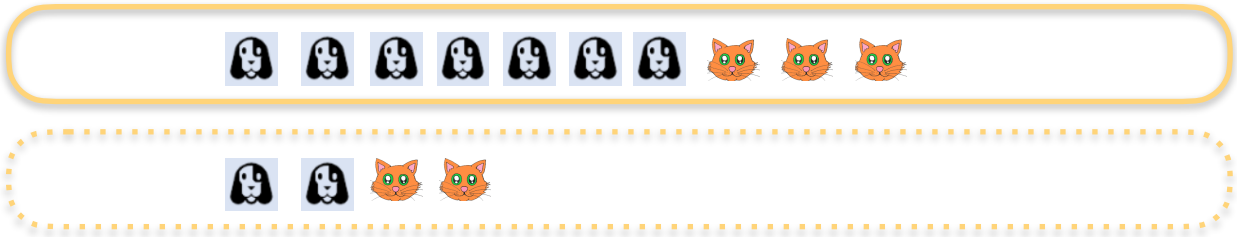
		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	2	0
	Negative	1	0

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



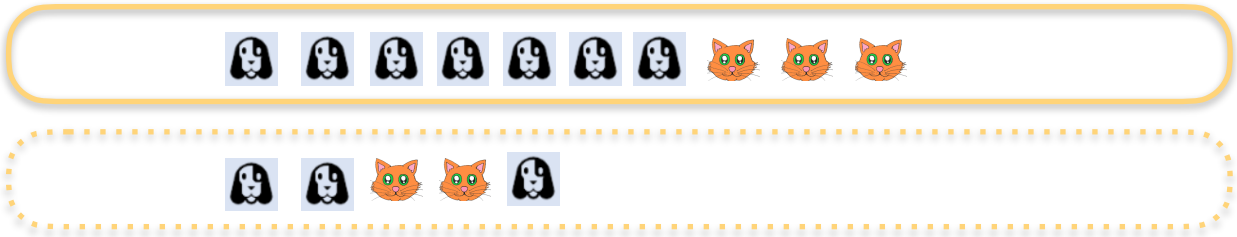
Instances with ground truth (true) labels

Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	2	0
	Negative	2	0

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix



Instances with ground truth (true) labels

Predictions

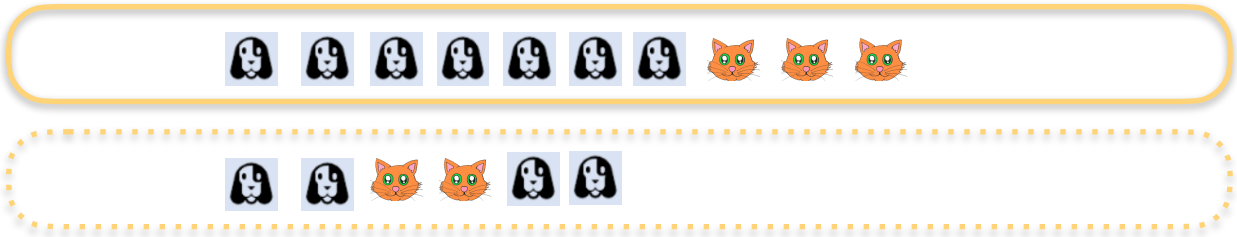
		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	0 3	0
	Negative	2	0

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



Instances with ground truth (true) labels

Predictions

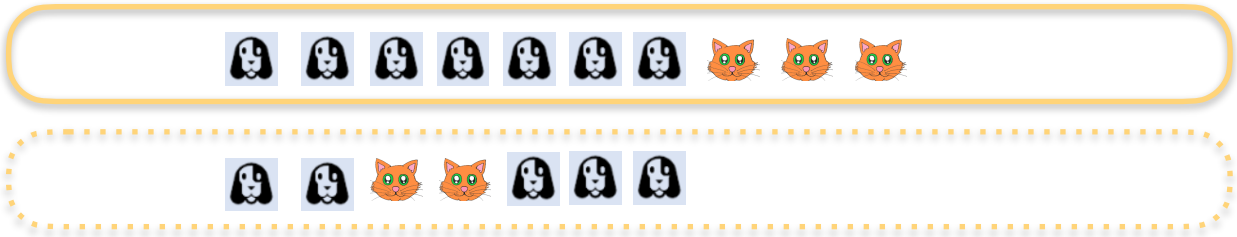
		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	4	0
	Negative	2	0

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



Instances with ground truth (true) labels

Predictions

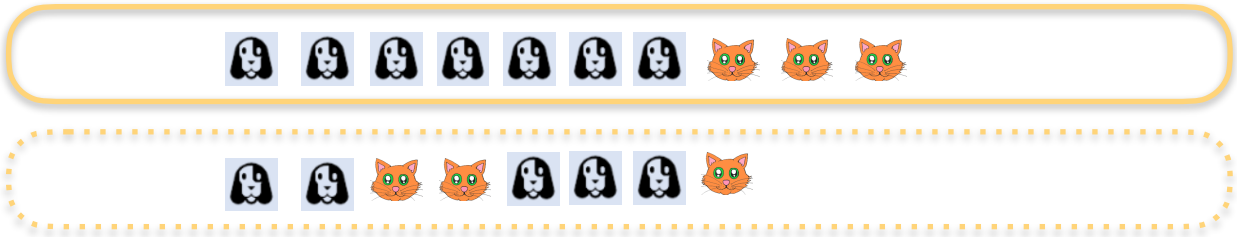
		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5	0
	Negative	2	0

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



Instances with ground truth (true) labels

Predictions

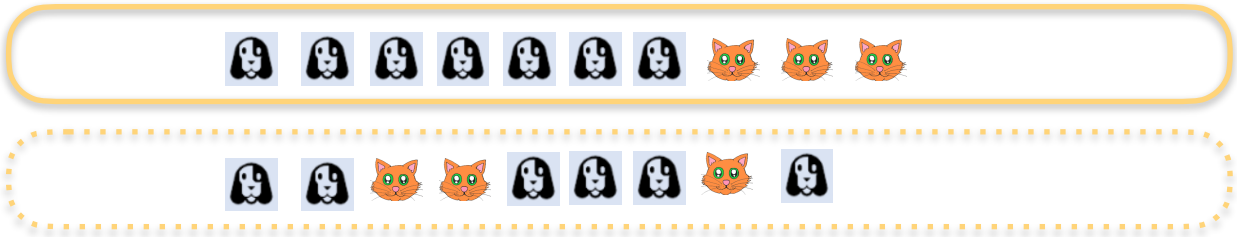
		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	0 5	0
	Negative	2	1

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



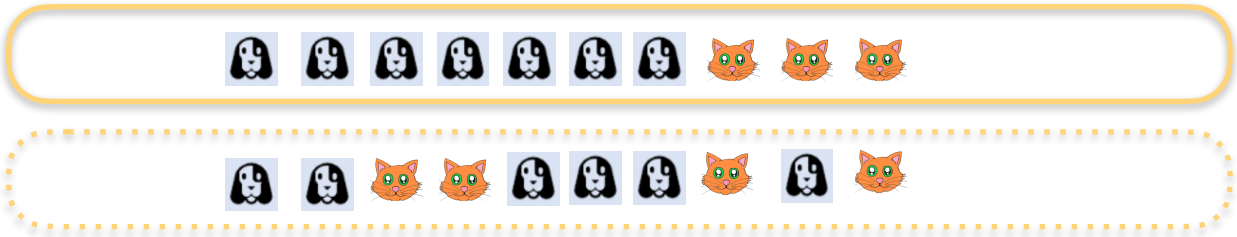
Instances with ground truth (true) labels

Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	0 5	1
	Negative	2	1

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix



Instances with ground truth (true) labels

Predictions

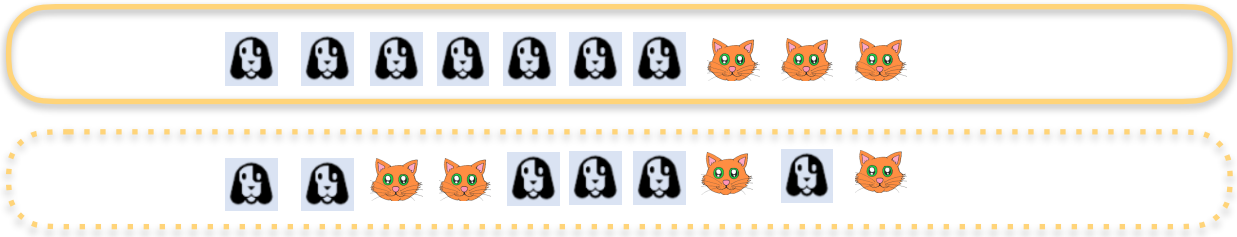
		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	0 5	1
	Negative	2	2

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



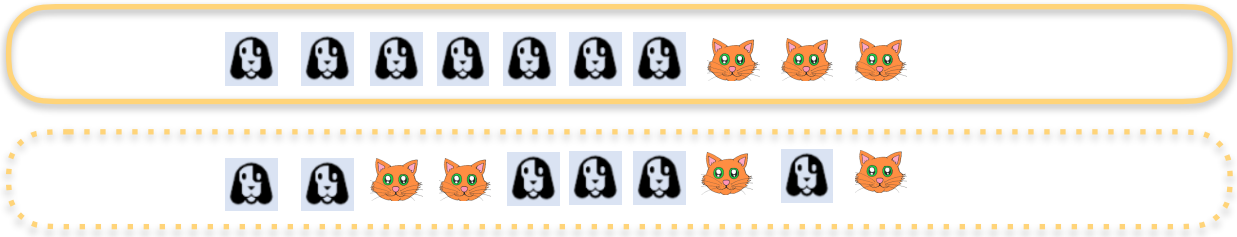
Instances with ground truth (true) labels

Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	0 5 (TP)	1
	Negative	2	2 (TN)

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Confusion Matrix



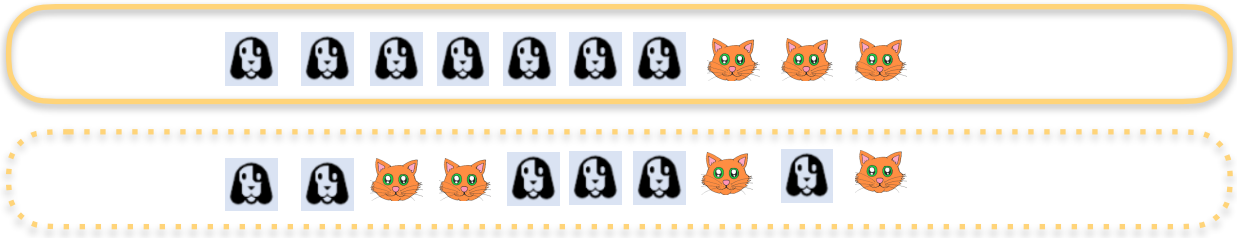
Instances with ground truth (true) labels

Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	0 5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Accuracy



Instances with ground truth (true) labels

Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1
	Negative	2	2 (TN)

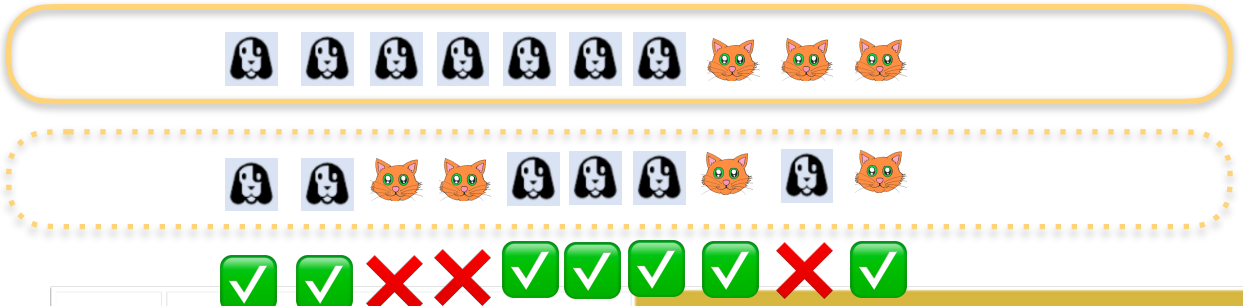
$$acc = \frac{5+2}{10} = \frac{7}{10} = 70\,\%$$

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Accuracy

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



Instances with ground truth (true) labels

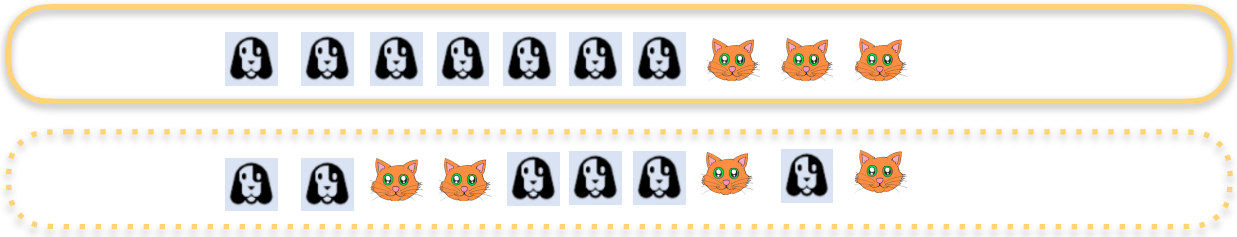
Predictions

		True/Actual: Positive ( )	True/Actual: Negative
Predicted	Positive ( )	5 (TP)	1
Predicted	Negative	2	2 (TN)

$$acc = \frac{5+2}{10} = 70\,\%$$

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Precision



Instances with ground truth (true) labels

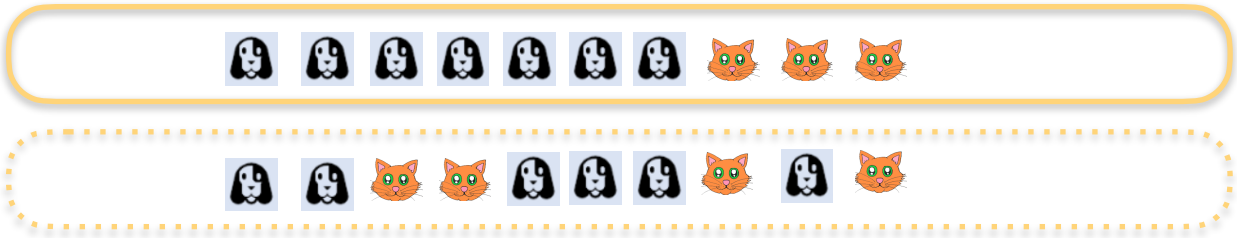
Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

Among all the instances I predicted positive (label 🐶), how often am I right?

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Precision



Instances with ground truth (true) labels

Predictions

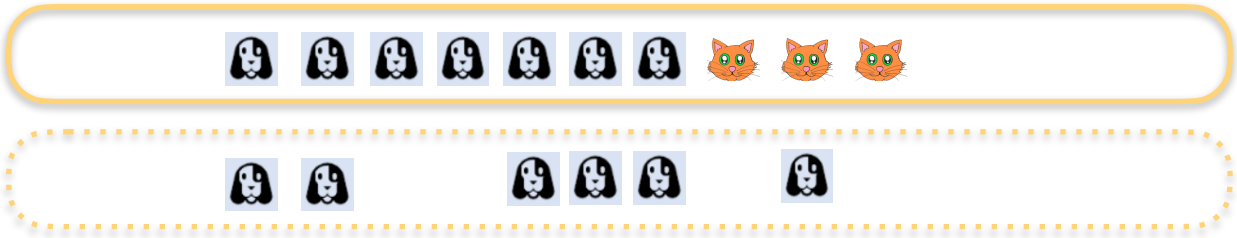
		True/Actual	
		Positive ( )	Negative
Predicted	Positive ( )	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

$$P = \frac{TP}{TP + FP} = \frac{5}{5+1} = \frac{5}{6}$$

Among all the instances I predicted positive (label), how often was I right?

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Precision



Instances with ground truth (true) labels

Predictions

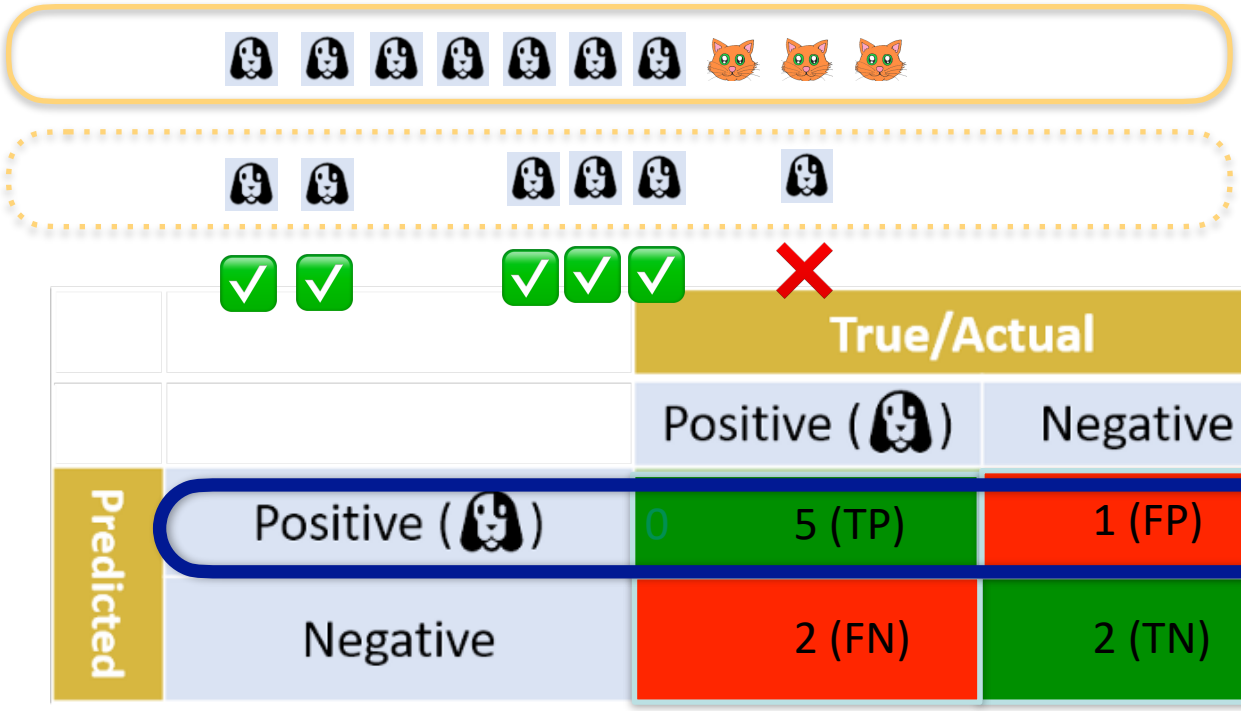
		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

$$P = \frac{TP}{TP + FP} = \frac{5}{5+1} = \frac{5}{6}$$

Among all the instances I predicted positive (label 🐶), how often was I right?

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Precision

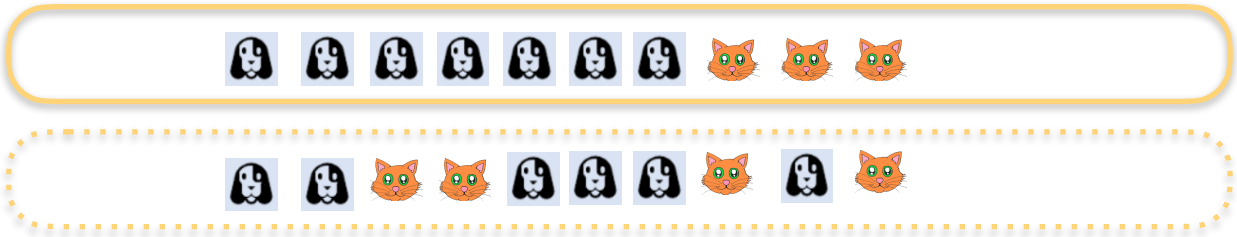


$$P = \frac{TP}{TP + FP} = \frac{5}{5+1} = \frac{5}{6} = 83.3\,\%$$

Among all the instances I predicted positive (label), how often was I right?

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Recall



Instances with ground truth (true) labels

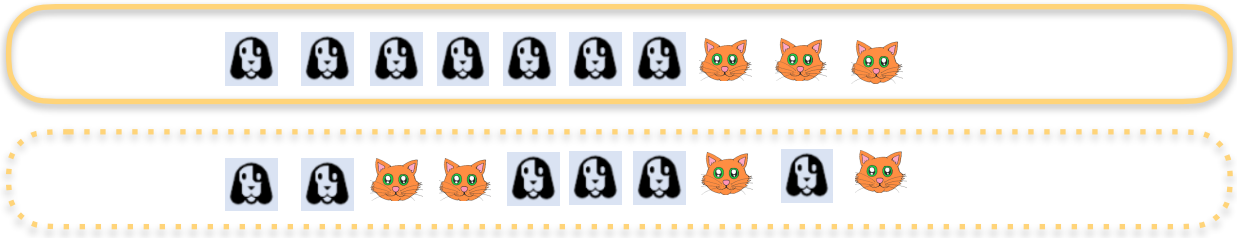
Predictions

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

How many of positive instances from ground truth could I find?

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Recall



Instances with ground truth (true) labels

Predictions

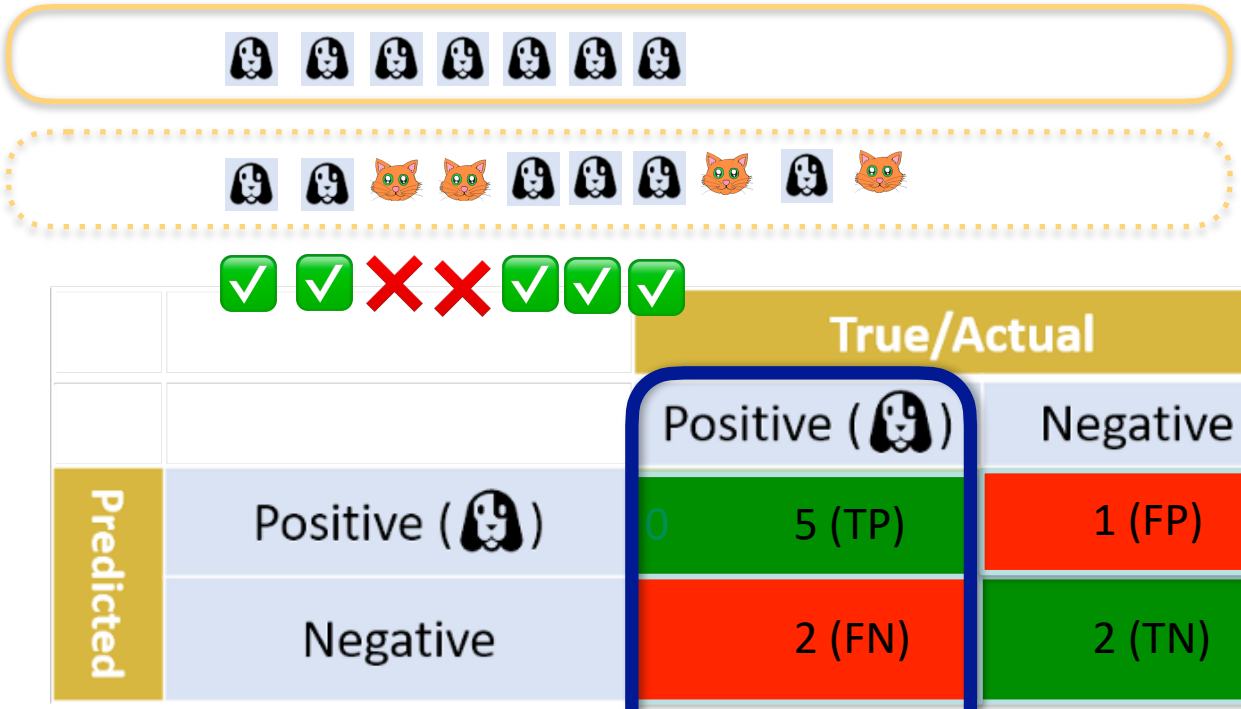
		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

How many of positive instances could I find?

$$R = \frac{TP}{TP + FN} = \frac{5}{5+2} = \frac{5}{7} = 71.4\,\%$$

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Example: Recall



How many of positive instances could I find?

$$R = \frac{TP}{TP + FN} = \frac{5}{5+2} = \frac{5}{7} = 71.4\,\%$$

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Summary

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- What is ML?
  - Sample representation is the core of ML
  - Feature definition and selection for sample encoding
- Two primary types of ML algorithms?
  - Supervised learning
    - Classification
    - Regression
  - Unsupervised learning
    - Clustering
    - Dimensionality reduction
- Evaluating performance of ML algorithms
  - K-fold cross validation
  - Test set
  - Evaluation metrics

Readings

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Mandatory

- Russel & Norvig: *Chapter 18 Learning from Examples (**except the section about decision trees)***

Thank You