Hochschule Düsseldorf
University of Applied Sciences

Fachbereich Medien
Faculty of Media

Classification

with Decision Trees

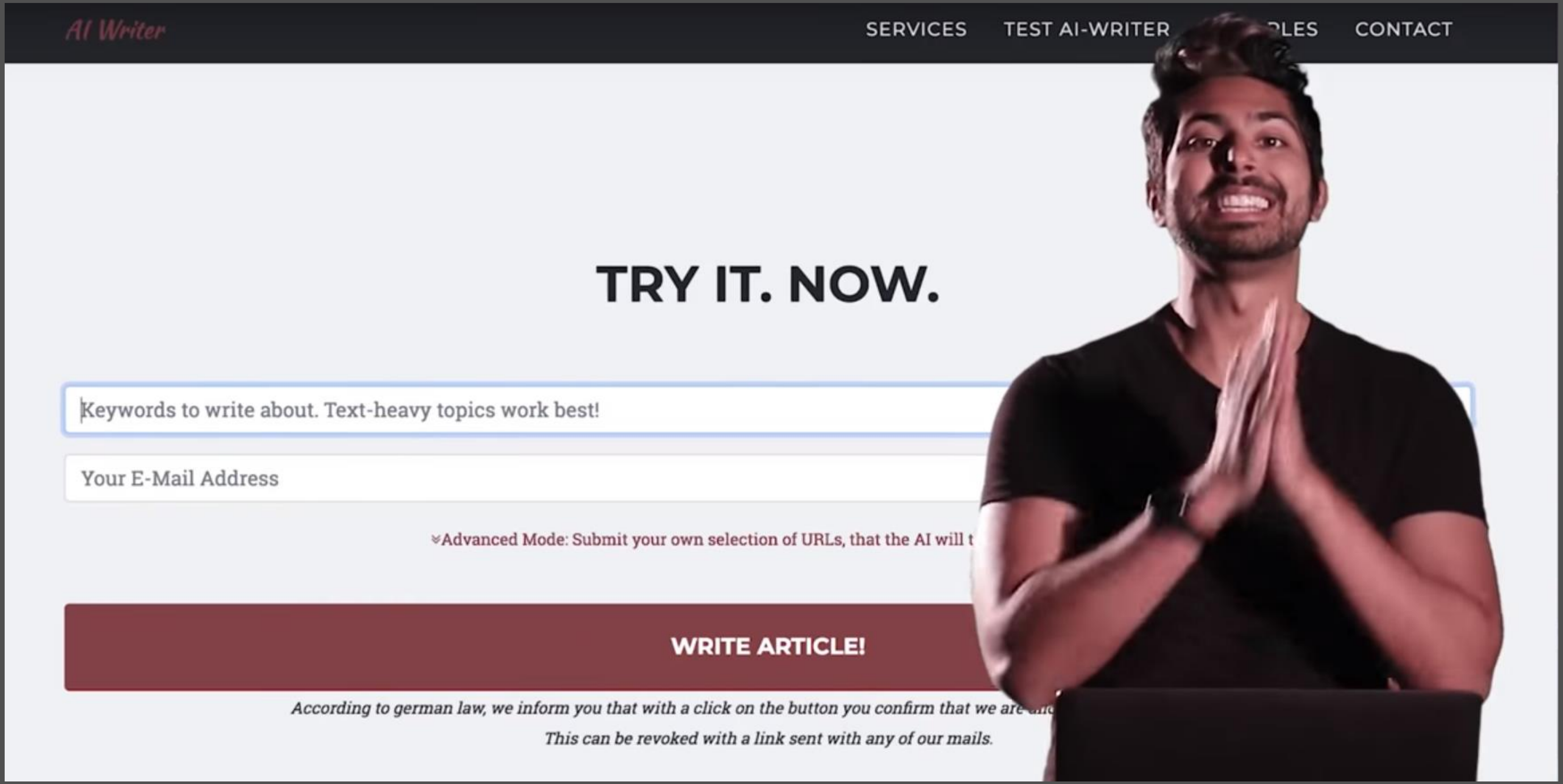
MODALG SoSe18

Facts & News

OpenAI Salaries (Source NYT):

- Ilya Sutskever: 1.9 MM $ (in 2016)
- Ian Goodfellow: 800 M $ (in 2016)
- Pieter Abbeel: 425 M $ (in 2016)

AI in Marketing



TRY IT. NOW.

Keywords to write about. Text-heavy topics work best!

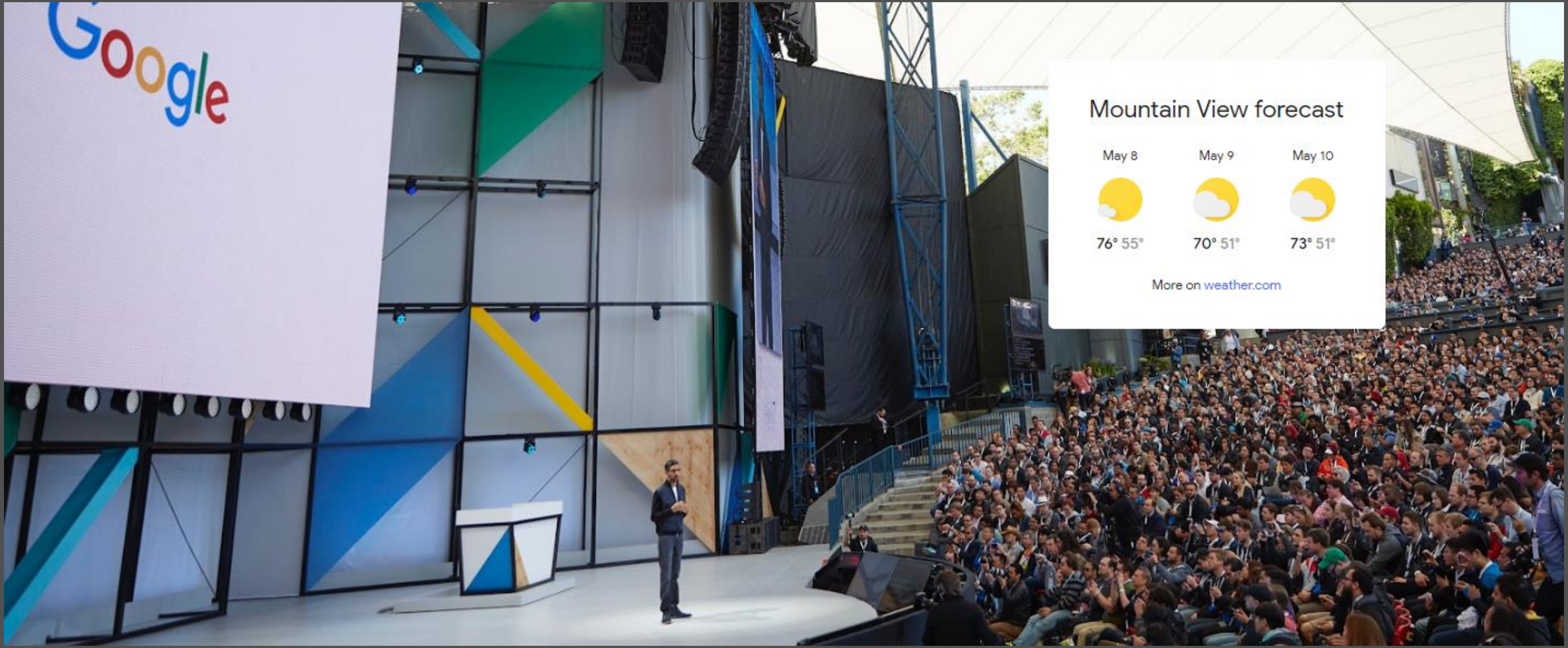
Your E-Mail Address

Advanced Mode: Submit your own selection of URLs, that the AI will t

According to german law, we inform you that with a click on the button you confirm that we are

This can be revoked with a link sent with any of our mails.

Start of Google I/O 2018



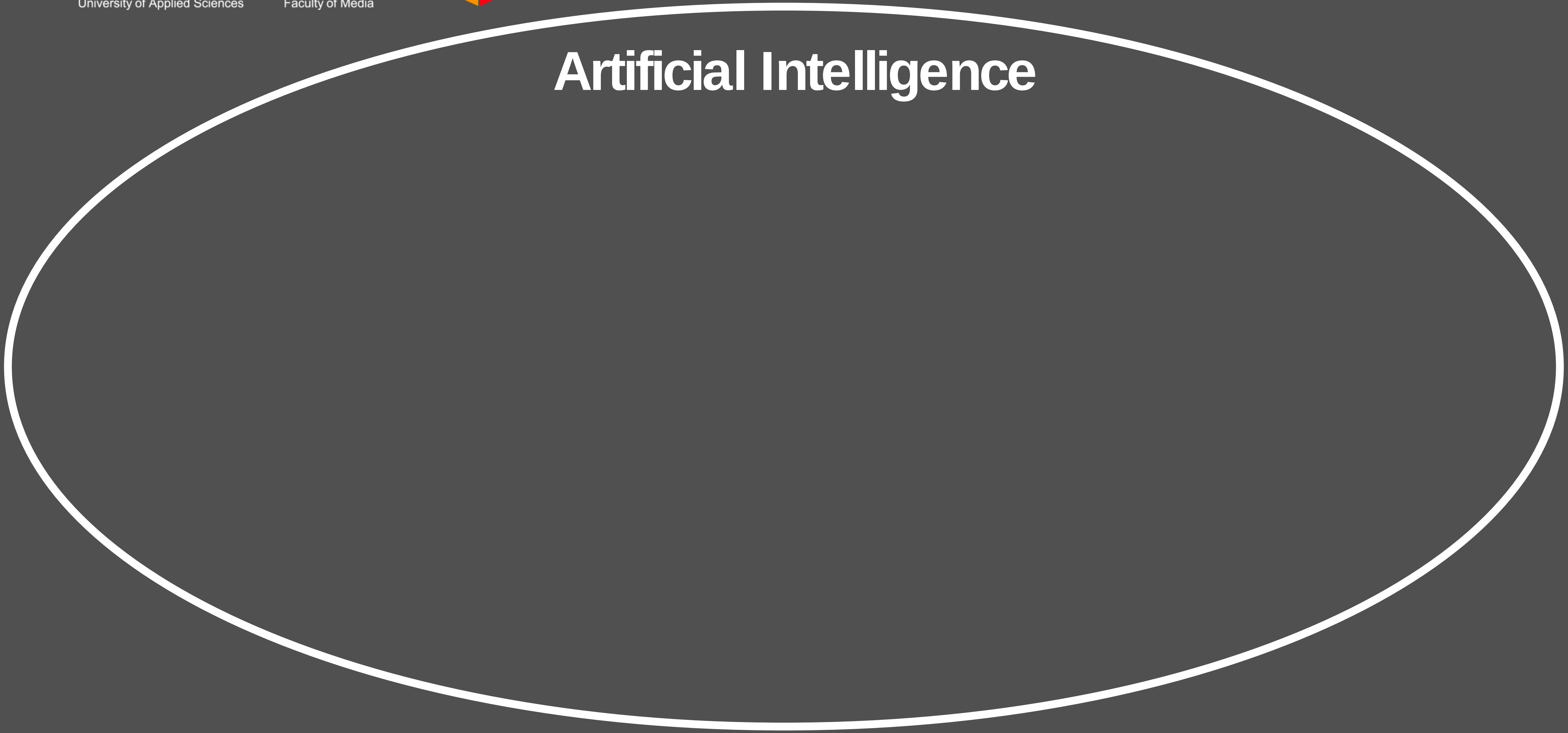
Mountain View forecast

May 8	May 9	May 10
76° 55°	70° 51°	73° 51°

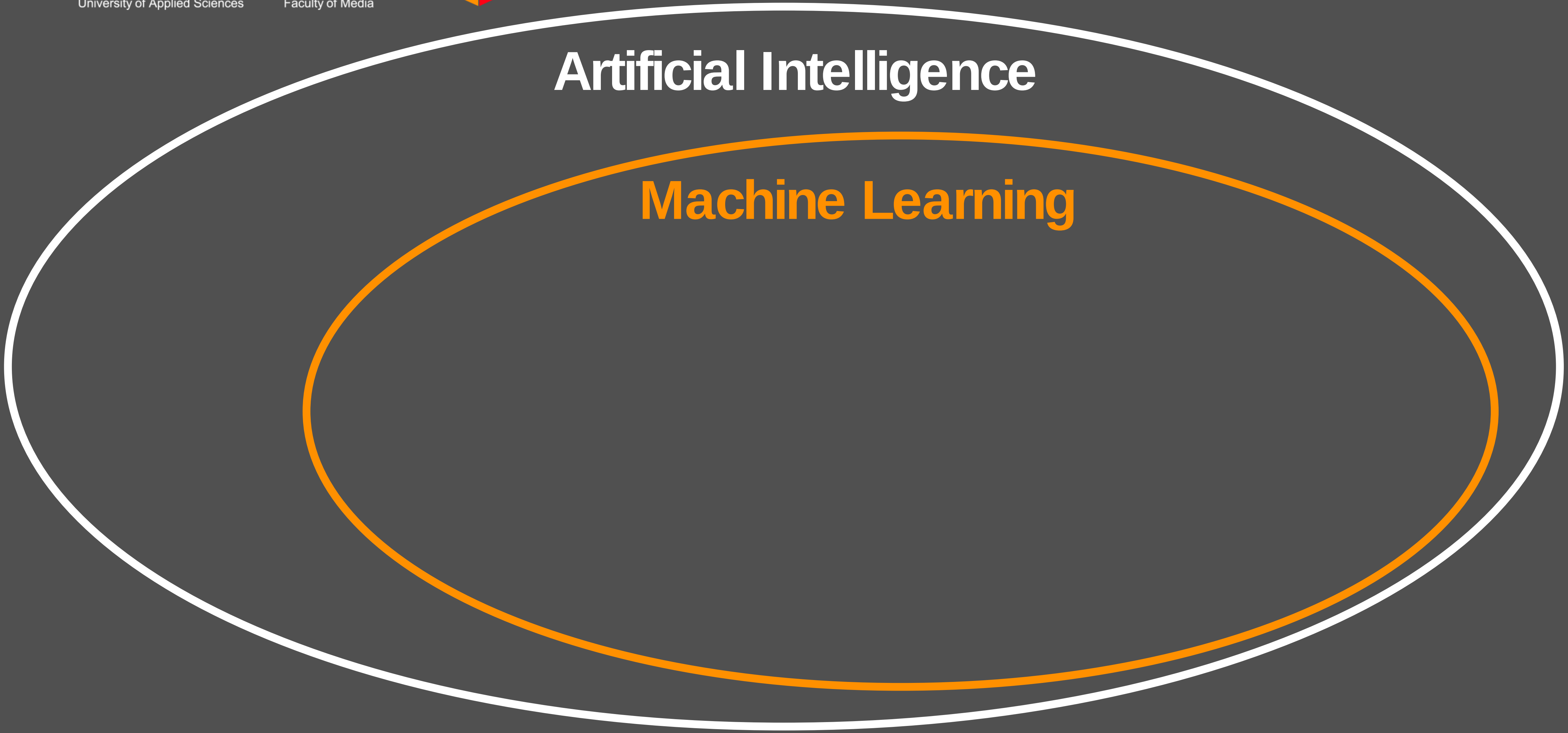
More on weather.com

kaggle™

Repitition

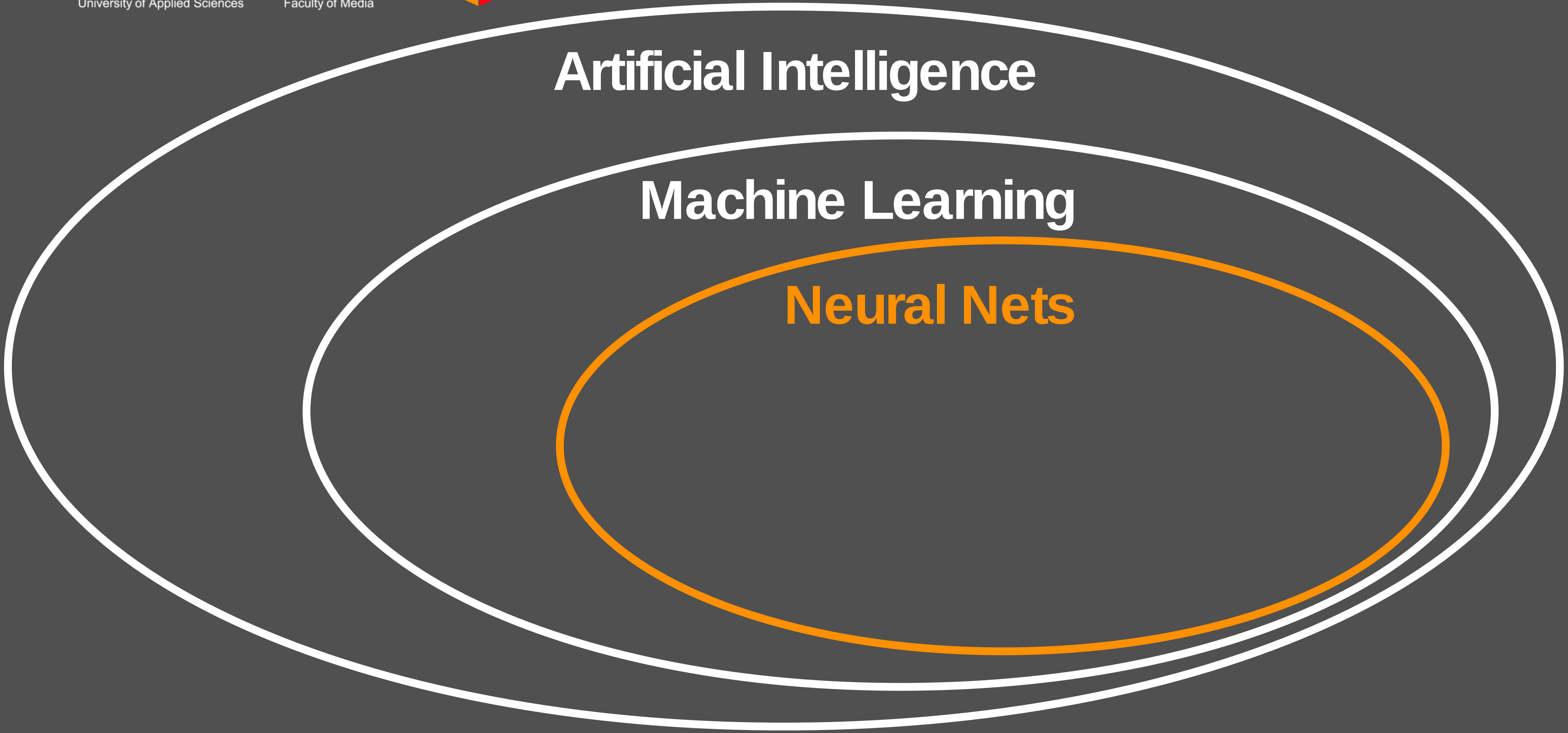


Artificial Intelligence



Artificial Intelligence

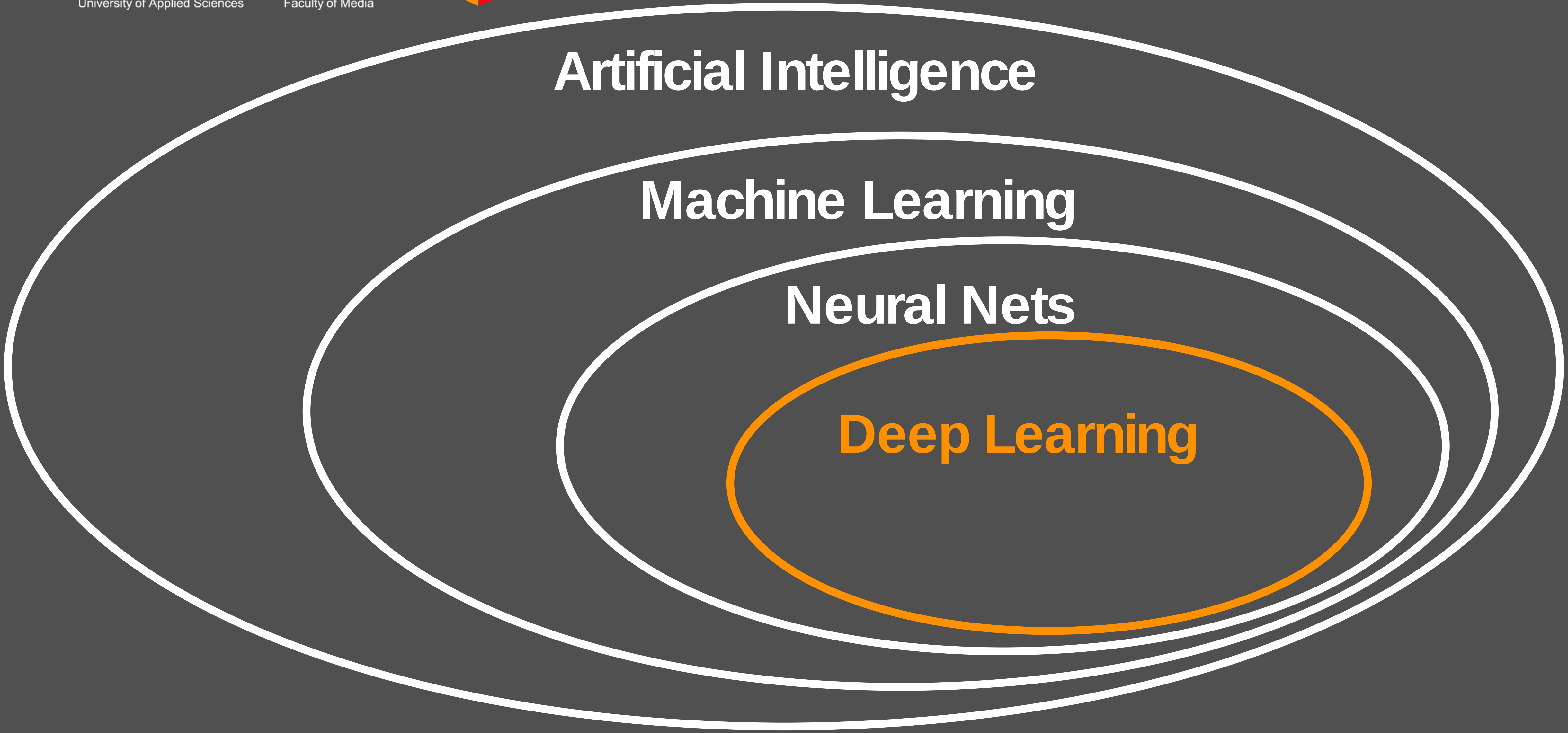
Machine Learning



Artificial Intelligence

Machine Learning

Neural Nets



Artificial Intelligence

Machine Learning

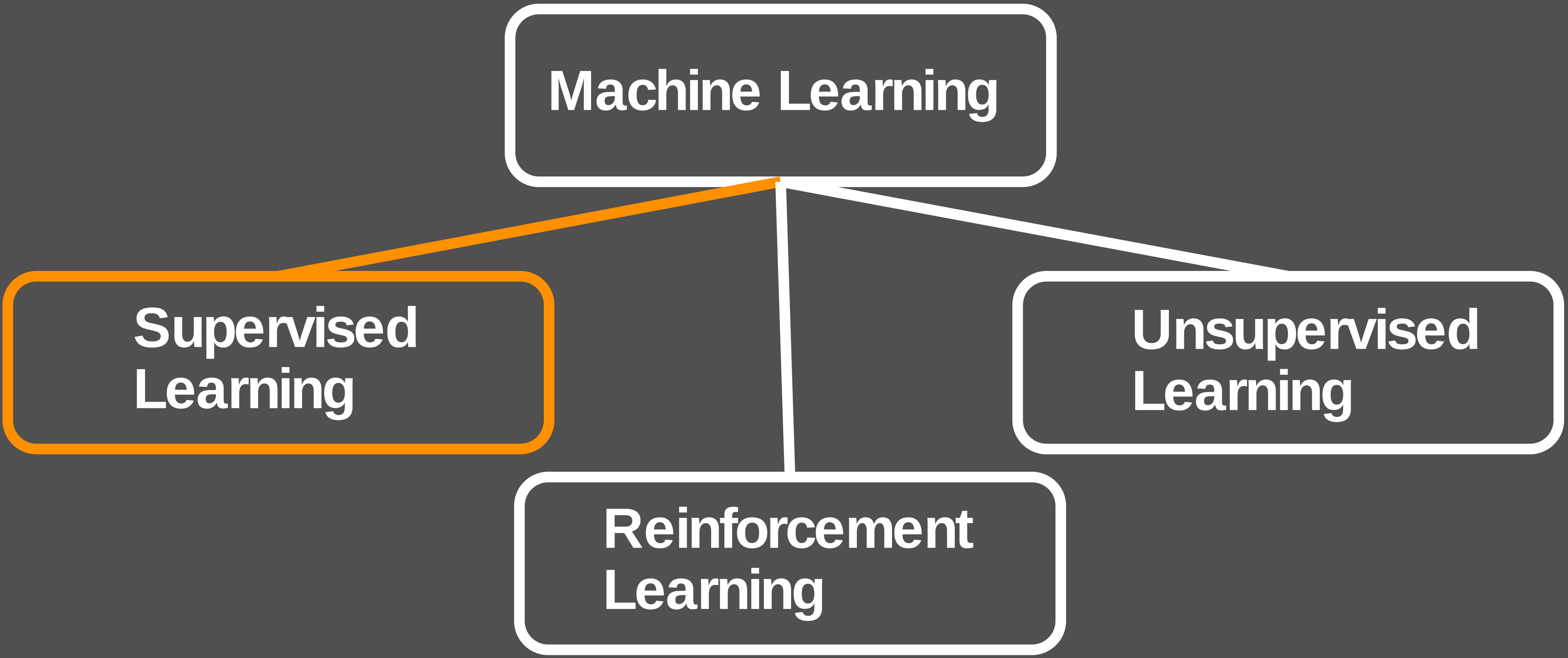
Neural Nets

Deep 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.” (Tom Mitchell)

Human Perspective:

- Task: Learn a text by heart!
- Performance: How many words are wrong?
- Experience: Learning Progress

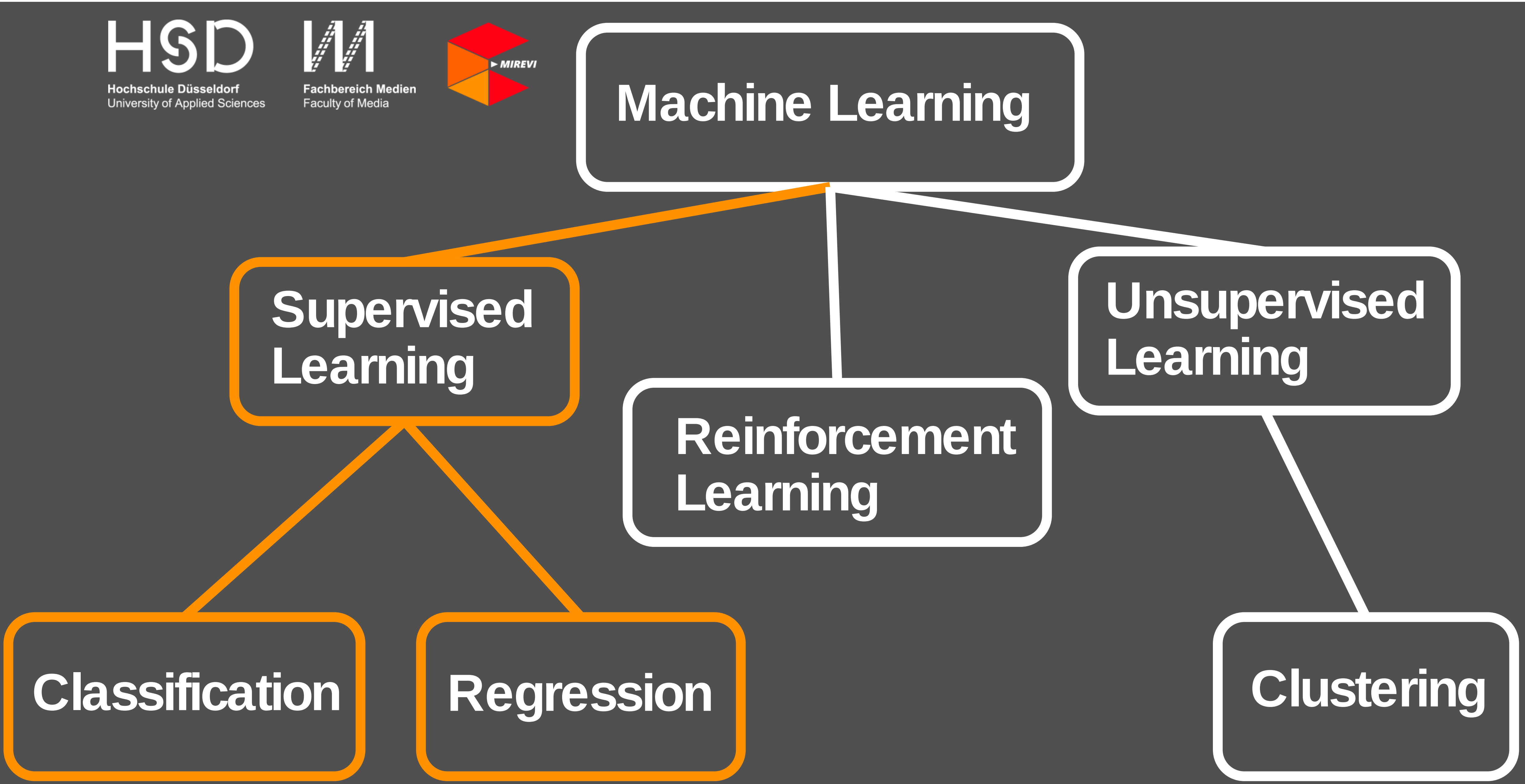


Machine Learning

Supervised Learning

Reinforcement Learning

Unsupervised Learning



Machine Learning

Supervised Learning

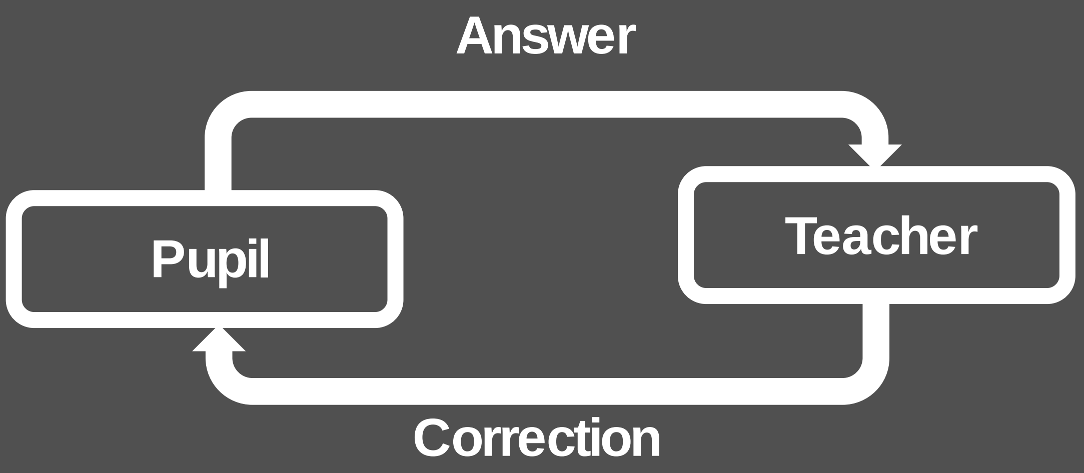
Reinforcement Learning

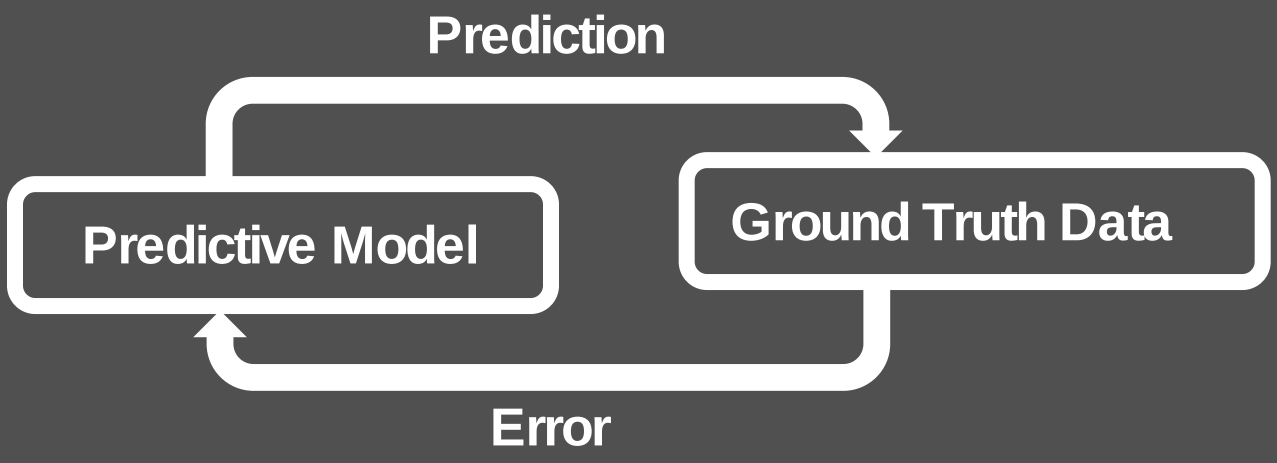
Unsupervised Learning

Classification

Regression

Clustering


Answer
Pupil
Teacher
Correction


Prediction
Predictive Model
Ground Truth Data
Error

Classification

5 Questions

What is the aim of classification?

What is a class?

Natural classification?

Artificial classification?

How we can do it?

What is the aim of classification?

- Differentiation

What is a class?

- Result of grouping things
- Compare with different types of personalities
- Data with the specific label

Natural classification?

- You do it daily!!!
- Learn something about the world
- Apply knowledge

Artificial classification?

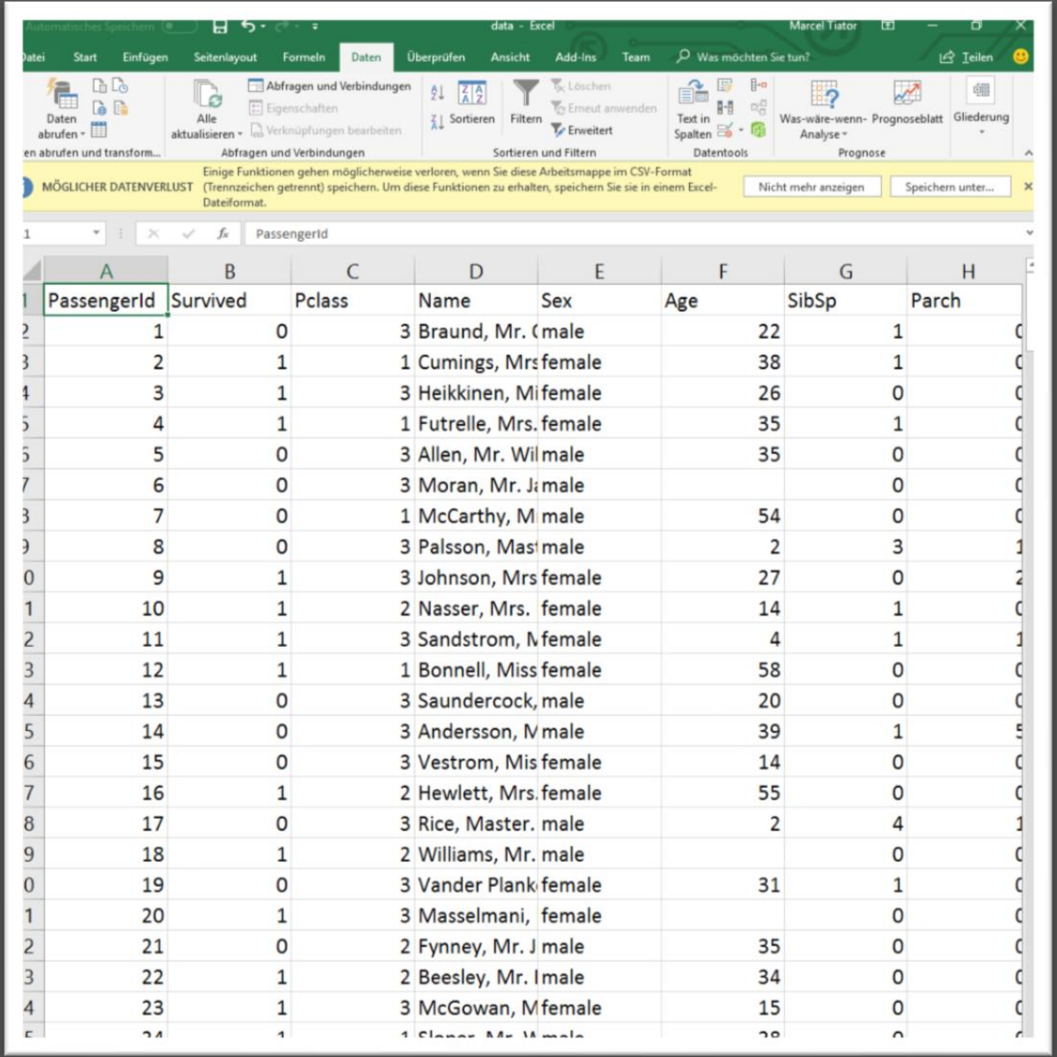
- Let an algorithm learn something about the world
- Apply knowledge

How we can do it?

- Data (knowledge about the world)
- Algorithm (something that can learn)

In Detail

Example Class Survived



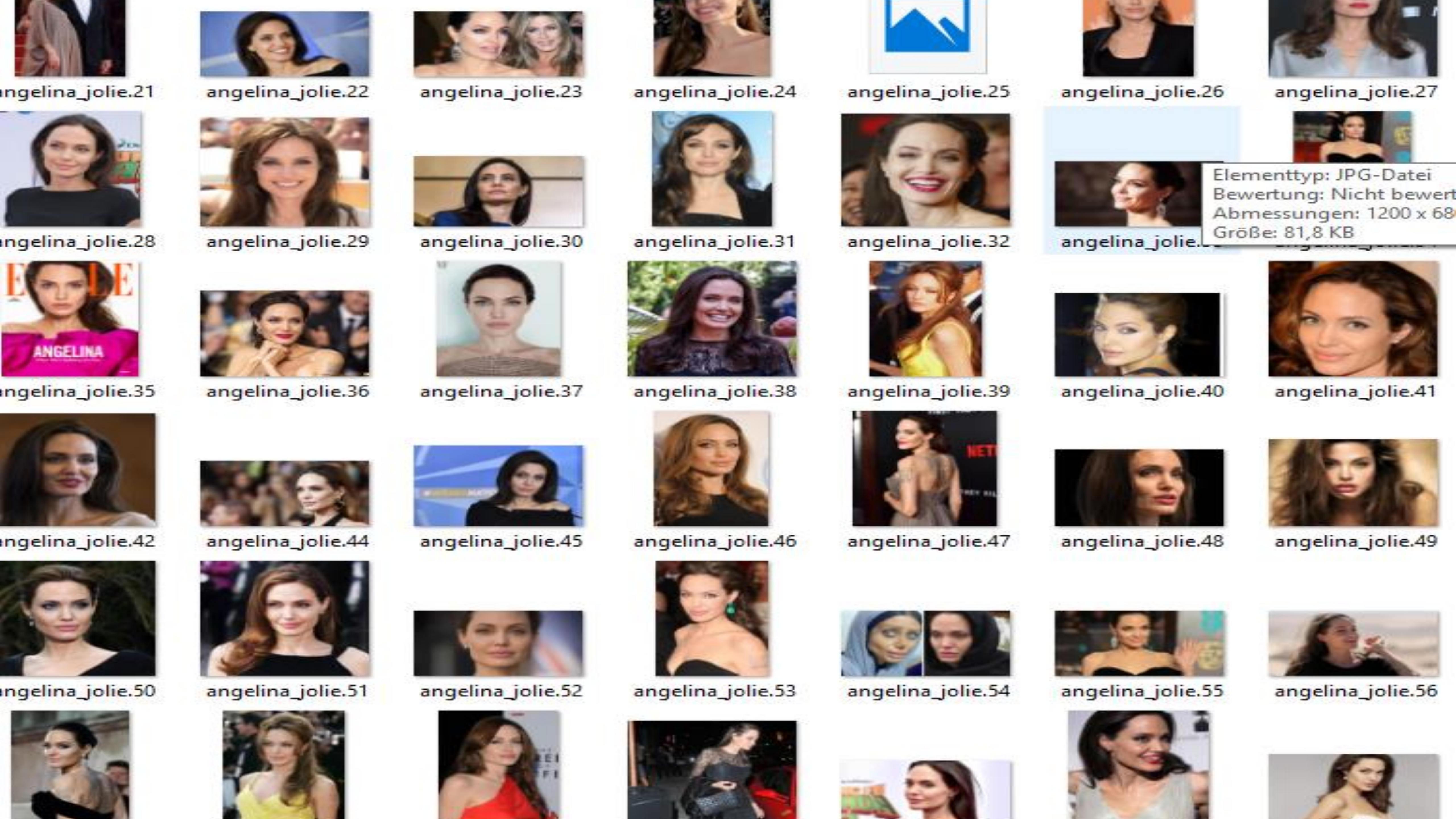
MÖGLICHER DATENVERLUST Einige Funktionen gehen möglicherweise verloren, wenn Sie diese Arbeitsmappe im CSV-Format (Trennzeichen getrennt) speichern. Um diese Funktionen zu erhalten, speichern Sie sie in einem Excel-Dateiformat.

1 | fx PassengerId

	A	B	C	D	E	F	G	H
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
2	1	0	3	Braund, Mr. (	male	22	1	0
3	2	1	1	Cumings, Mrs	female	38	1	0
4	3	1	3	Heikkinen, Mi	female	26	0	0
5	4	1	1	Futrelle, Mrs.	female	35	1	0
6	5	0	3	Allen, Mr. Wil	male	35	0	0
7	6	0	3	Moran, Mr. Ja	male		0	0
8	7	0	1	McCarthy, M	male	54	0	0
9	8	0	3	Palsson, Mast	male	2	3	1
10	9	1	3	Johnson, Mrs	female	27	0	2
11	10	1	2	Nasser, Mrs.	female	14	1	0
12	11	1	3	Sandstrom, M	female	4	1	1
13	12	1	1	Bonnell, Miss	female	58	0	0
14	13	0	3	Saundercock,	male	20	0	0
15	14	0	3	Andersson, M	male	39	1	5
16	15	0	3	Vestrom, Mis	female	14	0	0
17	16	1	2	Hewlett, Mrs.	female	55	0	0
18	17	0	3	Rice, Master.	male	2	4	1
19	18	1	2	Williams, Mr.	male		0	0
20	19	0	3	Vander Plank	female	31	1	0
21	20	1	3	Masselmani,	female		0	0
22	21	0	2	Fynney, Mr. J	male	35	0	0
23	22	1	2	Beesley, Mr. I	male	34	0	0
24	23	1	3	McGowan, M	female	15	0	0
25	24	1	1	Sloper, Mr. W	male	28	0	0

Can we predict if someone would survive?

- Look at the features!
  - Male or female
  - Number of class
  - Age of the person



ngelina_jolie.21 angelina_jolie.22 angelina_jolie.23 angelina_jolie.24 angelina_jolie.25 angelina_jolie.26 angelina_jolie.27

ngelina_jolie.28 angelina_jolie.29 angelina_jolie.30 angelina_jolie.31 angelina_jolie.32 angelina_jolie.

Elementtyp: JPG-Datei
Bewertung: Nicht bewert
Abmessungen: 1200 x 68
Größe: 81,8 KB

ngelina_jolie.35 angelina_jolie.36 angelina_jolie.37 angelina_jolie.38 angelina_jolie.39 angelina_jolie.40 angelina_jolie.41

ngelina_jolie.42 angelina_jolie.44 angelina_jolie.45 angelina_jolie.46 angelina_jolie.47 angelina_jolie.48 angelina_jolie.49

ngelina_jolie.50 angelina_jolie.51 angelina_jolie.52 angelina_jolie.53 angelina_jolie.54 angelina_jolie.55 angelina_jolie.56

Algorithm

Decision Tree

Algorithm

Decision Tree

Algorithm

Random Forest

Decision Tree

Gradient Boosting

Algorithm

Random Forest

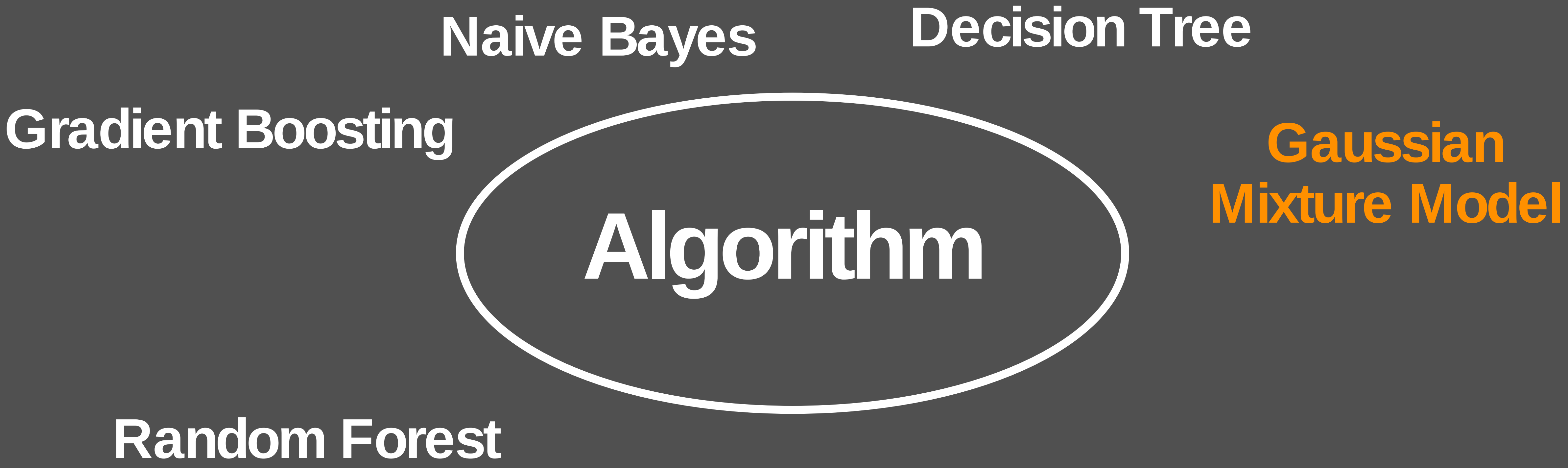
Naive Bayes

Decision Tree

Gradient Boosting

Algorithm

Random Forest



Naive Bayes

Decision Tree

Gradient Boosting

Algorithm

Gaussian Mixture Model

Random Forest

Naive Bayes

Decision Tree

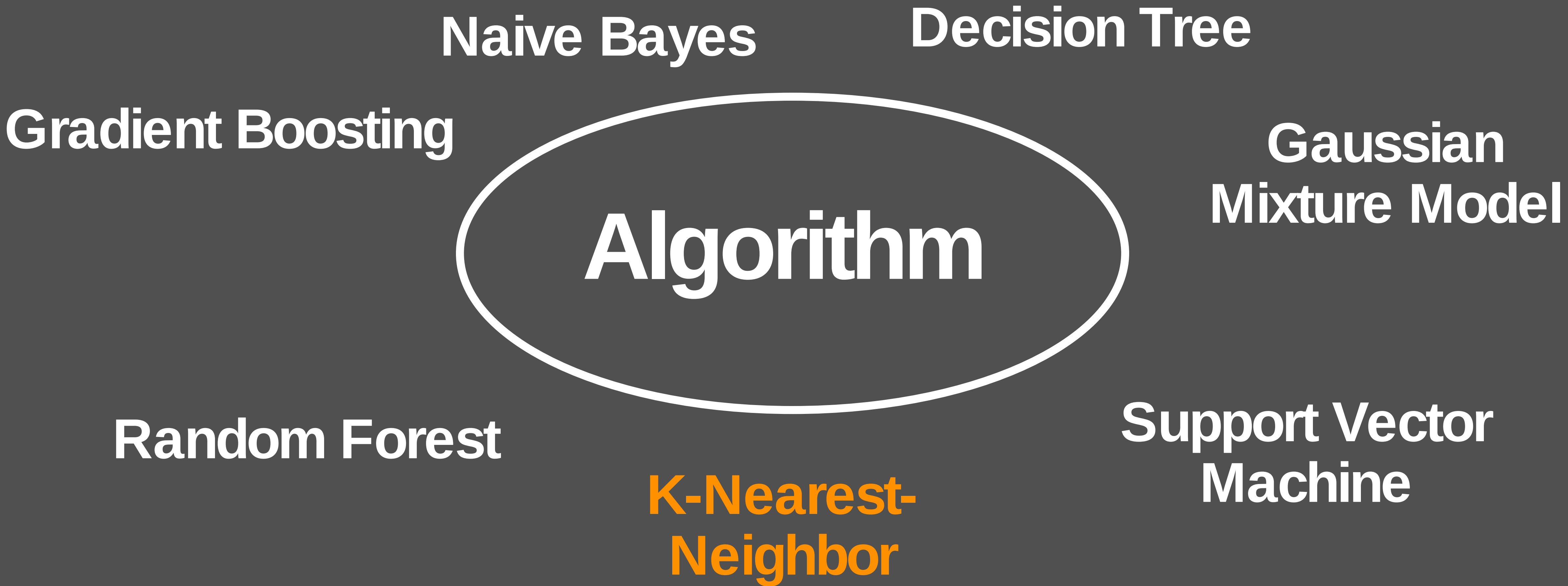
Gradient Boosting

Algorithm

Gaussian Mixture Model

Random Forest

Support Vector Machine



Naive Bayes

Decision Tree

Gradient Boosting

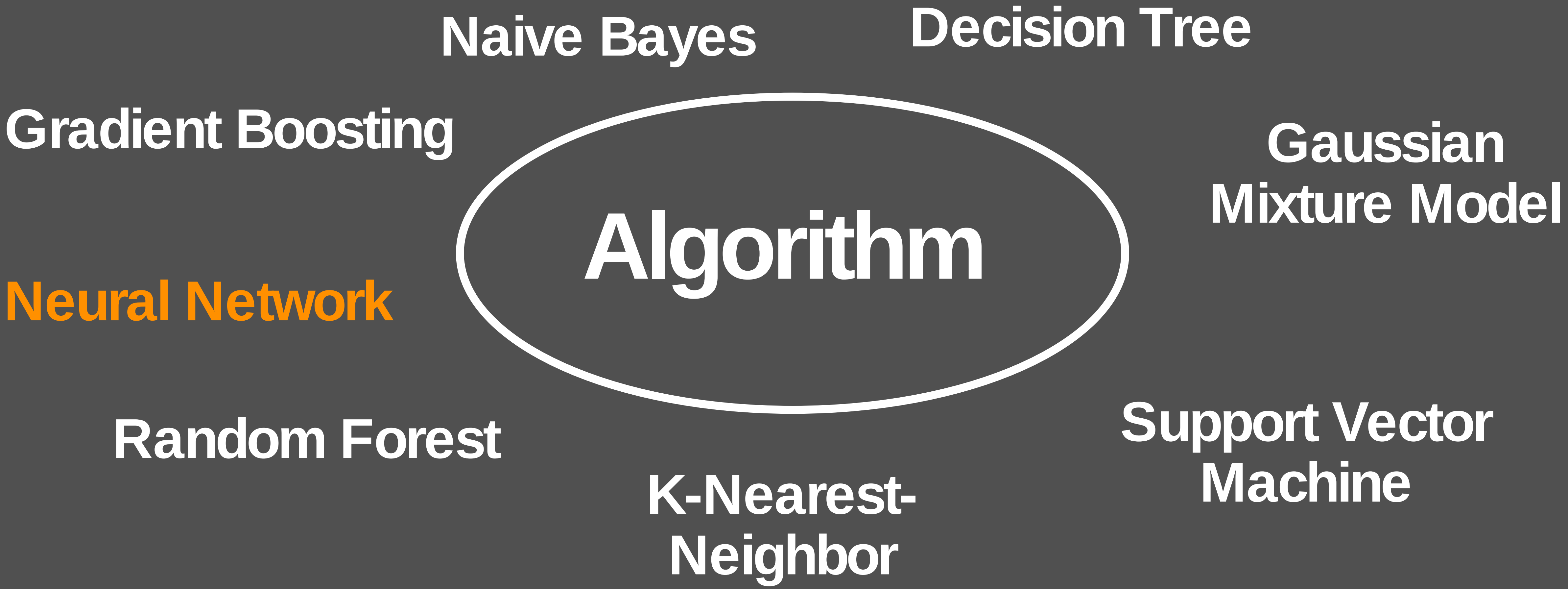
Algorithm

Gaussian Mixture Model

Random Forest

K-Nearest-Neighbor

Support Vector Machine



Naive Bayes

Decision Tree

Gradient Boosting

Gaussian Mixture Model

Algorithm

Neural Network

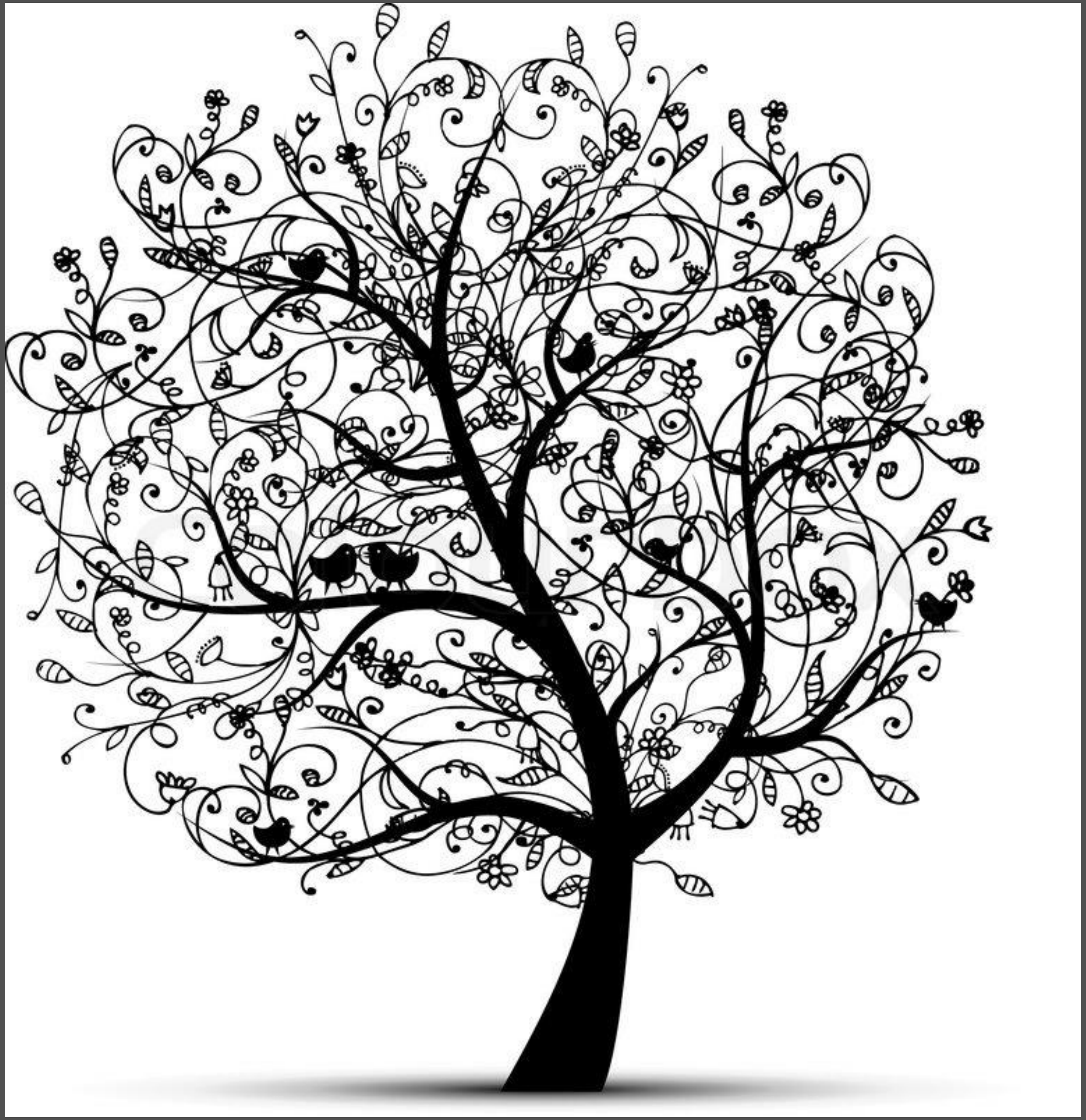
Random Forest

Support Vector Machine

K-Nearest-Neighbor

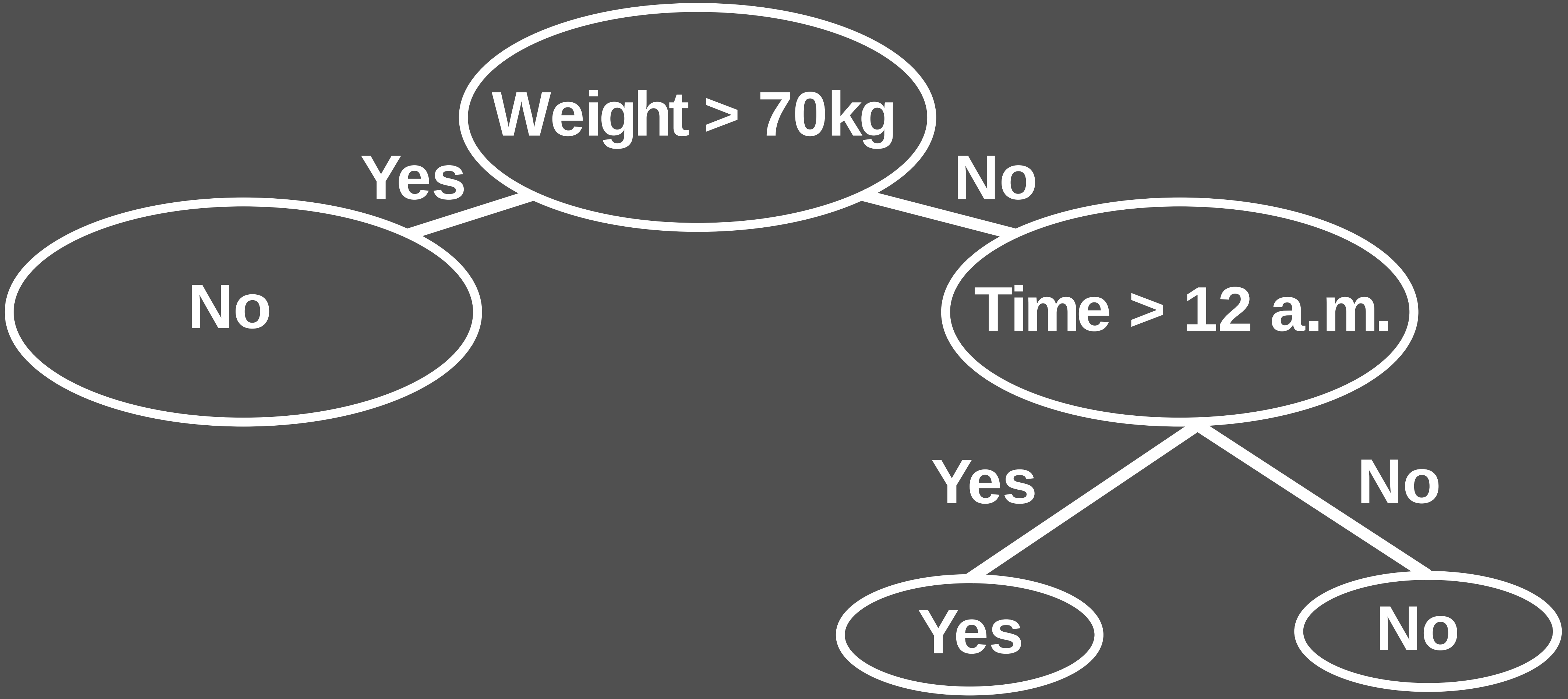
Algorithm = Model

Decision Tree

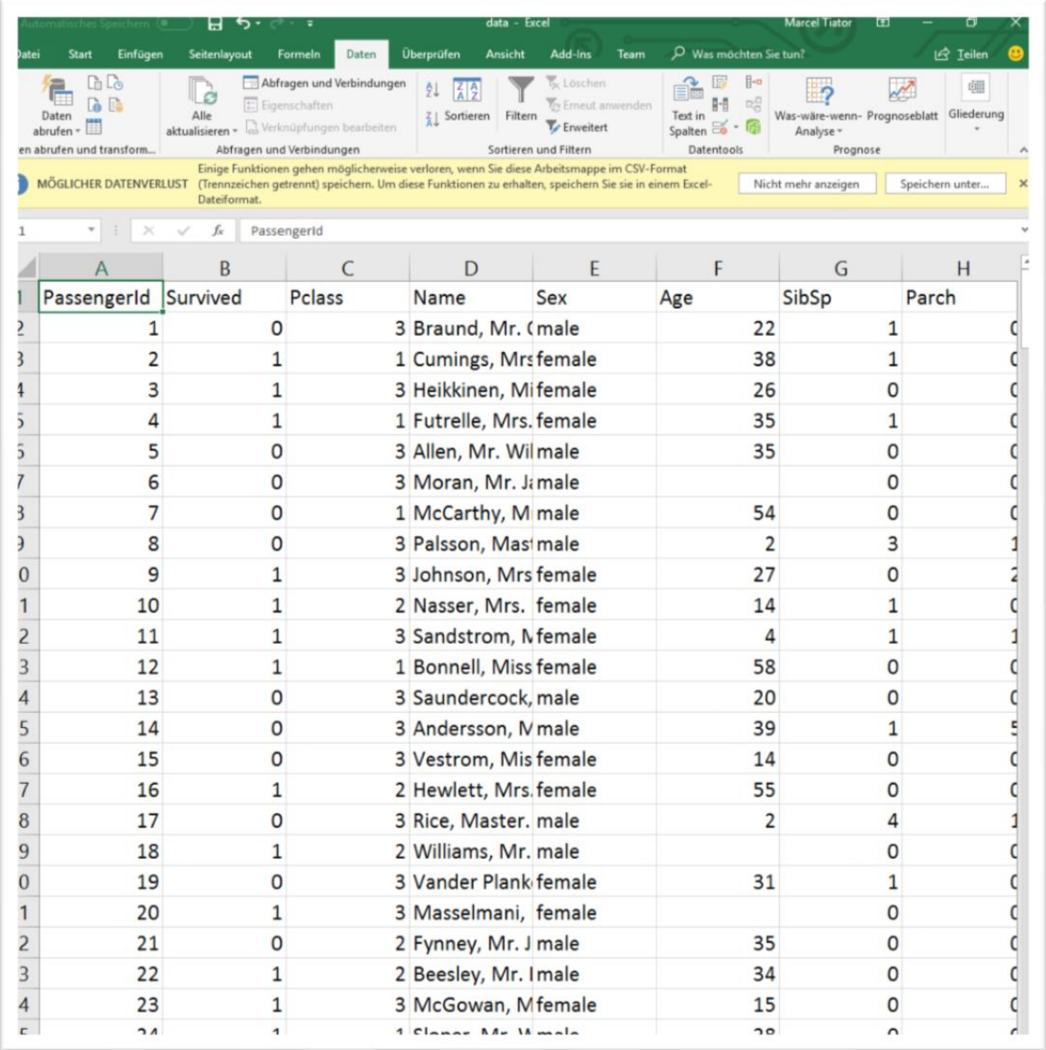


- Should I eat sweets?
- Possible answers: Yes/No
- Things to consider:
  - Weight
  - Condition of teeth
  - Diabetes
  - Time
  - Situation

Eat sweets?



Questions and Splits



MÖGLICHER DATENVERLUST Einige Funktionen gehen möglicherweise verloren, wenn Sie diese Arbeitsmappe im CSV-Format (Trennzeichen getrennt) speichern. Um diese Funktionen zu erhalten, speichern Sie sie in einem Excel-Dateiformat.

PassengerId

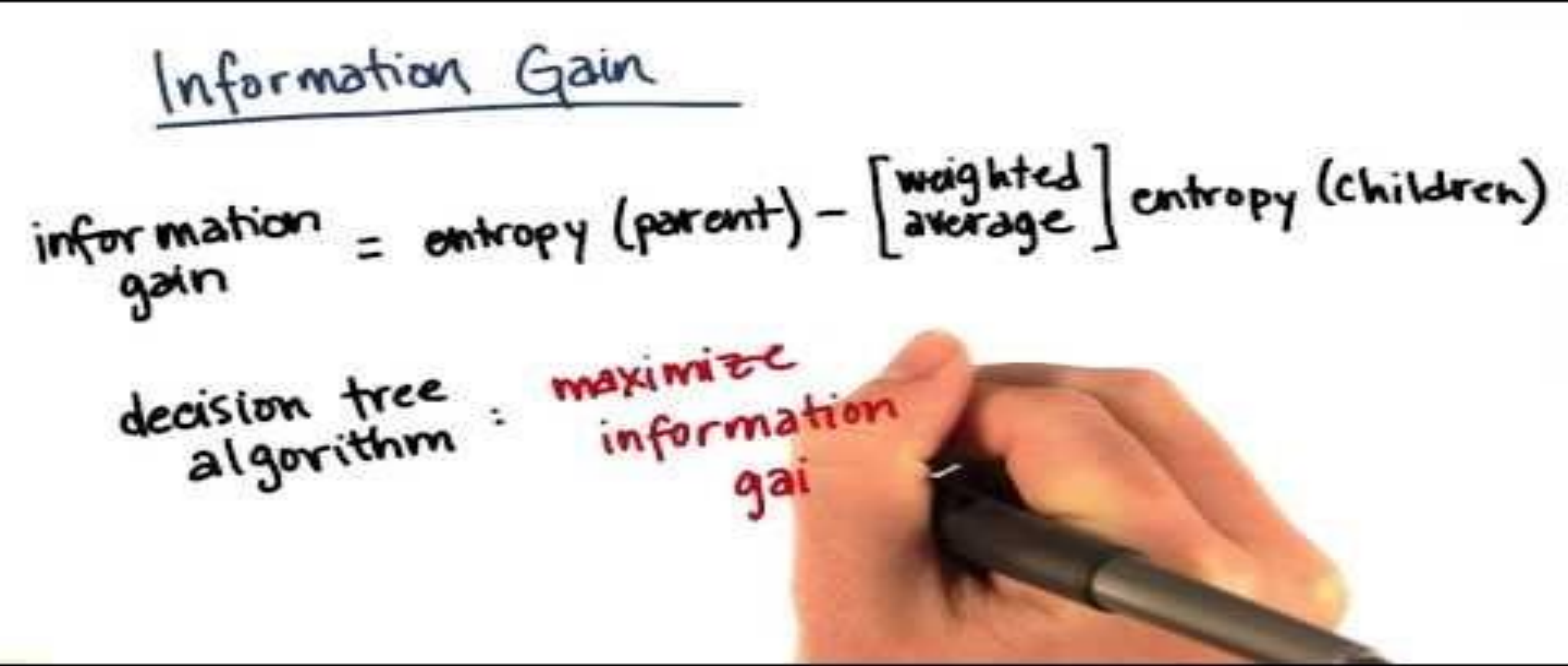
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
1	0	3	Braund, Mr. (	male	22	1	0
2	1	1	Cumings, Mrs	female	38	1	0
3	1	3	Heikkinen, Mi	female	26	0	0
4	1	1	Futrelle, Mrs.	female	35	1	0
5	0	3	Allen, Mr. Wil	male	35	0	0
6	0	3	Moran, Mr. J	male		0	0
7	0	1	McCarthy, M	male	54	0	0
8	0	3	Palsson, Mas	male	2	3	1
9	1	3	Johnson, Mrs	female	27	0	2
10	1	2	Nasser, Mrs.	female	14	1	0
11	1	3	Sandstrom, N	female	4	1	1
12	1	1	Bonnell, Miss	female	58	0	0
13	0	3	Saundercock,	male	20	0	0
14	0	3	Andersson, M	male	39	1	5
15	0	3	Vestrom, Mis	female	14	0	0
16	1	2	Hewlett, Mrs.	female	55	0	0
17	0	3	Rice, Master.	male	2	4	1
18	1	2	Williams, Mr.	male		0	0
19	0	3	Vander Plank	female	31	1	0
20	1	3	Masselmani,	female		0	0
21	0	2	Fynney, Mr. J	male	35	0	0
22	1	2	Beesley, Mr. I	male	34	0	0
23	1	3	McGowan, M	female	15	0	0
24	1	1	Sloper, Mr. W	male	28	0	0

- Given a dataset
- Assume we have some questions concerning the data
- How to decide where to split?

- Find the feature that best splits the target class into the purest possible children nodes
- If target class consist male and female
- Nodes that don't contain a mix of both male and female, rather pure nodes with only one class

Information Gain

- Information Gain is a measurement
- How much information do we gain by doing a split at particular feature?
- Compare to measure of quality when you do a split at specific question



Information Gain

information gain = entropy (parent) − [weighted average] entropy (children)

decision tree algorithm : maximize information gai

What is entropy?

- **Another measurement**
- **Measures impurity of data**

$$H(x) = -\sum_{i=1}^{n} p(x_i) * log_b(p(x_i))$$

H: **Entropy**

x: **Whole column of target feature (e.g. Eat sweets?)**

$x_i \in \{0,1\}$

e.g. $x = (0,1,0,0,1,0,1,0,1)^T$

$$H(x) = -\sum_{i=1}^{n} p(x_i) * log_b(p(x_i))$$

$p(x_i)$: How many times is $x_i = 0$? (relative value)

$p(x_i)$: How many times is $x_i = 1$? (relative value)

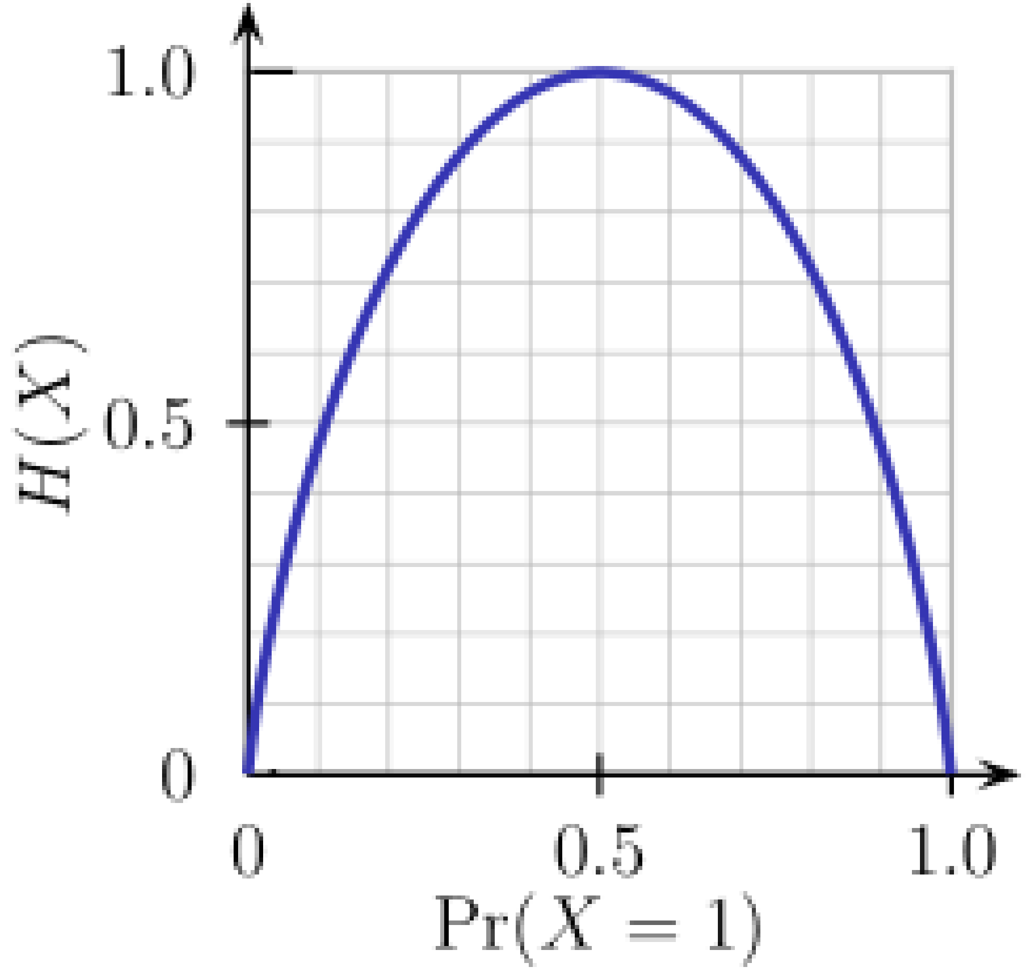
x_i: Answer of eating sweets

$$H(x) = -\sum_{i=1}^{n} p(x_i) * log_b(p(x_i))$$

b: number of classes (binary: $b = 2$)

n: length of vector x

$$H(x) = -\sum_{i=1}^{n} p(x_i) * log_b(p(x_i))$$


1.0
0.5
0
$H(X)$
0
0.5
1.0
$\Pr(X = 1)$

Color	Diam	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

What else?

- Recursive algorithm (Iterative Dichotomiser - ID3)
- Assume binary problem:

```
build_tree(data, questions){
}
```

- Recursive algorithm (Iterative Dichotomiser - ID3)
- Assume binary problem:

```
build_tree(data, questions){
    foreach(question) gain, question = information_gain(data, questions)
}
```

- Recursive algorithm (Iterative Dichotomiser - ID3)
- Assume binary problem:

```
build_tree(data, questions){
        foreach(question) gain, question = information_gain(data, questions)
        take question with argmax gain
}
```

- Recursive algorithm (Iterative Dichotomiser - ID3)
- Assume binary problem:

```
build_tree(data, questions){
    foreach(question) gain, question = information_gain(data, questions)
    take question with argmax gain
    if(gain == 0):
      return leaf_node(data)
}
```

- Recursive algorithm (Iterative Dichotomiser - ID3)
- Assume binary problem:

```
build_tree(data, questions){
        foreach(question) gain, question = information_gain(data, questions)
        take question with argmax gain
        if(gain == 0):
           return leaf_node(data)
        true_data, false_data = split(data, question)
}
```

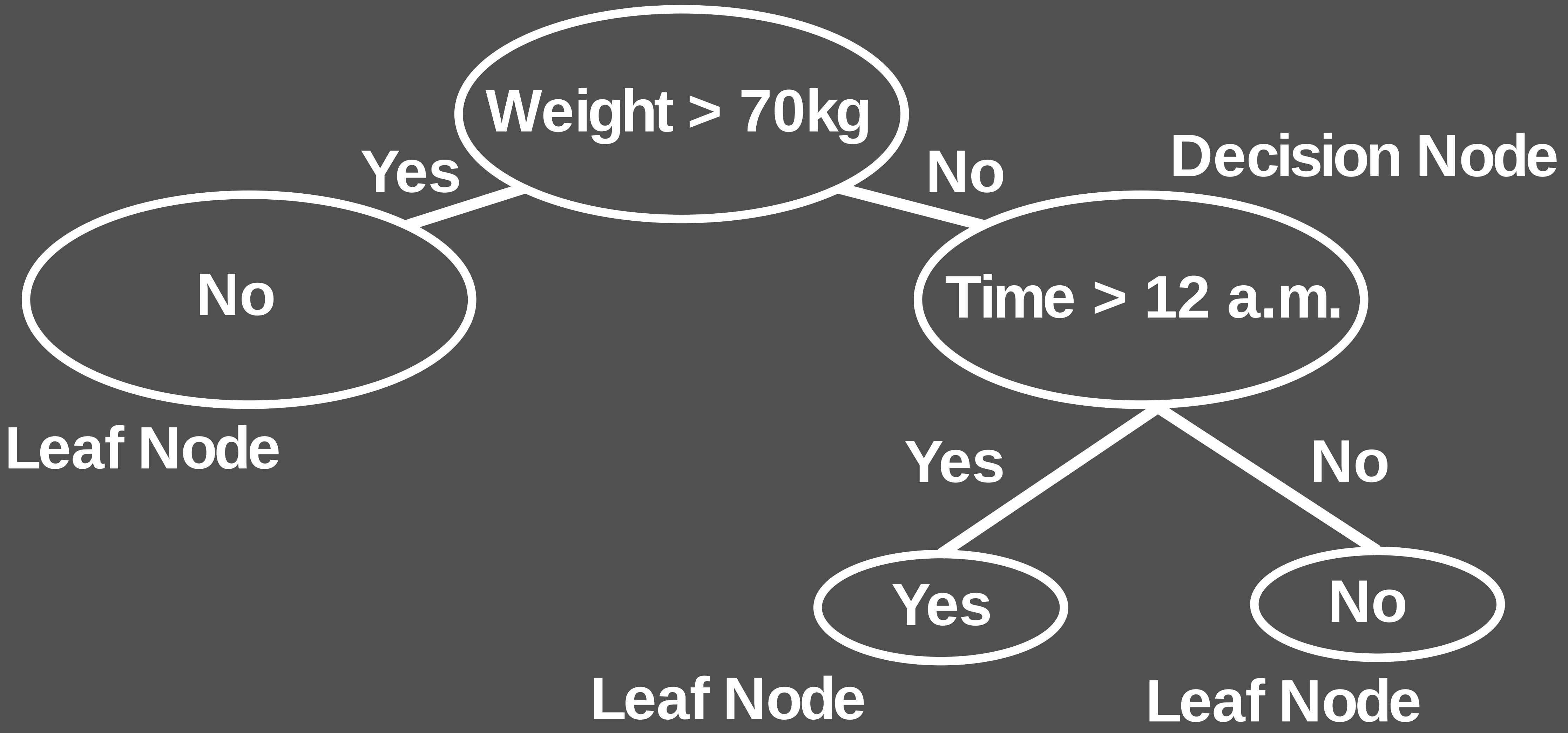
- Recursive algorithm (Iterative Dichotomiser - ID3)
- Assume binary problem:

```
build_tree(data, questions){
        foreach(question) gain, question = information_gain(data, questions)
        take question with argmax gain
        if(gain == 0):
           return leaf_node(data)
        true_data, false_data = split(data, question)
        build_tree(true_data, questions)
        build_tree(false_data, questions)
}
```

- Recursive algorithm (Iterative Dichotomiser - ID3)
- Assume binary problem:

```
build_tree(data, questions){
    foreach(question) gain, question = information_gain(data, questions)
    take question with argmax gain
    if(gain == 0):
       return leaf_node(data)
    true_data, false_data = split(data, question)
    build_tree(true_data, questions)
    build_tree(false_data, questions)
    return decision_node(question, data)
 }
```

Eat sweets?



Weight > 70kg

Yes

No

Decision Node

No

Time > 12 a.m.

Leaf Node

Yes

No

Yes

No

Leaf Node

Leaf Node

There is more!

- Multiclass-Problem (Eat sweets? Yes/No/Maybe)
- Pruning of the tree
- Cross Validation
- Bias
- Metrics – Important: Accuracy

Accuracy = #Correct Classified / #Examples

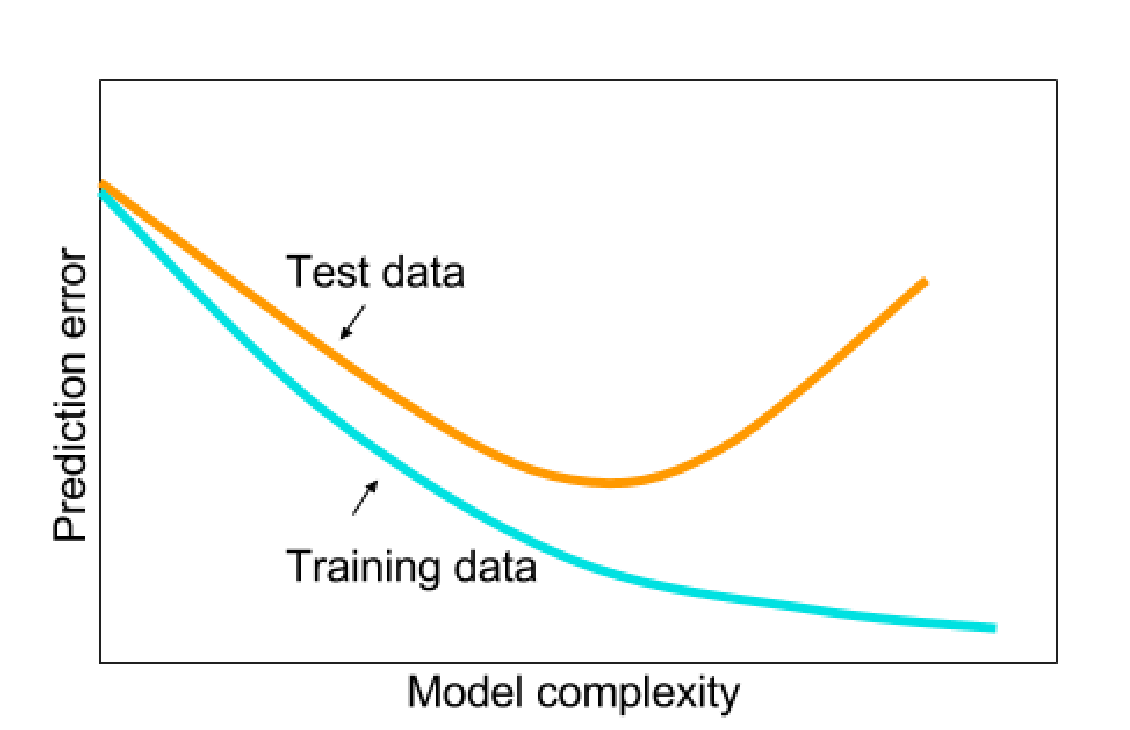
What is the aim of classification?

- Differentiation
- You want to be a master
- Generalise!

Problem of generalisation:

- The world is full of bias
- You may a master concerning the training set, but what happens if I give you a another test set?
- Accuracy can decrease!

Overfitting



Conclusion:

- You can compare classification to grouping
- We can use algorithms (models)
- A decision tree is a human readable model
- Measure the performance with classification accuracy

Your Turn!

github.com/mati3230/modalg181