WORKSHOP MACHINE LEARNING

Dania International Days

13 – 15 March 2024

Andy Louwyck

Vives University of applied sciences – Association KU Leuven – Kortrijk, Belgium

DANIA ACADEMY

vives University of Applied Sciences

Andy Louwyck

andy.louwyck@vives.be

- **Doctor & Master of Science: Geology**
- **Associate Degree: IT & Programming**
- **Micro Degree: AI & Data Science**

2024 - …	**Voluntary Post-Doctoral Researcher** (Ghent University)
2023 - …	**IT Architect** (Flanders Environment Agency)
2020 - …	**Lecturer AI** (Vives University of Applied Sciences)
2020 - 2022	**Research Associate AI** (Vives University of Applied Sciences)
2009 - 2020	**Groundwater Modeler & Data Expert** (Flanders Environment Agency)
2007 - 2008	**Project Engineer Water Management** (International Marine & Dredging Consultants)
2006	**Science Teacher** (HH Ninove Secondary School)
2001 - 2005	**PhD Fellow Hydrogeology** (Ghent University)

GHENT UNIVERSITY

FLANDERS ENVIRONMENT AGENCY

Flanders State of the Art

vives University of Applied Sciences

Dania International Days 2024

Workshop Machine Learning

MACHINE LEARNING: OVERVIEW

DANIA ACADEMY

University of Applied Sciences

Data

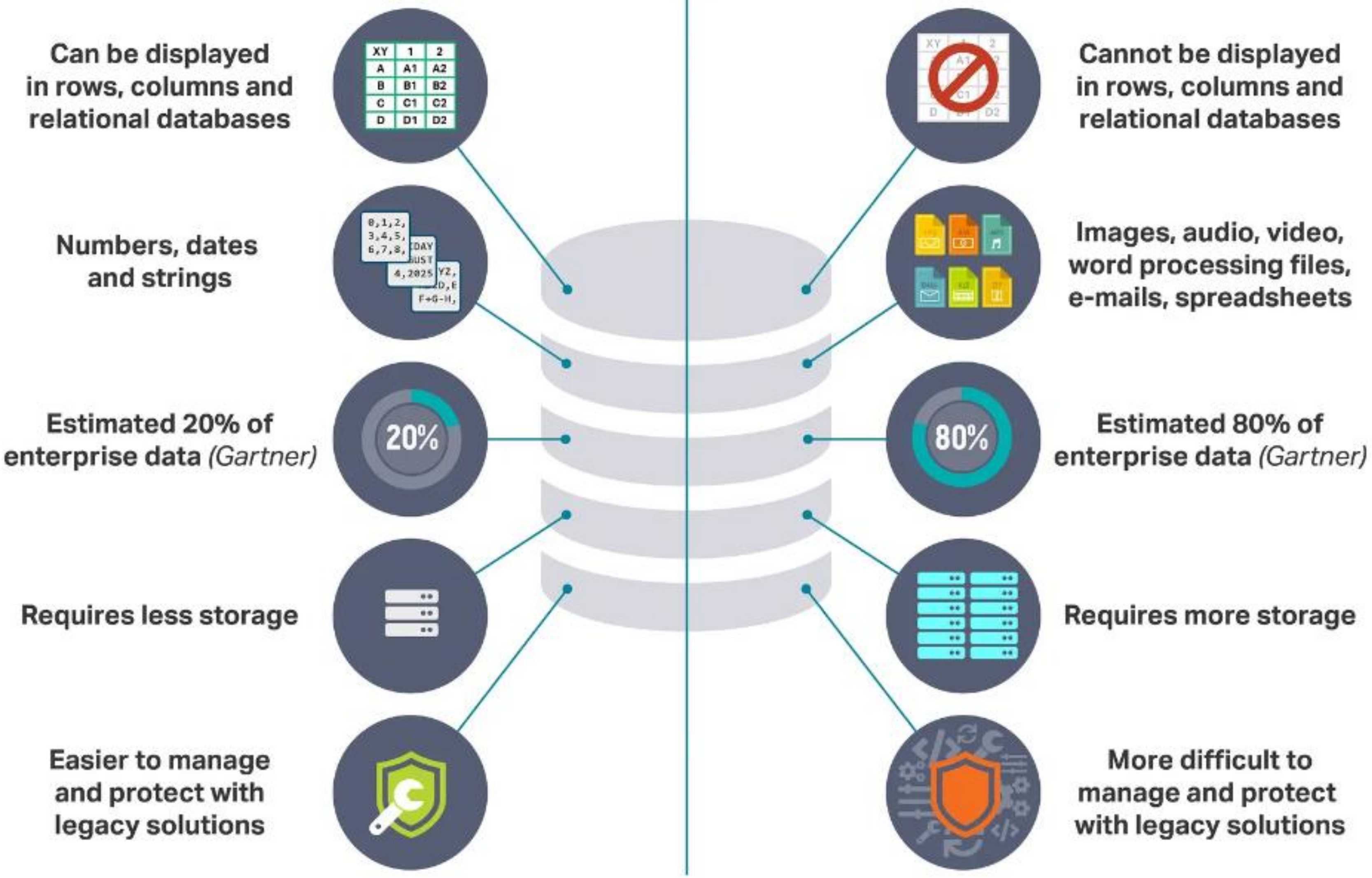
“We are drowning in data but starving for knowledge”

[Naisbitt, 1982]

- – A lot of data is gathered, but never used
- – It is easier to generate data than to analyze data

→ **MACHINE LEARNING**

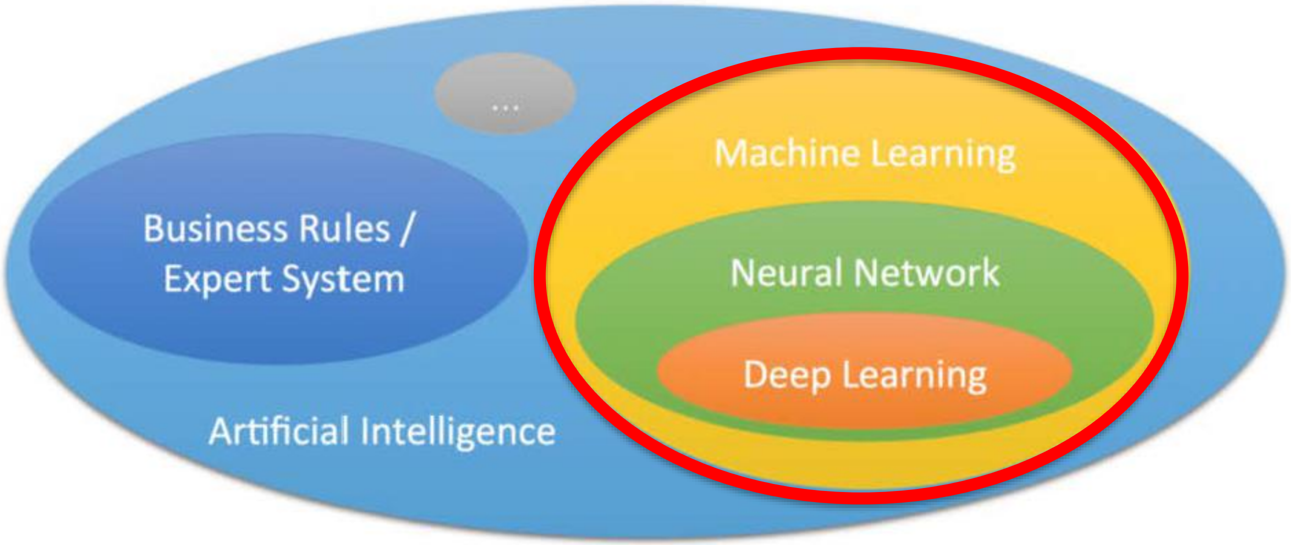
Structured Data VS Unstructured Data



https://www.igneous.io/blog/structured-data-vs-unstructured-data

Machine Learning & Artificial Intelligence

- **Artificial Intelligence** (AI):
  "The set of all tasks in which a computer can make decisions."
- **Machine Learning** (ML):
  "The set of all tasks in which a computer can make decisions *based on data*."
- **Deep Learning** (DL):
  "The field of machine learning that uses certain objects called *neural networks*."



Machine Learning

- Core domain of AI, concerned with automatic learning

intelligence

noun

UK /ɪnˈtel.ɪ.dʒ^{ə}ns/ US /ɪnˈtel.ə.dʒ^{ə}ns/

intelligence ***noun*** **(ABILITY)**

B2 [U]

the ability to learn, understand, and make judgments or have opinions that are based on reason:

- *an intelligence test*
- *a child of high/average/low intelligence*
- *It's the intelligence of her writing that impresses me.*

- A computer is said to be able to learn if its performance in solving some task improves with its experience

DANIA ACADEMY

vives University of Applied Sciences

Machine Learning

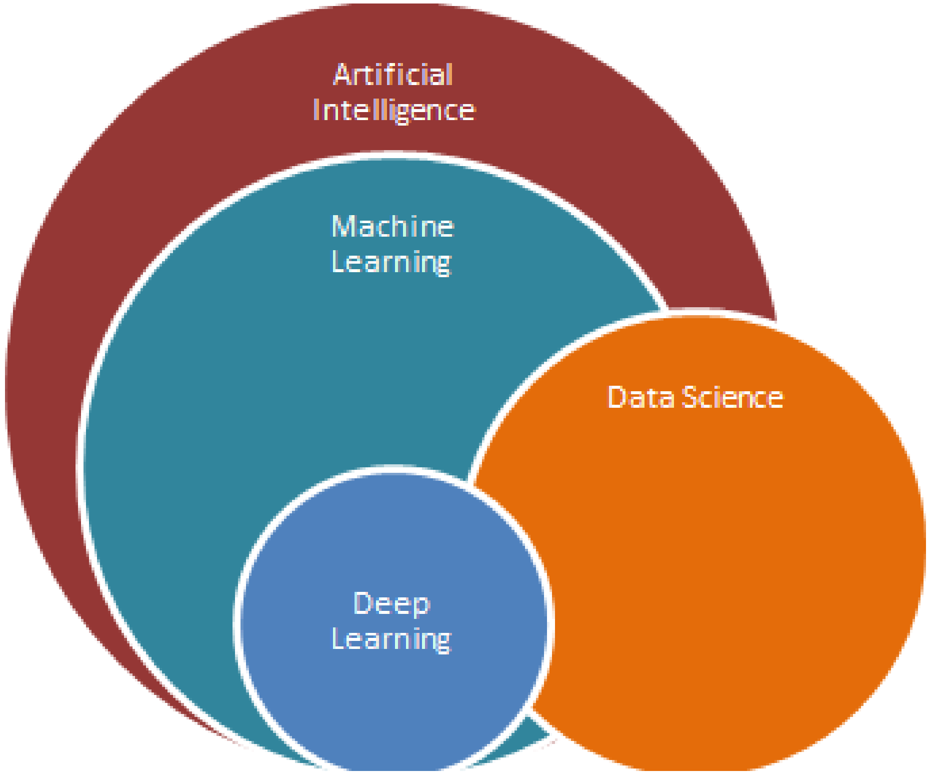
Example: buying a new car

- How do we make decisions?
  - by logical **reasoning**
  - by relying on previous **experiences** (either our own or those of others)

- For a computer: **experiences = data**

"Machine learning is common sense, except done by a computer"

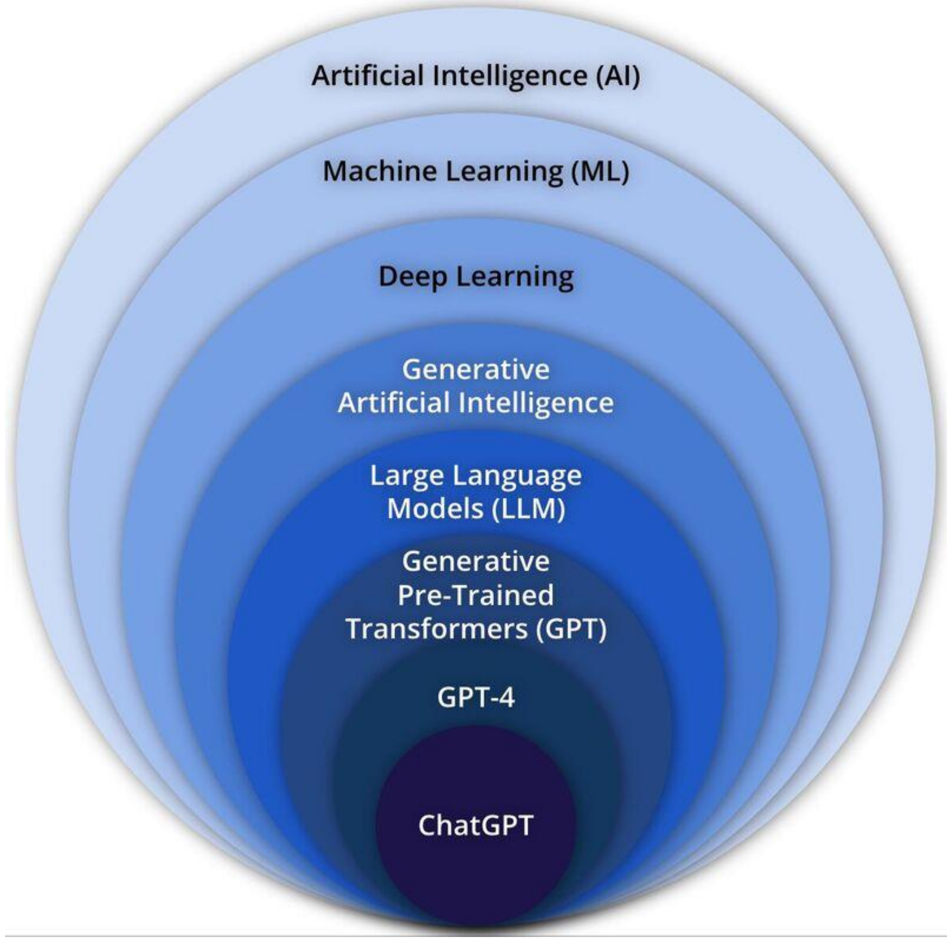
Machine Learning ≠ Data Science



In practice:

- ML team: delivers software
- DS team: provides new insights

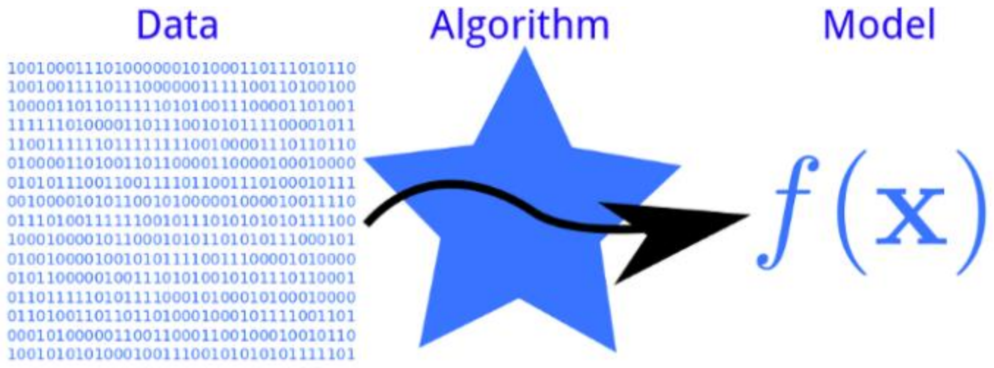
What about ChatGPT?



https://www.linkedin.com/posts/edoardoquercidellarovere_spesso-mi-capita-di-leggere-e-sentire-conversazioni-activity-7126095949738721280-amQI/

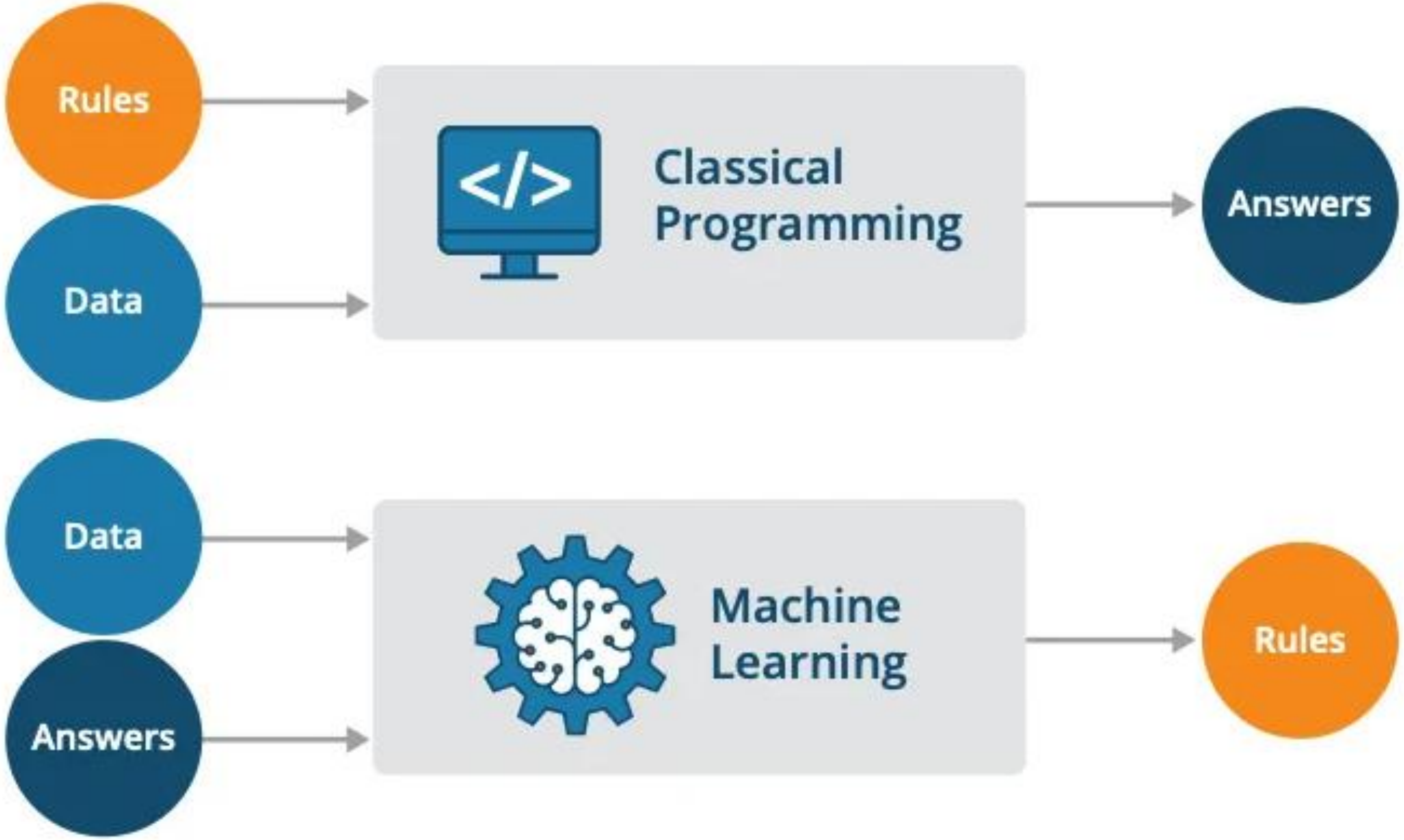
Algorithm vs Model

- **Model**: A set of rules that represent our data and can be used to make predictions.
- **Algoritme**: A procedure, or a set of steps, used to build a model.



"An algorithm is run on data to create a model"

Machine Learning vs Classical Programming



Thermostat example

Traditional approach

- The rule is given:

  "If temperature is smaller than 17°C, then heating is on, otherwise it's off"
- The algorithm implements the rule
- No data required to derive the rule!

```
threshold = 17
temperature = float(input("What is the temperature?\n"))  # data
heating = 'on' if temperature < threshold else 'off'      # rule
print(f'The heating is {heating}!')                       # answer
```

```
What is the temperature?
18
The heating is off!
```

Machine learning

The rule is not known and must be derived from data!

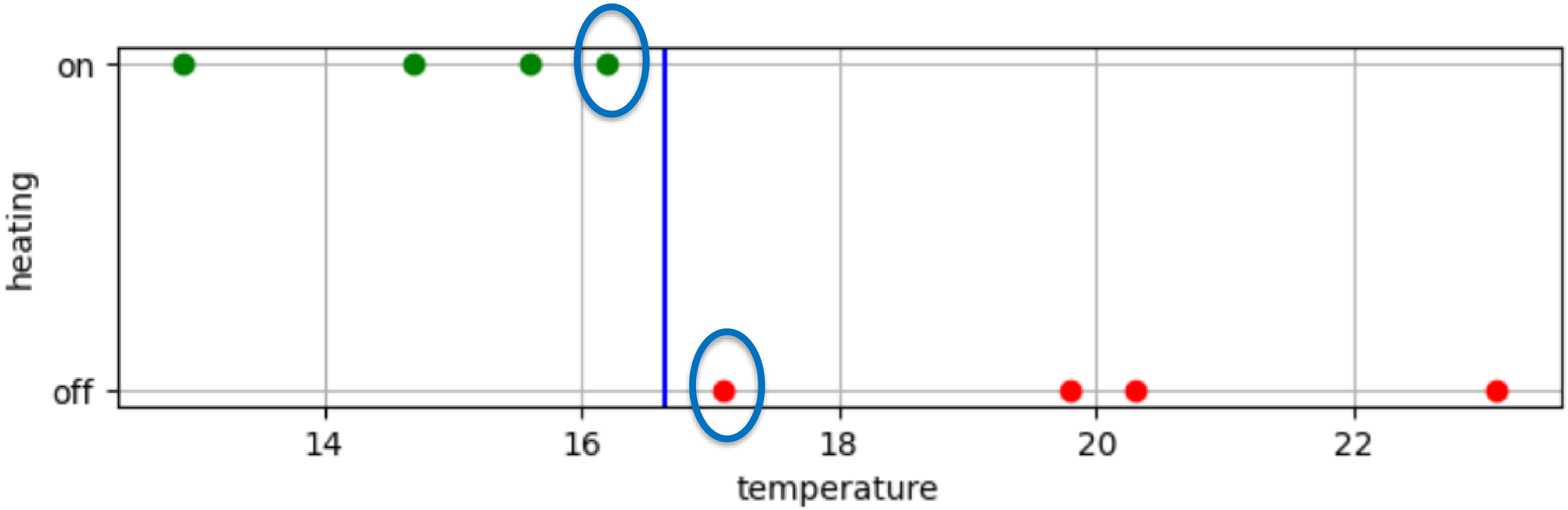
```python
import pandas as pd
temperature = [17.1, 15.6, 23.1, 19.8, 12.9, 20.3, 14.7, 16.2]  # data
heating = ['off', 'on', 'off', 'off', 'on', 'off', 'on', 'on']  # answers
table = pd.DataFrame(dict(temperature=temperature, heating=heating))
```

	temperature	heating
0	17.1	off
1	15.6	on
2	23.1	off
3	19.8	off
4	12.9	on
5	20.3	off
6	14.7	on
7	16.2	on

Naive algorithm

```
max_temperature_on = table[table.heating=='on']['temperature'].max()
min_temperature_off = table[table.heating=='off']['temperature'].min()
threshold = (max_temperature_on + min_temperature_off) / 2
print(f'maximum temperature if heating is on: {max_temperature_on}°C')
print(f'minimum temperature if heating is off: {min_temperature_off}°C')
print(f'threshold is {threshold}°C')
```

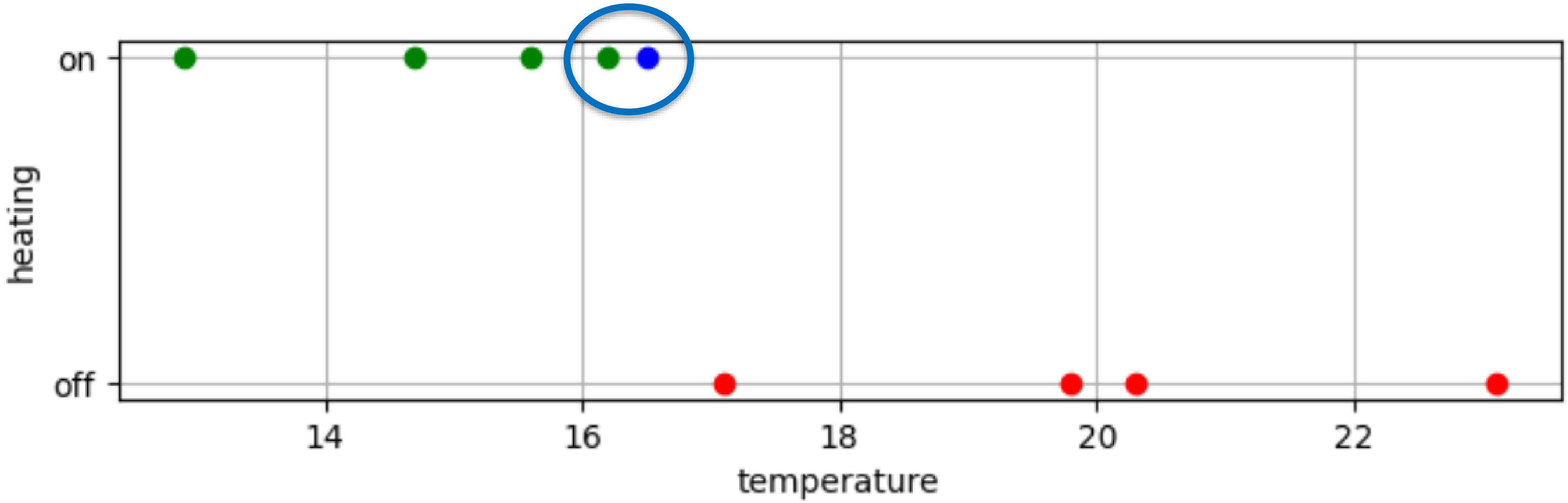
```
maximum temperature if heating is on: 16.2°C
minimum temperature if heating is off: 17.1°C
threshold is 16.65°C
```



Nearest neighbor

```
temperature = float(input("What is the temperature?\n"))  # input temperature
abs_difference = (temperature - table.temperature).abs()  # absolute difference
heating = table.heating.iloc[abs_difference.argmin()]     # label of nearest neighbor
print(f'The heating is {heating}!')                       # answer
```

```
What is the temperature?
16.5
The heating is on!
```



Some issues

- Real-life datasets are typically much larger:
  - more data points
  - more variables
- Real-life datasets may contain outliers and/or errors

- **Therefore we need more robust algorithms**
  - that use more than 1 or 2 samples only
  - that quantify and minimize the errors
- Examples:
  - **Logistic regression**: separates all data points instead of 2
  - **K Nearest Neighbors**: considers K nearest data points instead of 1

Dania International Days 2024

Workshop Machine Learning

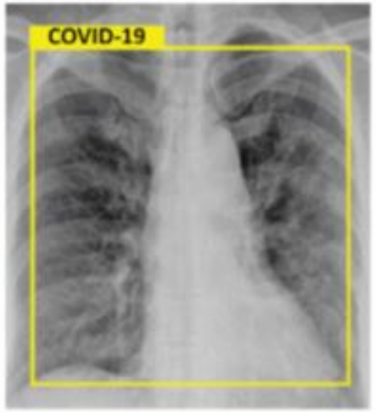
MACHINE LEARNING: APPLICATIONS & TASKS

DANIA ACADEMY

University of Applied Sciences

Machine Learning Applications

- Spam filters
- Recommender systems
- Personalized shopping
- Voice assistants
- Self-driving cars
- Search engines
- Chatbots
- Fraud prevention
- Face recognition
- Medical imaging
- Robotics
- Route planning
- Sales forecasting
- …



Machine Learning Tasks

- Classification
- Regression
- Forecasting
- Prediction
- Anomaly detection
- Association rule mining
- Clustering
- …

supervised learning

= A to B mapping

= Input to output mapping

= learning from (input, output) pairs

unsupervised learning

= learning from data without output

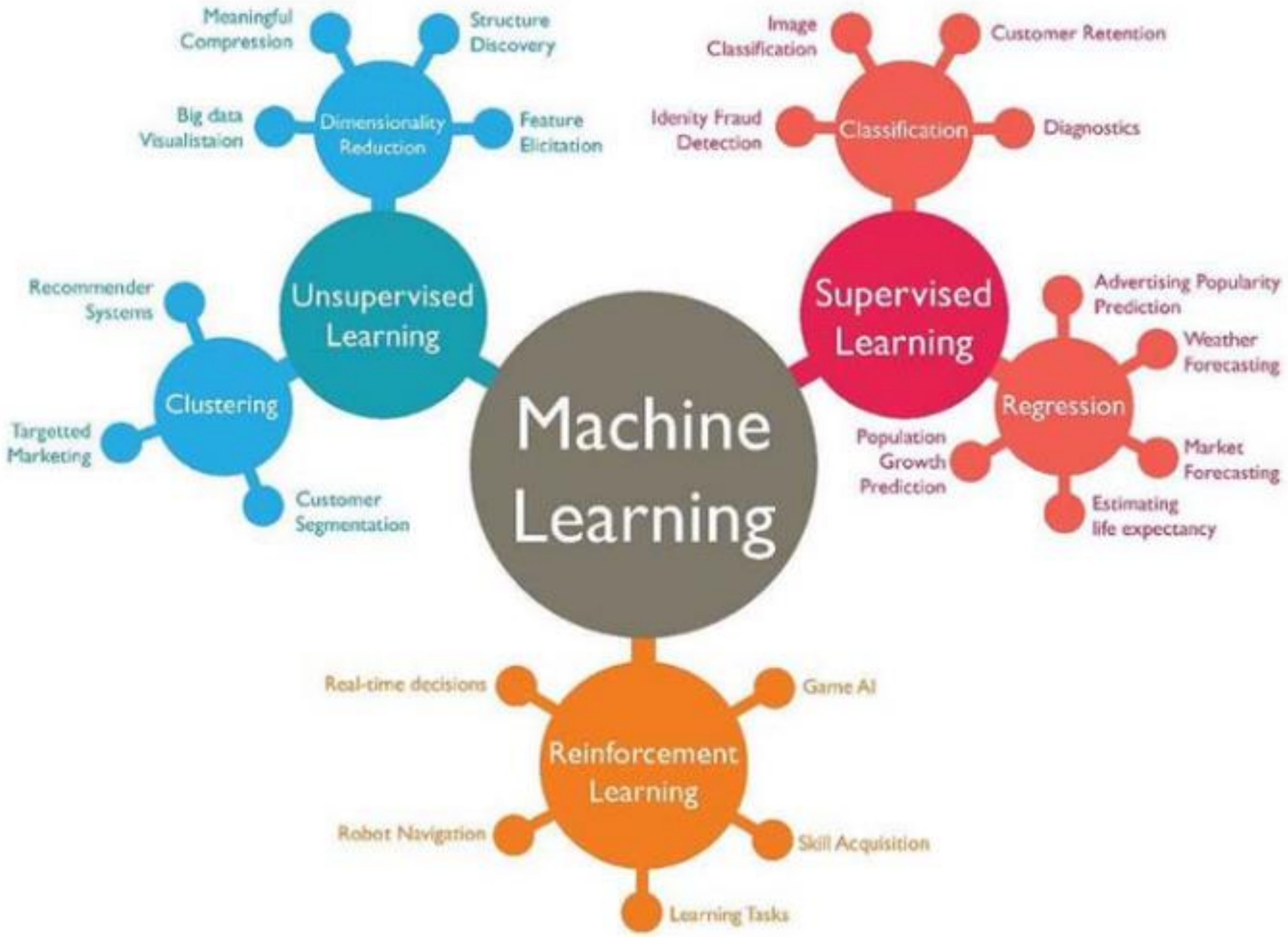
DANIA ACADEMY

vives University of Applied Sciences

Supervised Learning

Input (A)	Output (B)	Application
email	spam? (0/1)	spam filtering
audio	text transcript	speech recognition
English	Chinese	machine translation
ad, user info	click? (0/1)	online advertising
image, radar info	position of other cars	self-driving car
image of phone	defect? (0/1)	visual inspection

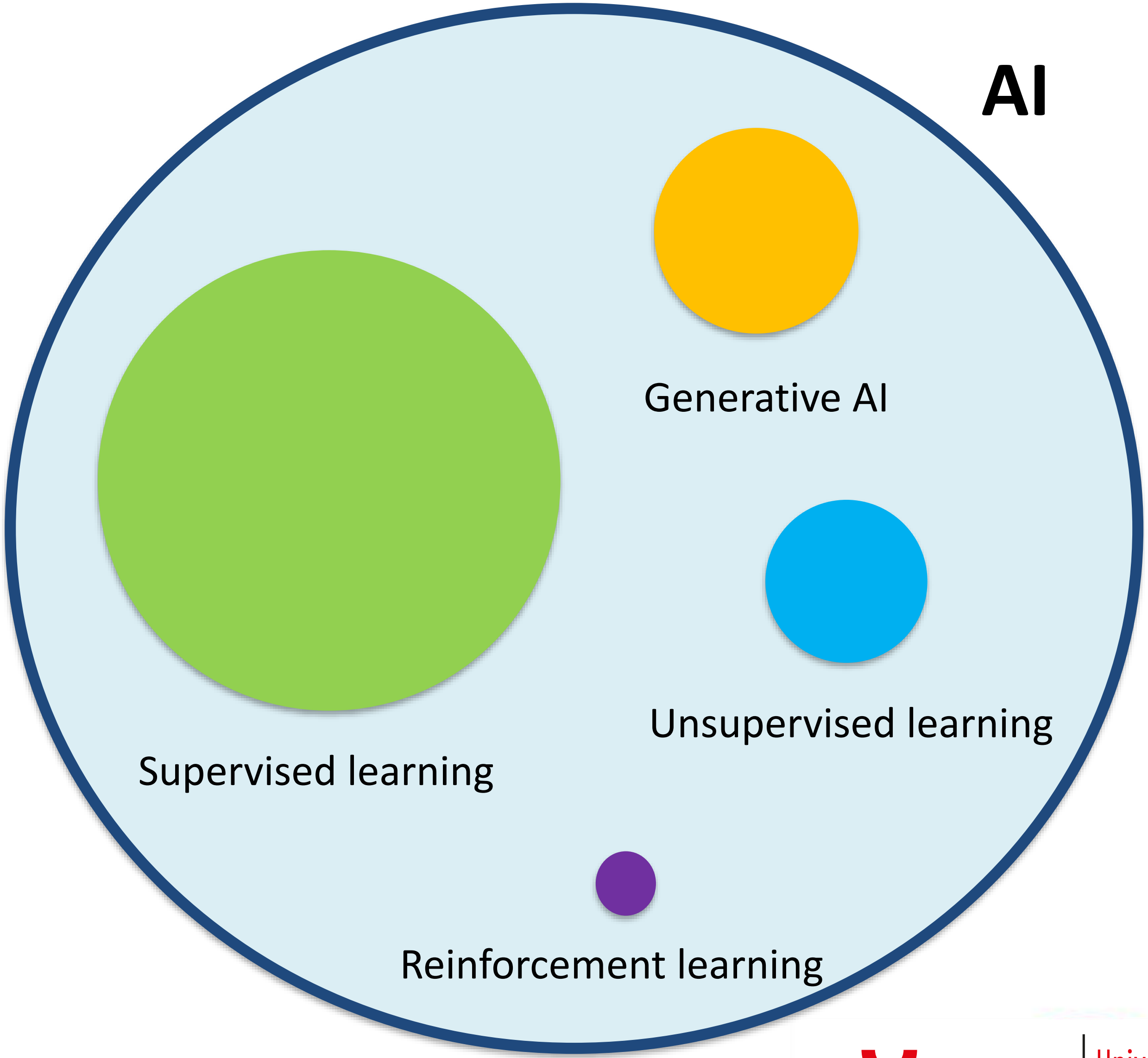
The Big Three



Category of machine learning. Image by https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/

What about GenAI?

- **Supervised learning**
  = learning from <u>labeled</u> data
- **Unsupervised learning**
  = learning from <u>unlabeled</u> data
- **Reinforcement learning**
  = learning from rewards
- **Generative AI**
  = generating new data



DANIA ACADEMY

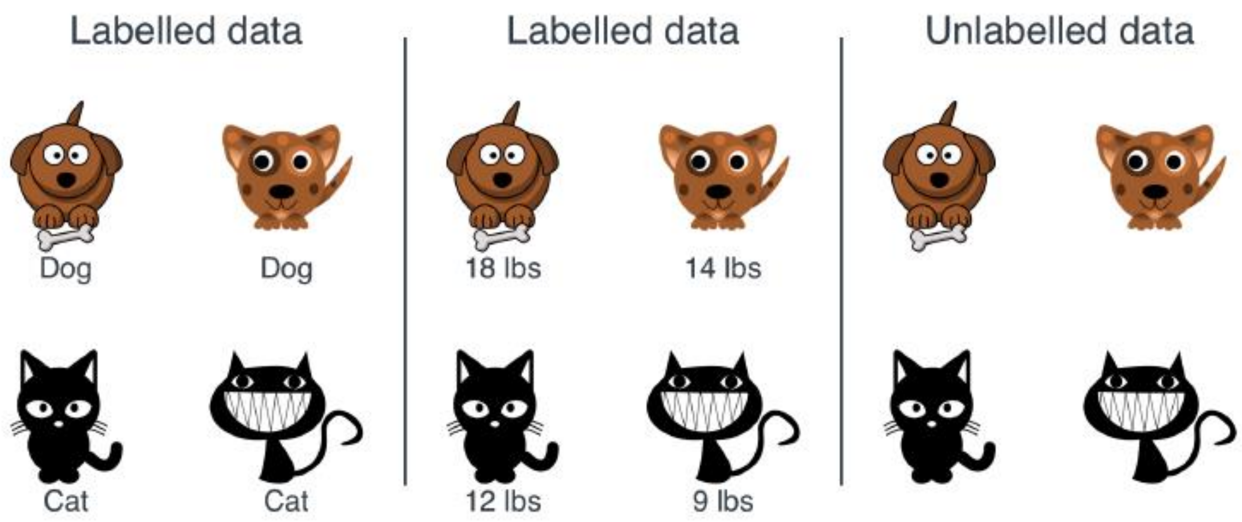
https://youtu.be/5p248yoa3oE?si=M6TY-fAtB2l1S7iF

vives University of Applied Sciences

Supervised vs Unsupervised

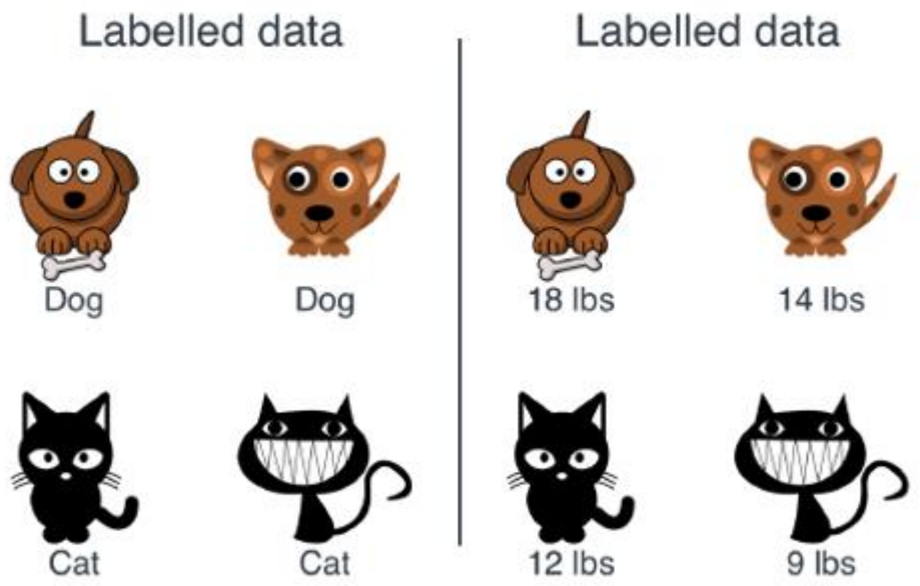
- **Labeled data**: data with label
  → **SUPERVISED** LEARNING

- **Unlabeled data**: data without label
  → **UNSUPERVISED** LEARNING

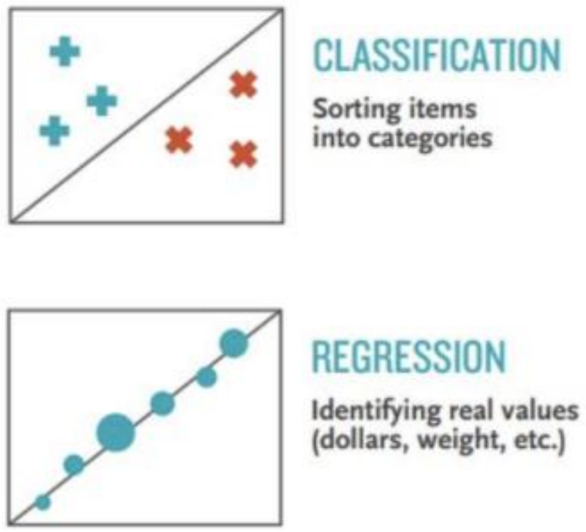


Classification vs Regression

- **Categorical** target ➔ **Classification**
- **Numerical** target ➔ **Regression**



Categorical **Numerical**



Structured Data

- **Data** = information (= table)
- **Example** = sample = instance = data point (= table row/record)
- **Feature** = independent variable (= table column/attribute)
- **Target** = labels = dependent variable = feature we want to predict

data(set)

	A	B	C	D	E	F
1	id	date	size	typos	recipients	spam
2	0	12/01/2021	2.5	0	1	False
3	1	13/01/2021	1.3	0	2	False
4	2	14/01/2021	12.1	3	15	True
5	3	15/01/2021	7.8	2	19	True
6	4	16/01/2021	4.6	1	5	False
7	5	17/01/2021	9.8	5	1	True
8	6	18/01/2021	11.6	3	63	True

example

feature

target

Dania International Days 2024

Workshop Machine Learning

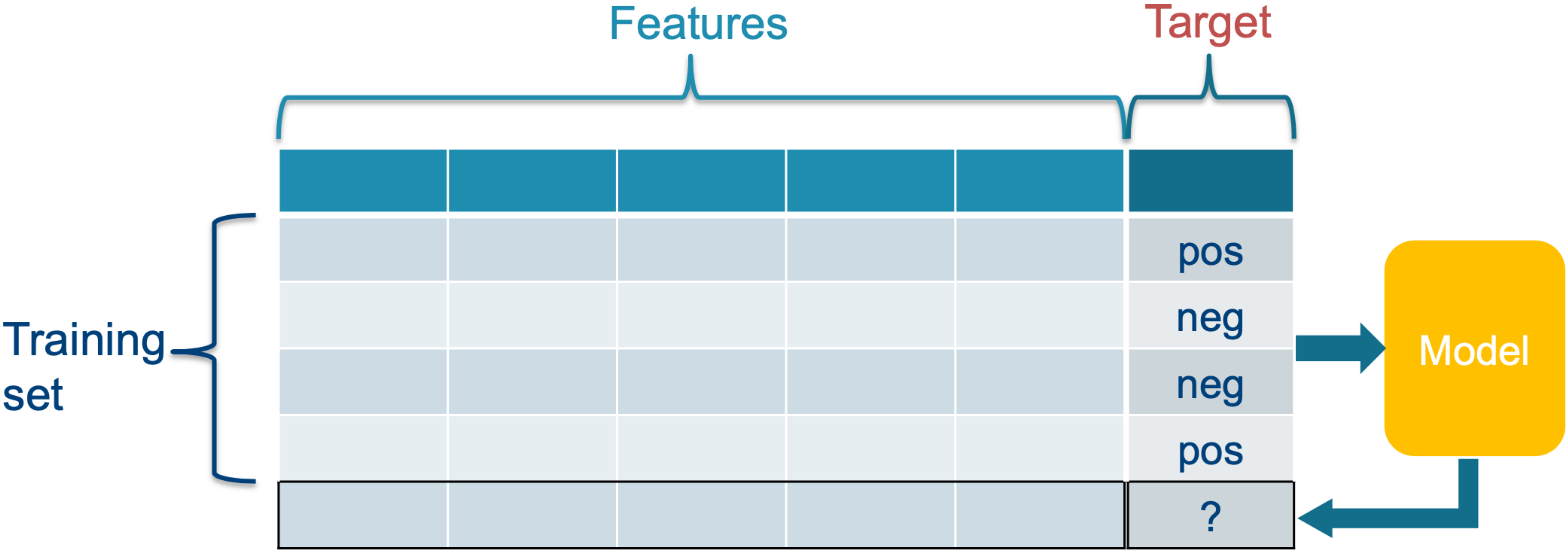
MACHINE LEARNING: SUPERVISED LEARNING

DANIA ACADEMY

University of Applied Sciences

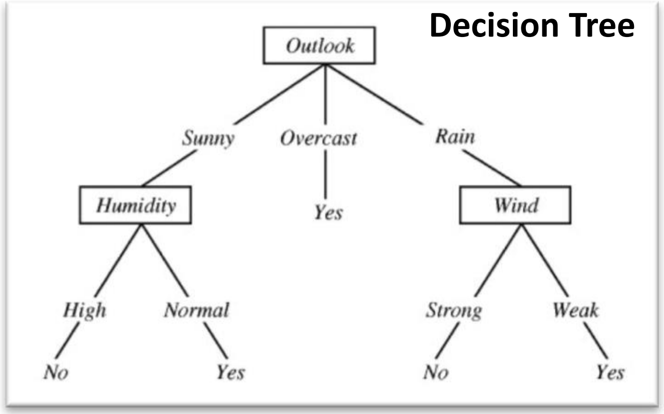
Supervised Learning

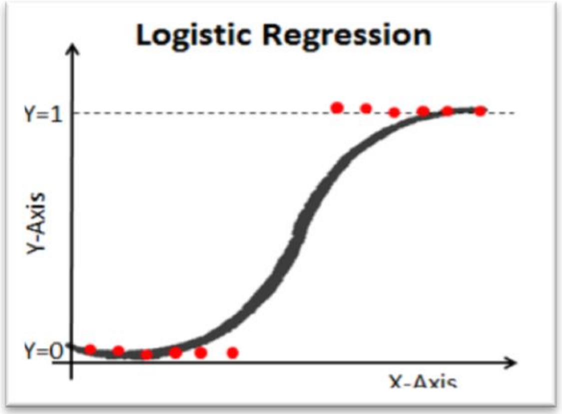
- <u>Task</u>: learn a model to predict a target for new data instances, based on a training set of data instances for which the target is known

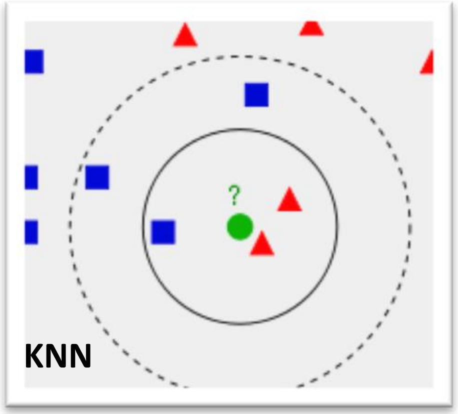


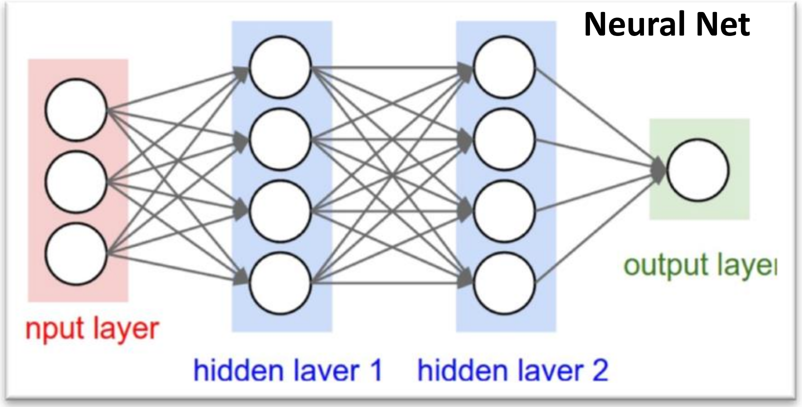
Supervised Learning Algorithms

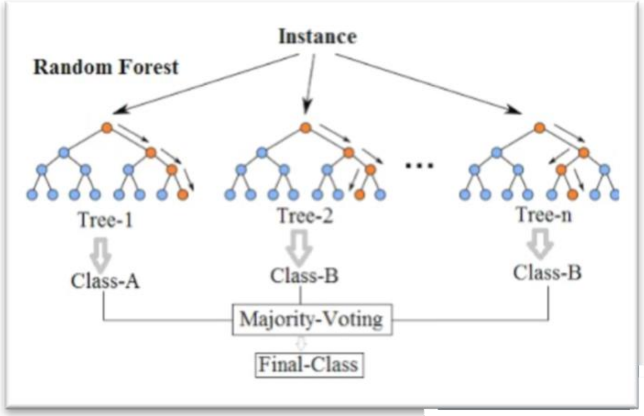
- There exist plenty of supervised learning algorithms
- **No free lunch**: there is no algorithm that works best for every problem







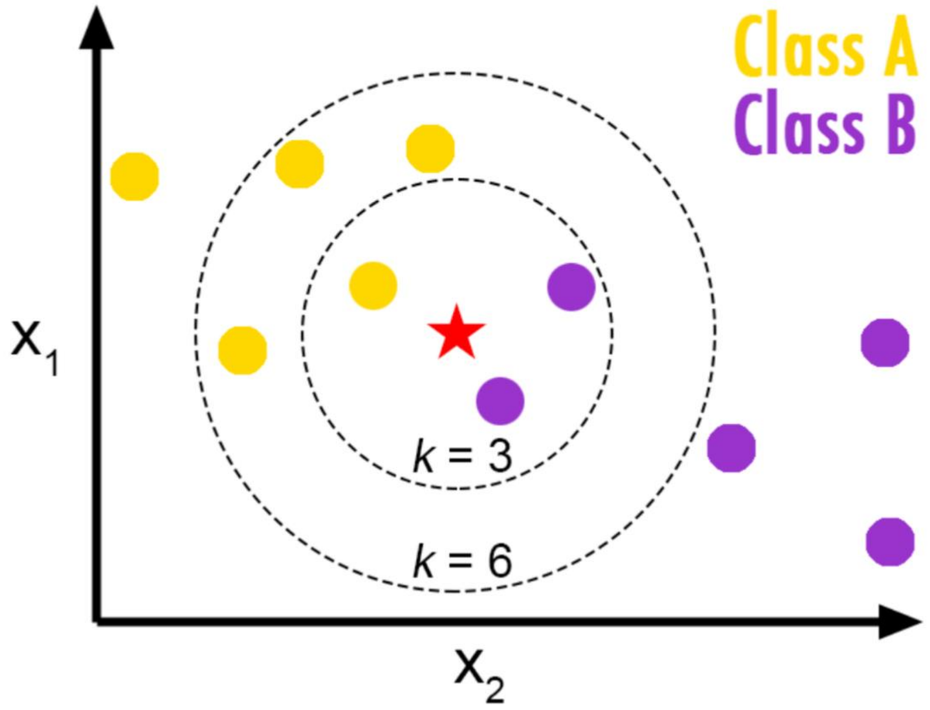




...

K Nearest Neighbors (KNN)

- Classification (regression is also possible)
- Requires no training (= lazy learning, as opposed to eager learning)
- Main task: find suitable distance function (Euclidean, Manhattan, …)

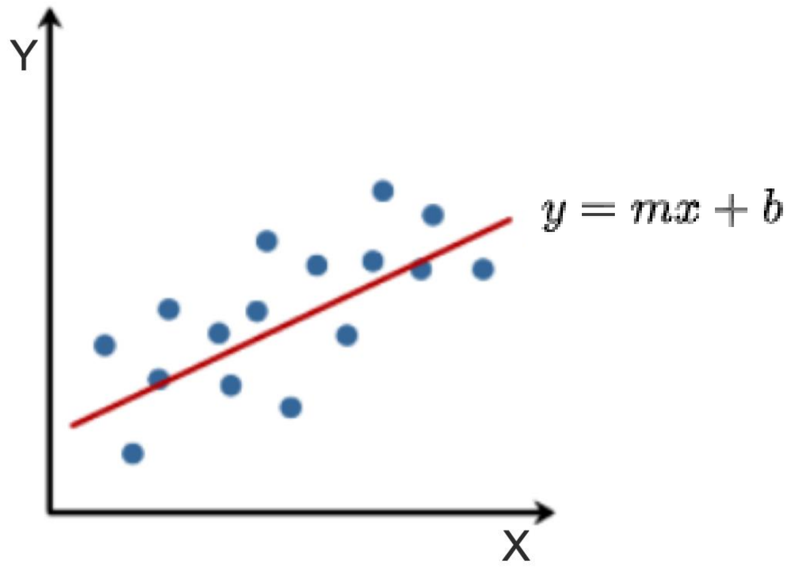


DANIA ACADEMY

vives University of Applied Sciences

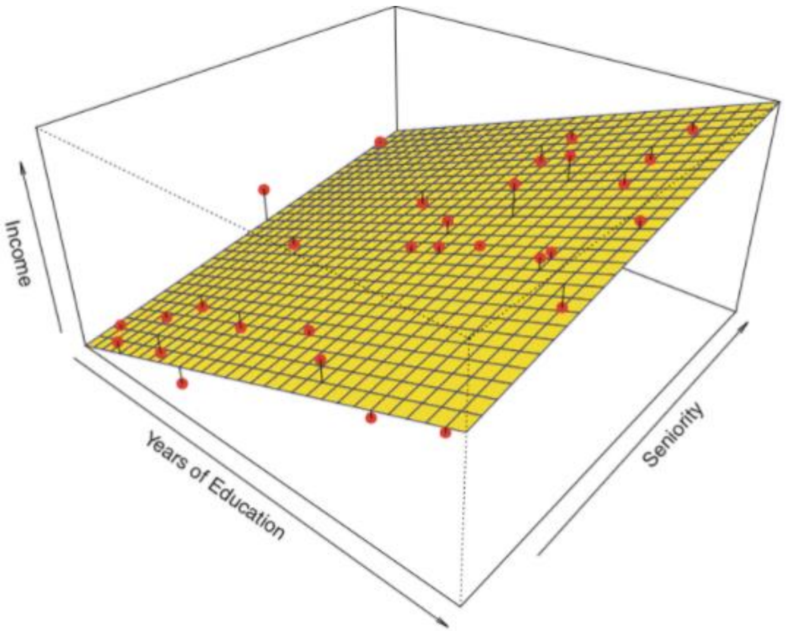
Simple Linear Regression

- Regression for numeric targets
- 1 independent variable (feature *x*) and 1 dependent variable (target *y*)
- Main task: estimate parameters *m* and *b*, such that predictions (red line) and targets (blue dots) are as close as possible (= best-fitting straight line)

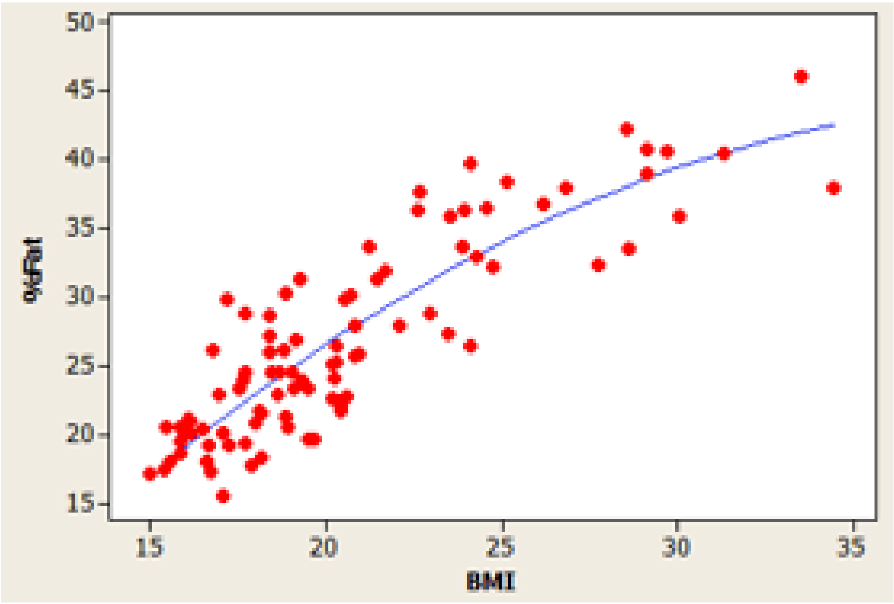


Linear & Nonlinear Regression

- Linear Regression 2 features and 1 target (left)
- Nonlinear regression 1 feature and 1 target (right)



Source: James et al. *Introduction to Statistical Learning* (Springer 2013)



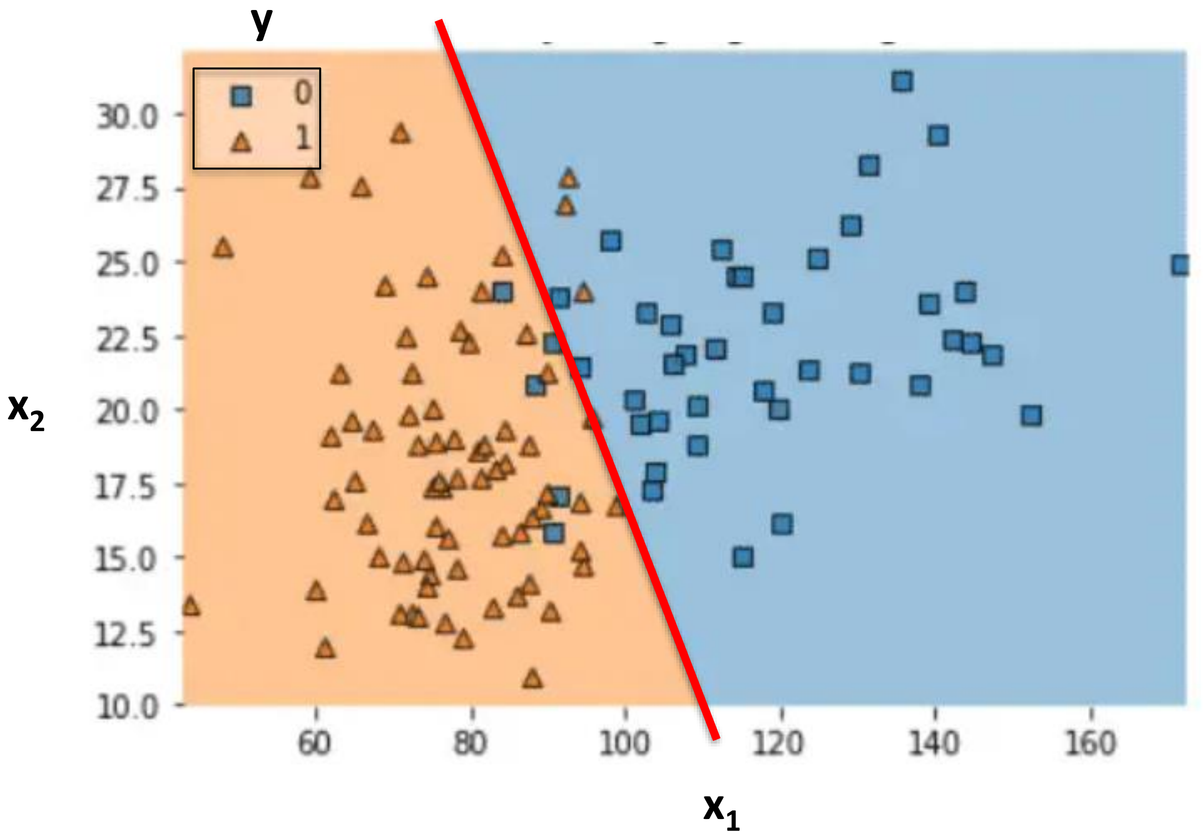
Source: the minitab blog

DANIA ACADEMY

vives University of Applied Sciences

Logistic Regression

- Regression for binary targets
- Features x_i and target y
- Main task: find a separating straight line
  = binary classification
- N dimensions: separating hyperplane

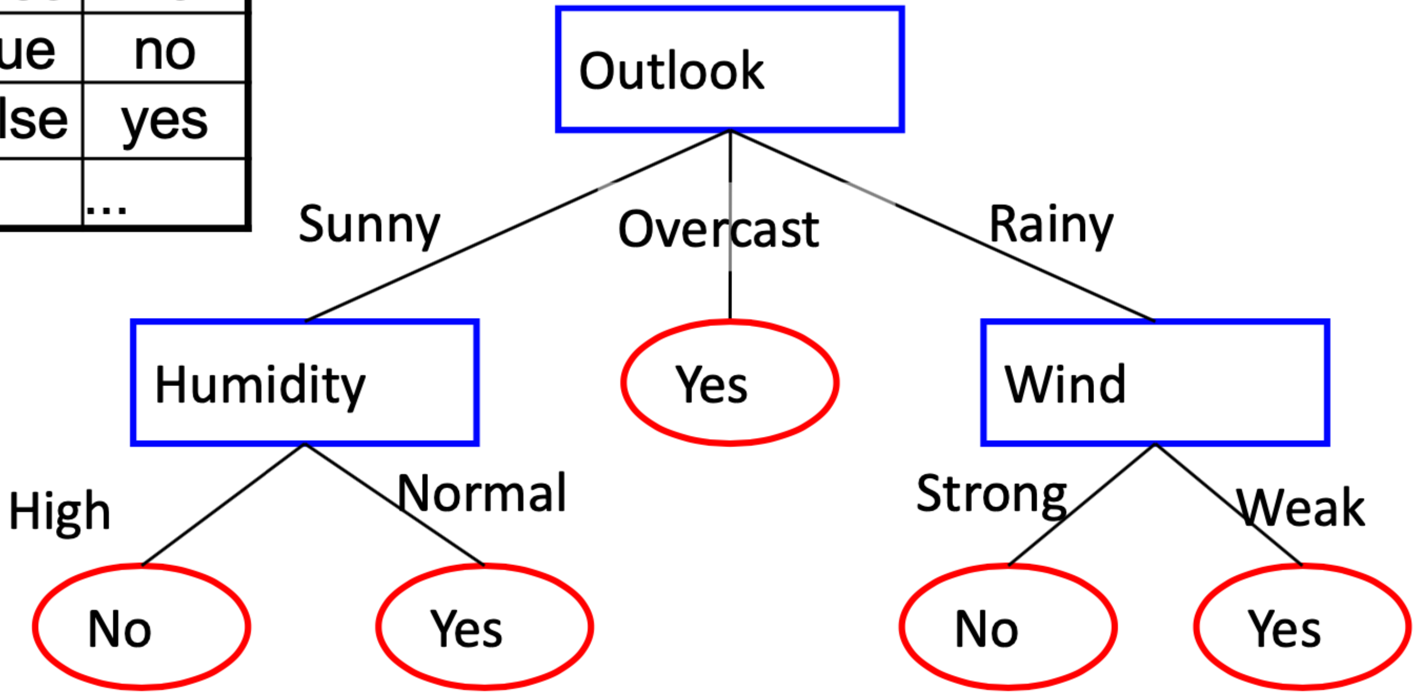


source: https://www.jcchouinard.com/logistic-regression/

Decision Tree

- Classification (regression is also possible)
- Example: Play tennis or not? (depending on weather conditions)

Outlook	Temp.	Hum.	Wind	Play?
Sunny	85	85	False	no
Sunny	80	90	True	no
Overcast	83	86	False	yes
...	...	...	...	...

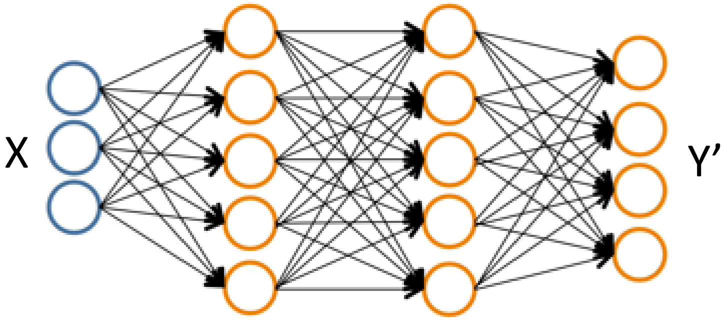


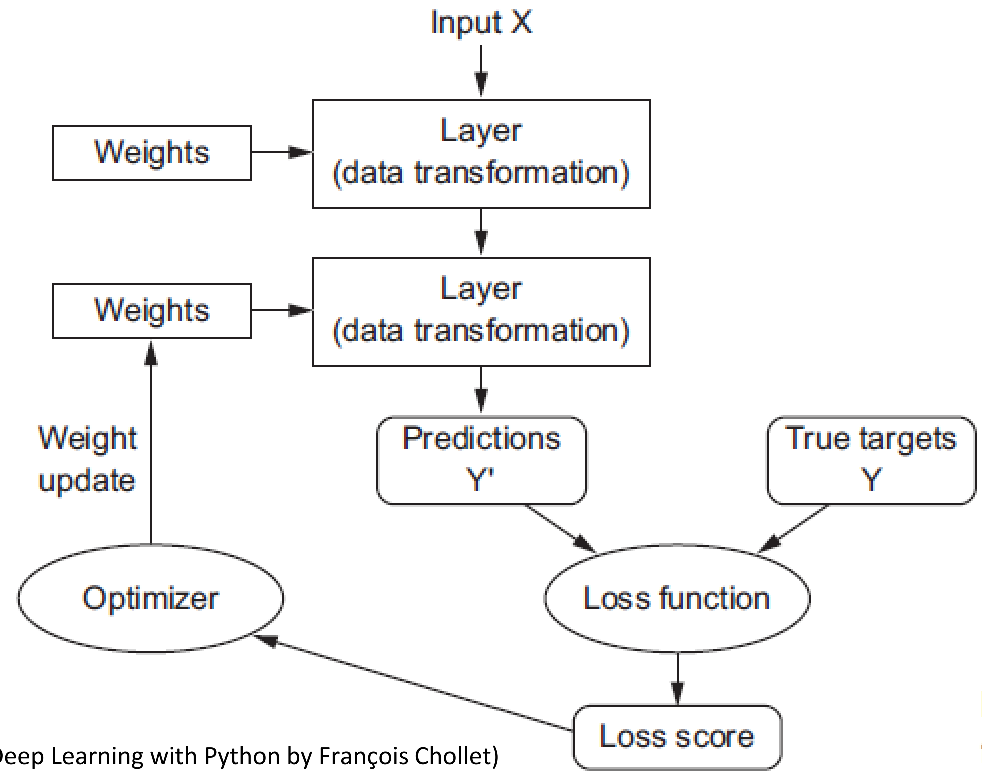
- Leaf nodes versus internal nodes

  = labels = features

Artificial Neural Network

- Regression or classification
- Features X and targets Y
- **Loss**: function quantifying differences
  between targets Y and predictions Y’
- Main task: find optimal weights
  that minimize the loss





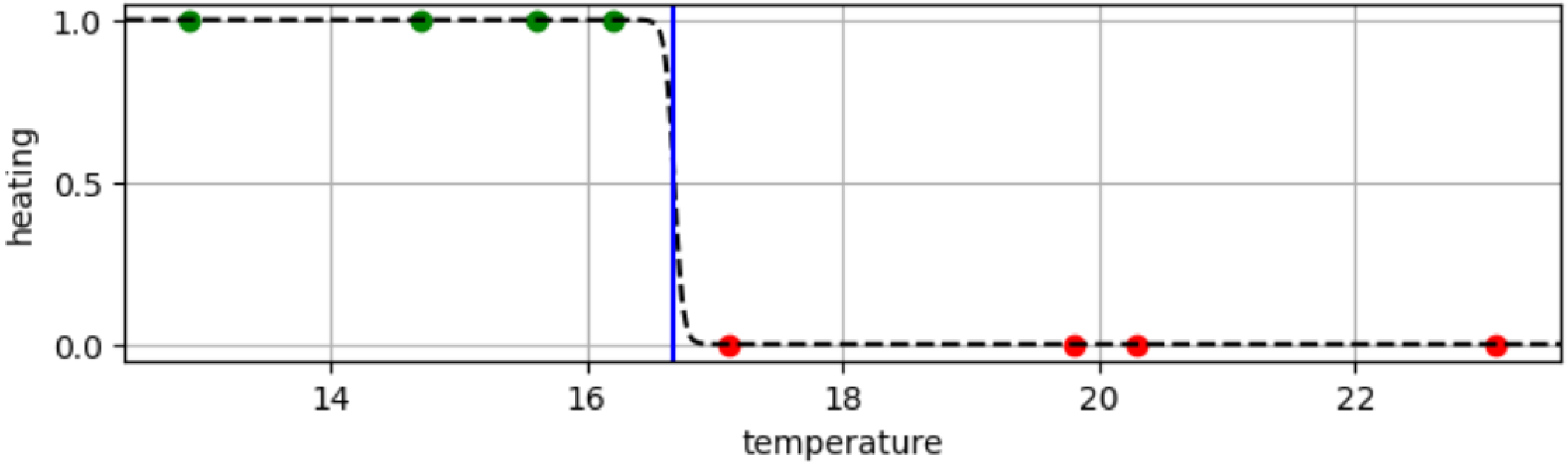
(Source: Deep Learning with Python by François Chollet)

Thermostat example

Logistic Regression

```python
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(penalty=None)  # instantiate
model.fit(table[['temperature']].values, table.heating=='on')  # fit data
threshold = -model.intercept_.item() / model.coef_.item()  # determine threshold
print(f'threshold is {threshold}°C')
model.predict([[17]]).item()  # predict label for new temperature value
```

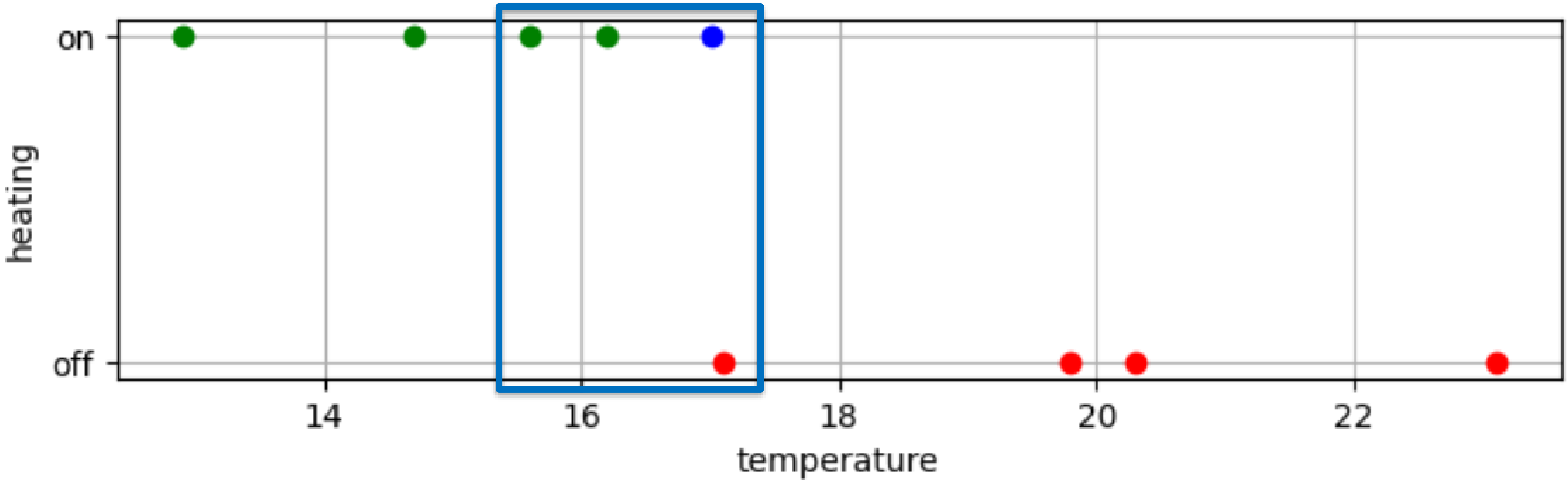
```
threshold is 16.681991552397978°C
False
```



K Nearest Neighbors

```
from sklearn.neighbors import KNeighborsClassifier
model = KNeighborsClassifier(n_neighbors=3)  # instantiate with K = 3
model.fit(table[['temperature']].values, table.heating=='on')  # fit data
model.predict([[17.0]]).item()  # predict label for new temperature value
```

```
True
```

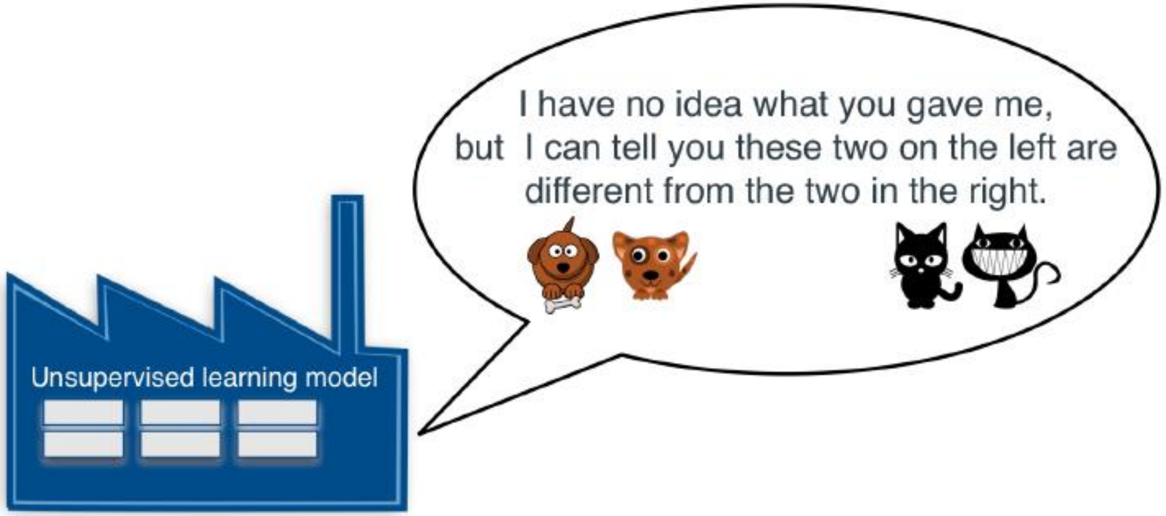


Dania International Days 2024

Workshop Machine Learning

MACHINE LEARNING: UNSUPERVISED LEARNING

Unsupervised Learning



- Data are **not labeled**
- Often used during data **preprocessing**

- Clustering: grouping data based on similarities
- Dimensionality reduction: reducing the number of features while retaining as much meaningful information as possible
- Matrix factorization: decomposing the data in order to discover latent features

Clustering

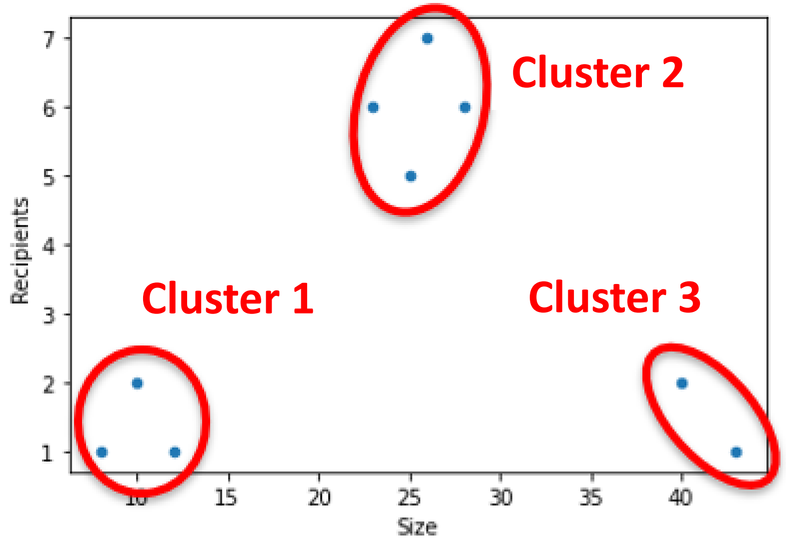
Applications:

- Genetics: grouping species based on similarities
- Medical imaging: partitioning images based on tissue structures
- Market segmentation: clustering customers based on demographics, income, etc.
- Mails:

No labels!

E-mail	Size	Recipients
1	8	1
2	12	1
3	43	1
4	10	2
5	40	2
6	25	5
7	23	6
8	28	6
9	26	7

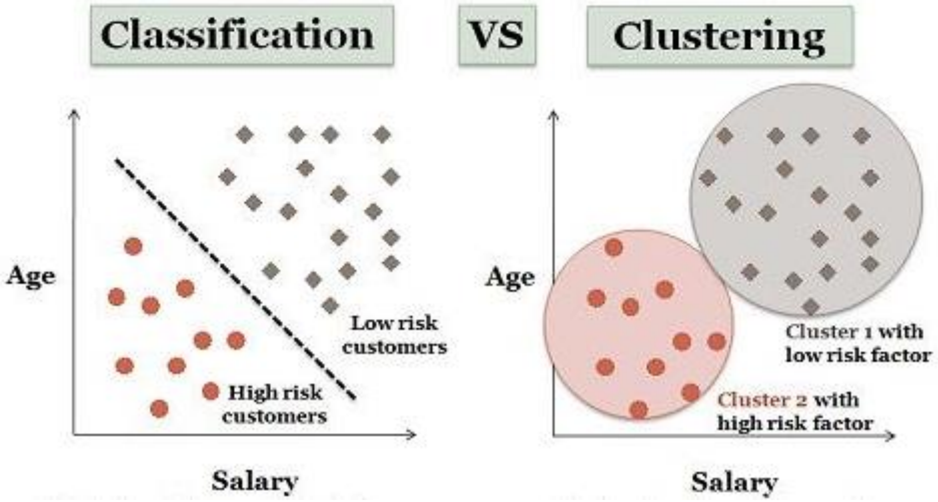
mails.csv



Clustering vs Classification

- **Classification**: labeled data → classes already exist
- **Clustering**: unlabeled data → classes don't exist yet

customer	age	salary	risk
0	23	1500	high
1	51	2500	low
2	42	3100	low
3	36	1900	high
4	67	2100	low



customer	age	salary	risk
0	23	1500	?
1	51	2500	?
2	42	3100	?
3	36	1900	?
4	67	2100	?

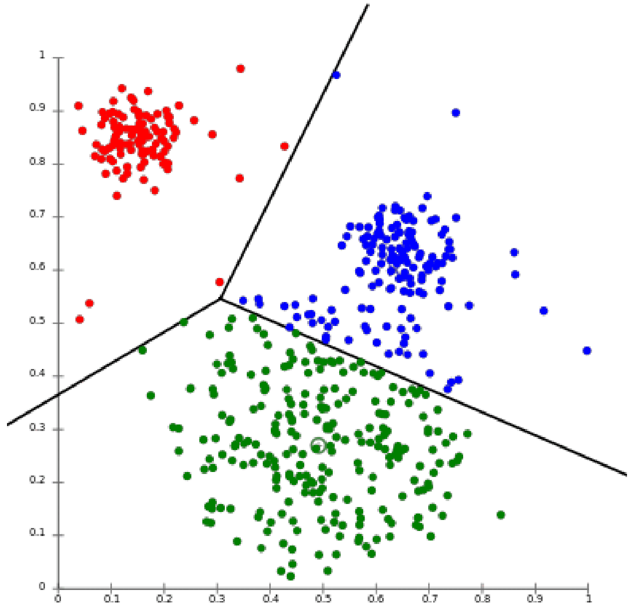
(source: https://techdifferences.com/difference-between-classification-and-clustering.html)

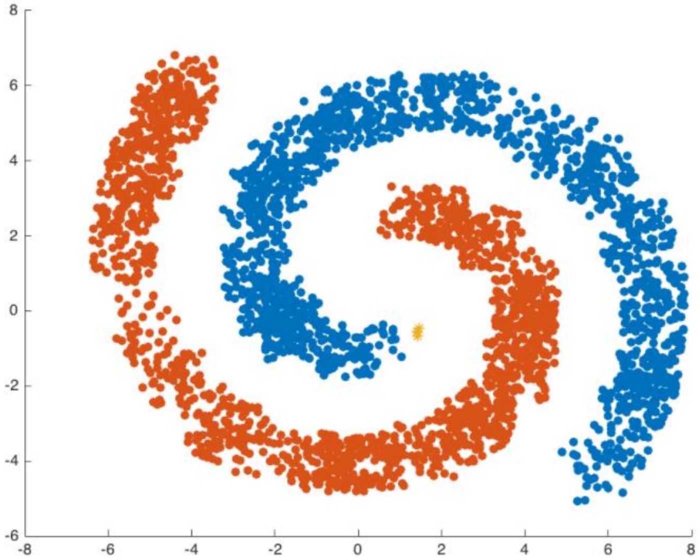
Clustering Algorithms

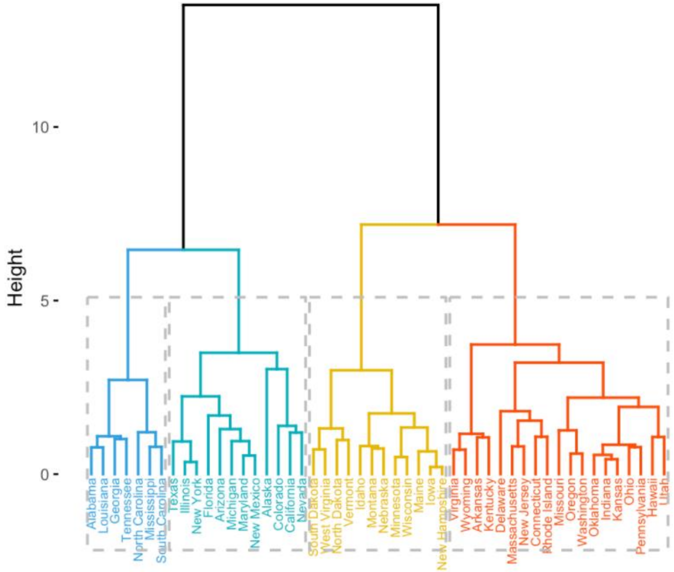
- K-means clustering

  https://youtu.be/nXY6PxAaOk0

- Hierarchical clustering (dendrogram)
- Gaussian mixture models
- DBSCAN
- …

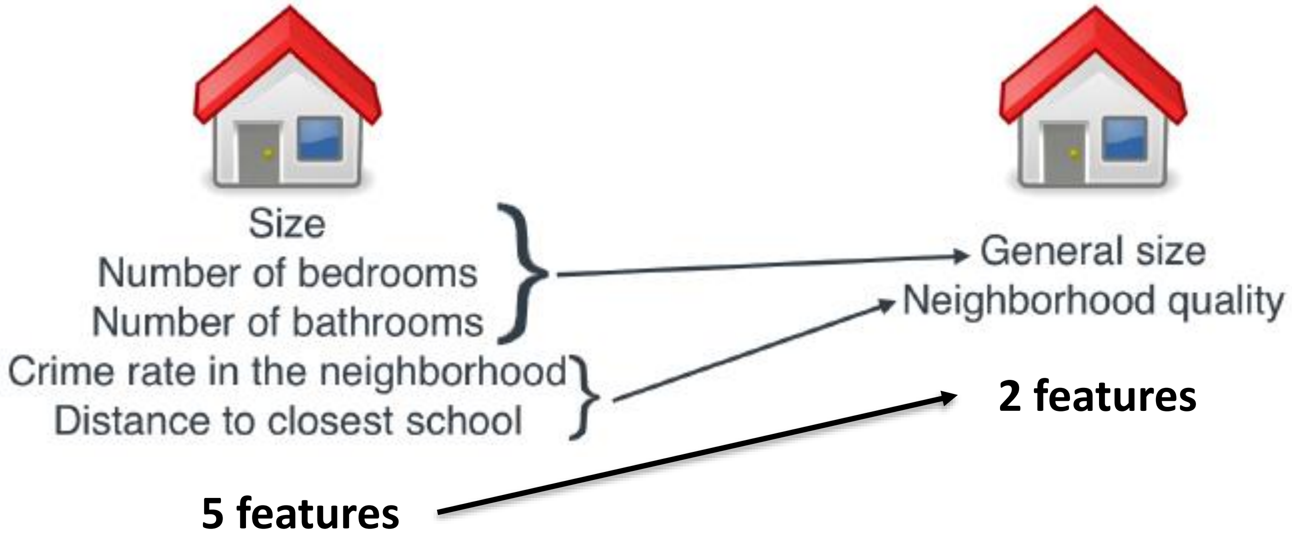




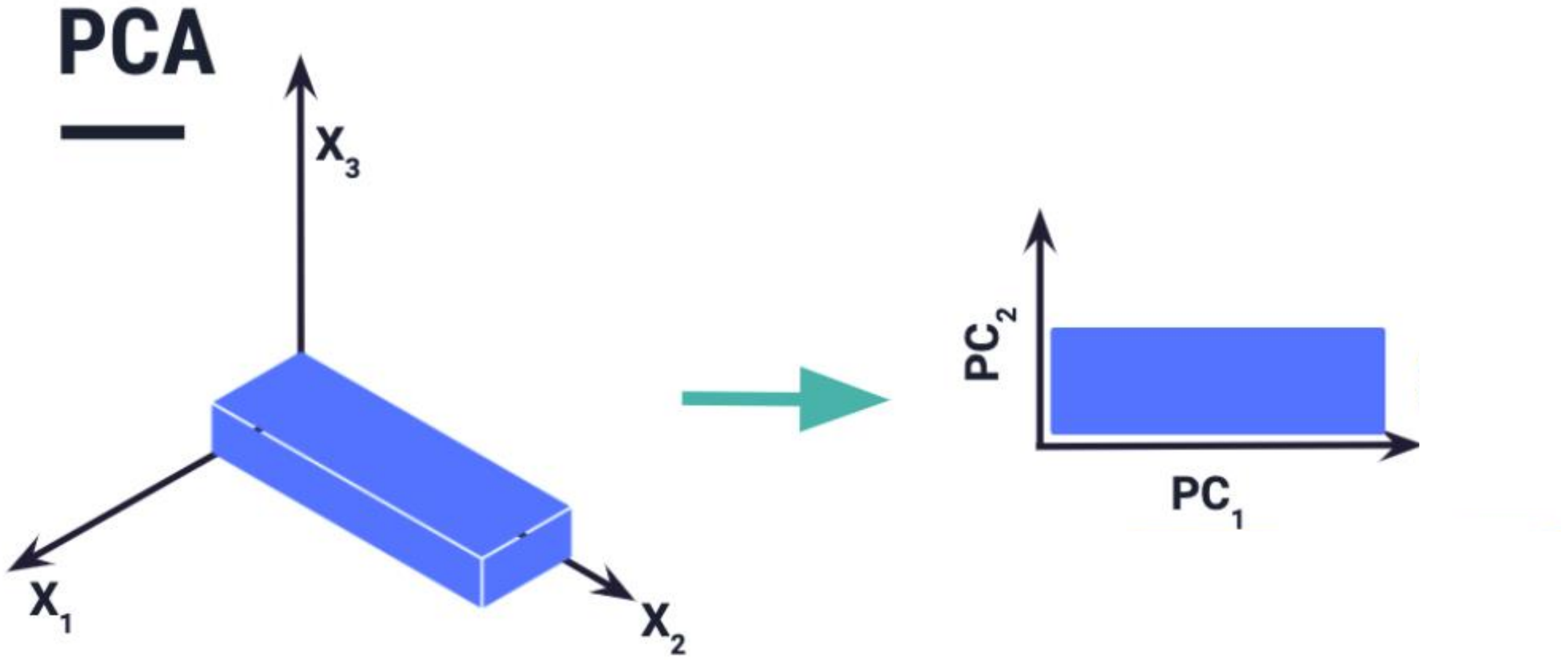


Dimensionality Reduction

- Number of dimensions = number of features
- Reducing the number of features

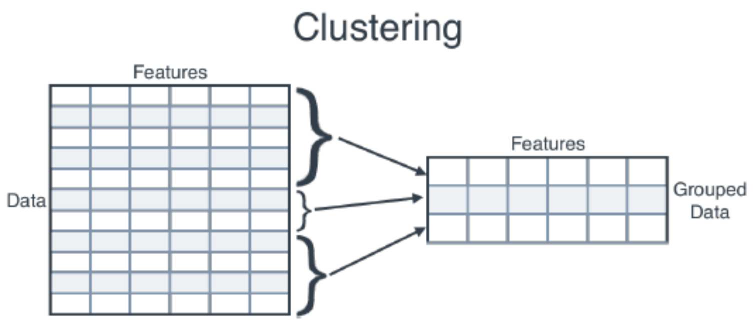


Example: Principal Component Analysis

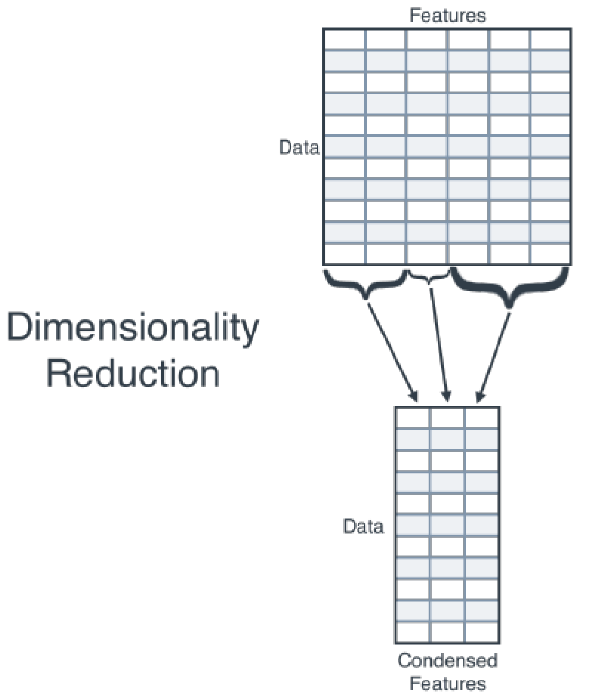


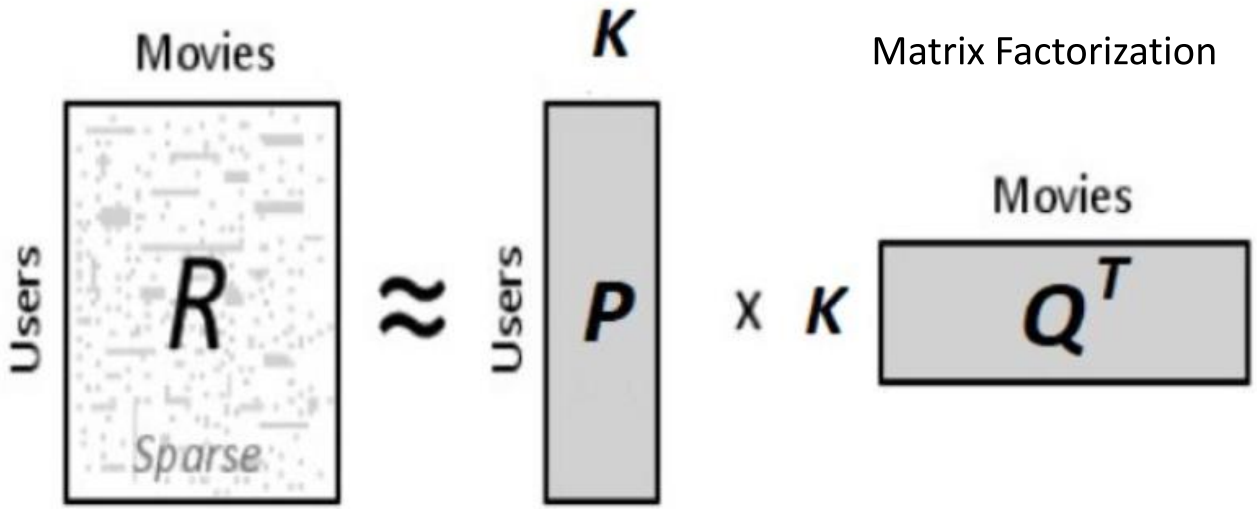
(source: https://knowledge.dataiku.com/latest/ml-analytics/statistics/concept-principal-component-analysis-pca.html)

Matrix Factorization



- **Clustering**: reducing samples (= rows)
- **Dimensionality Reduction**: reducing features (= columns)
- **Matrix Factorization**: reducing both rows and columns





Matrix Factorization

(source: https://www.kaggle.com/code/residentmario/notes-on-matrix-factorization-machines)

Example: Recommender Systems

Matrix Factorization

	M1	M2	M3	M4	M5
Comedy	3	1	1	3	1
Action	1	2	4	1	3

	Comedy	Action
A	✓	✗
B	✗	✓
C	✓	✗
D	✓	✓

	M1	M2	M3	M4	M5
	3	1	1	3	1
	1	2	4	1	3
	3	1	1	3	1
	4	3	5	4	4

NETFLIX

https://youtu.be/ZspR5PZemcs

Dania International Days 2024

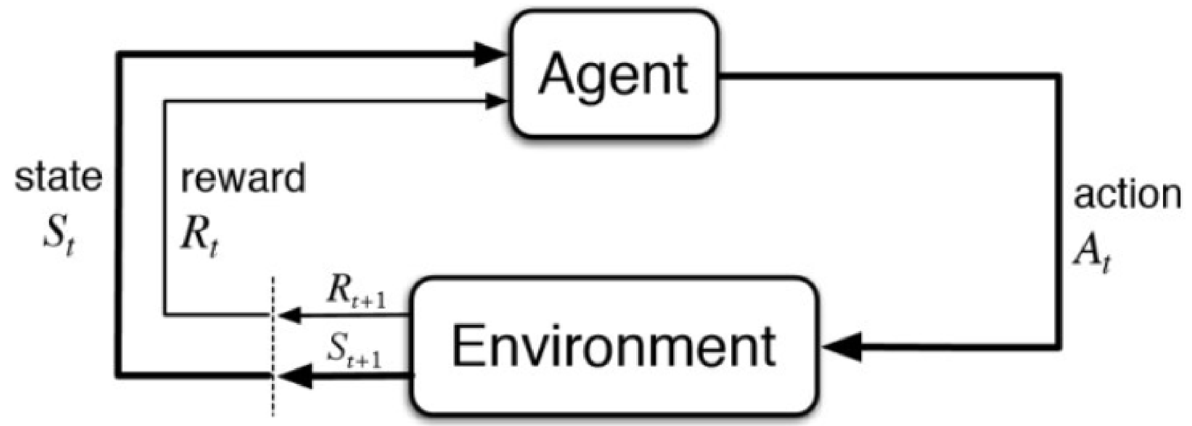
Workshop Machine Learning

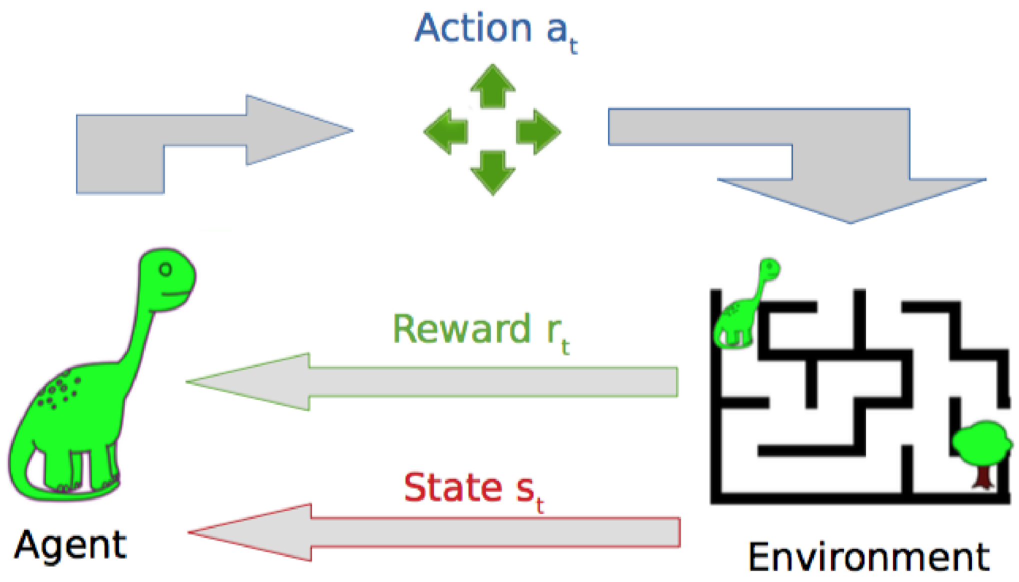
MACHINE LEARNING: REINFORCEMENT LEARNING

DANIA ACADEMY

vives University of Applied Sciences

Reinforcement Learning





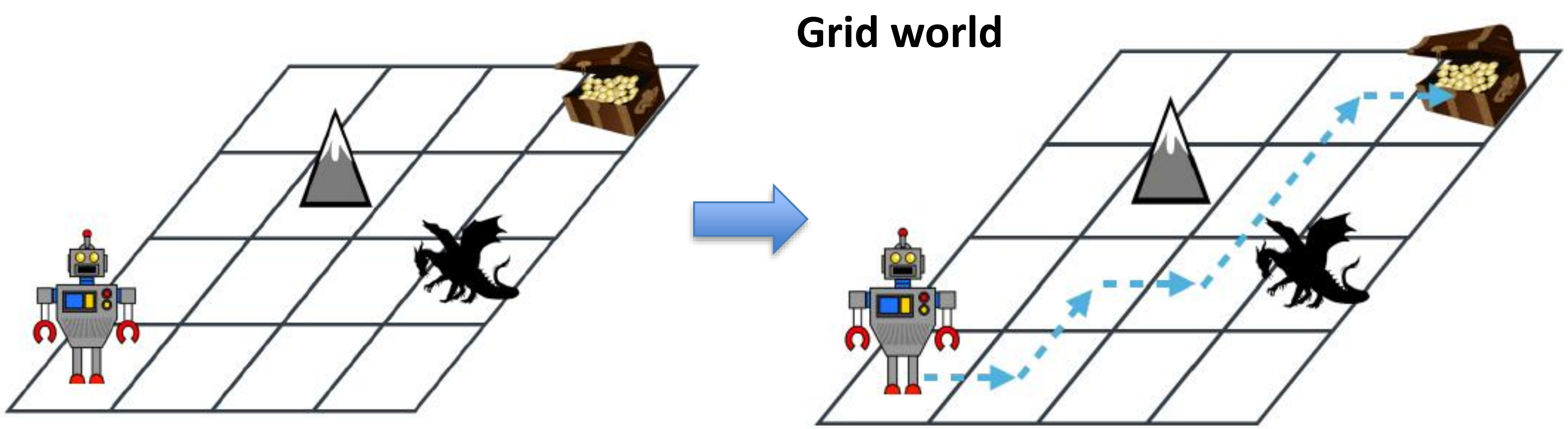
(source: https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292)

DANIA ACADEMY

vives University of Applied Sciences

Applications

- Robotics
- Self-driving cars
- Games
- …



AlphaGo en AlphaZero

https://deepmind.google/technologies/alphago/

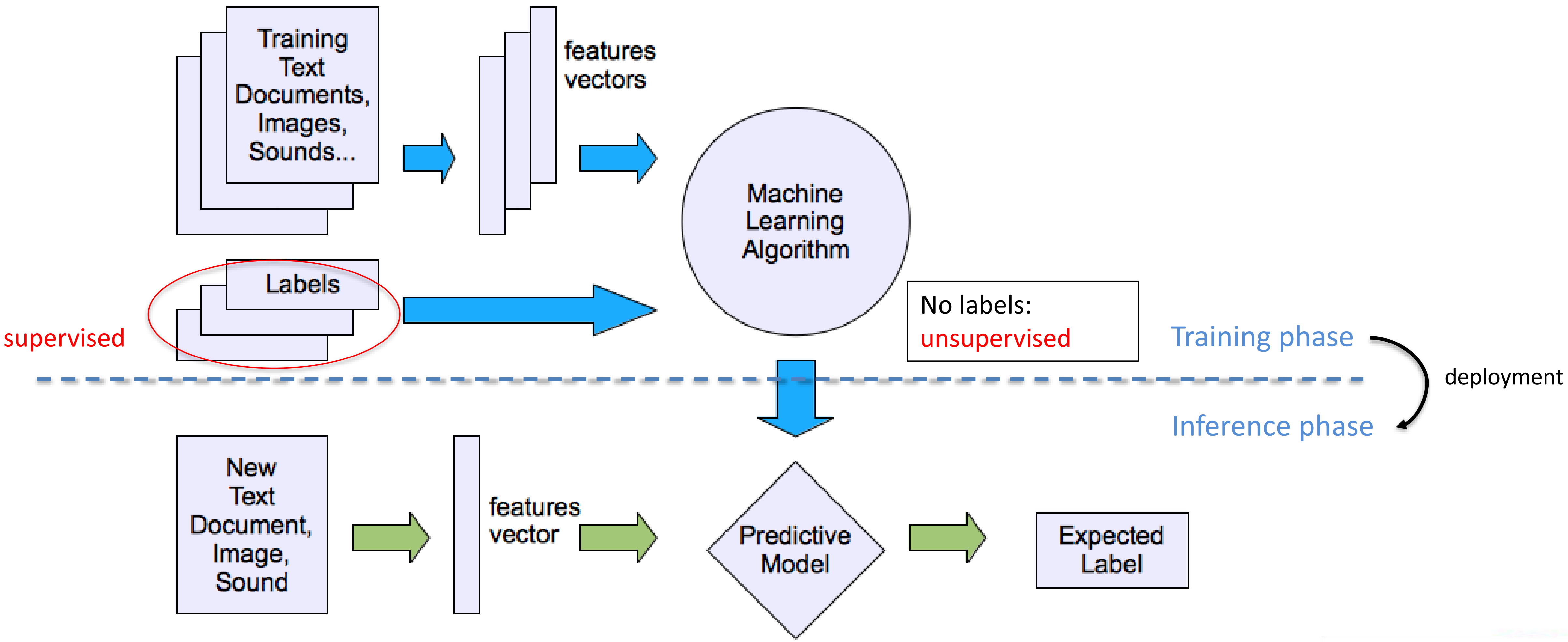
AlphaGo - the movie

Dania International Days 2024

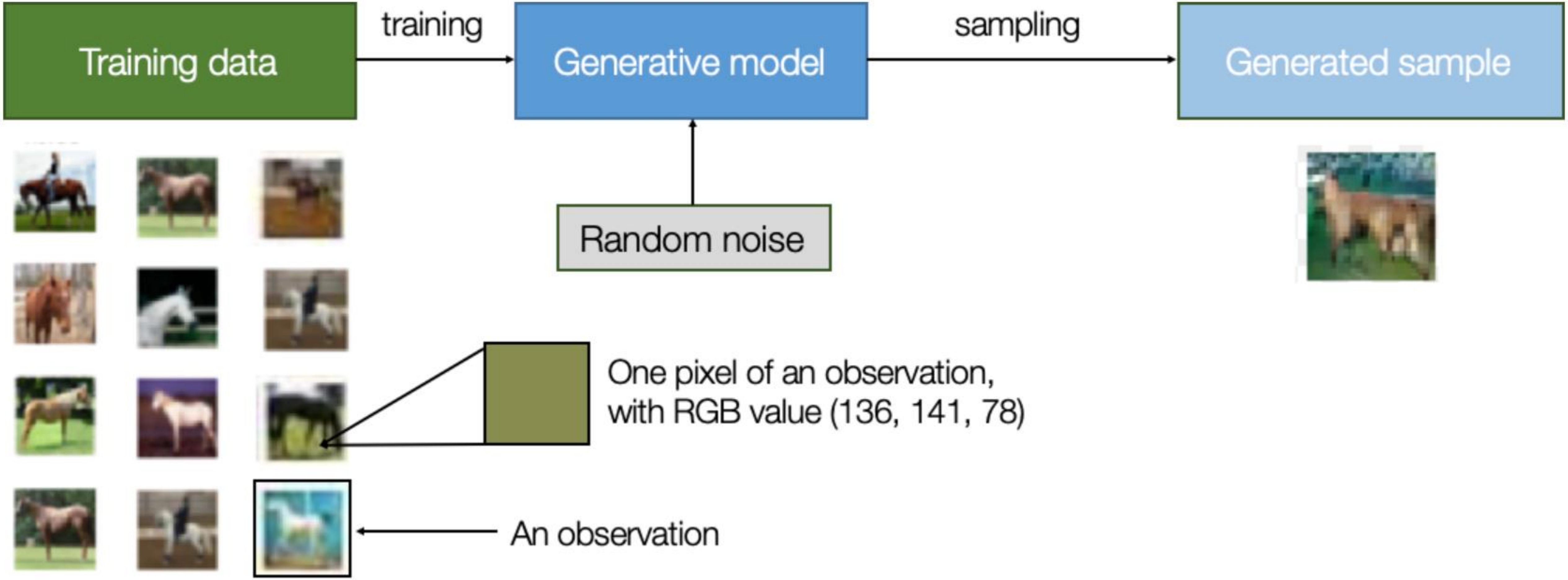
Workshop Machine Learning

MACHINE LEARNING: TRAINING AND EVALUATION

Training vs Inference



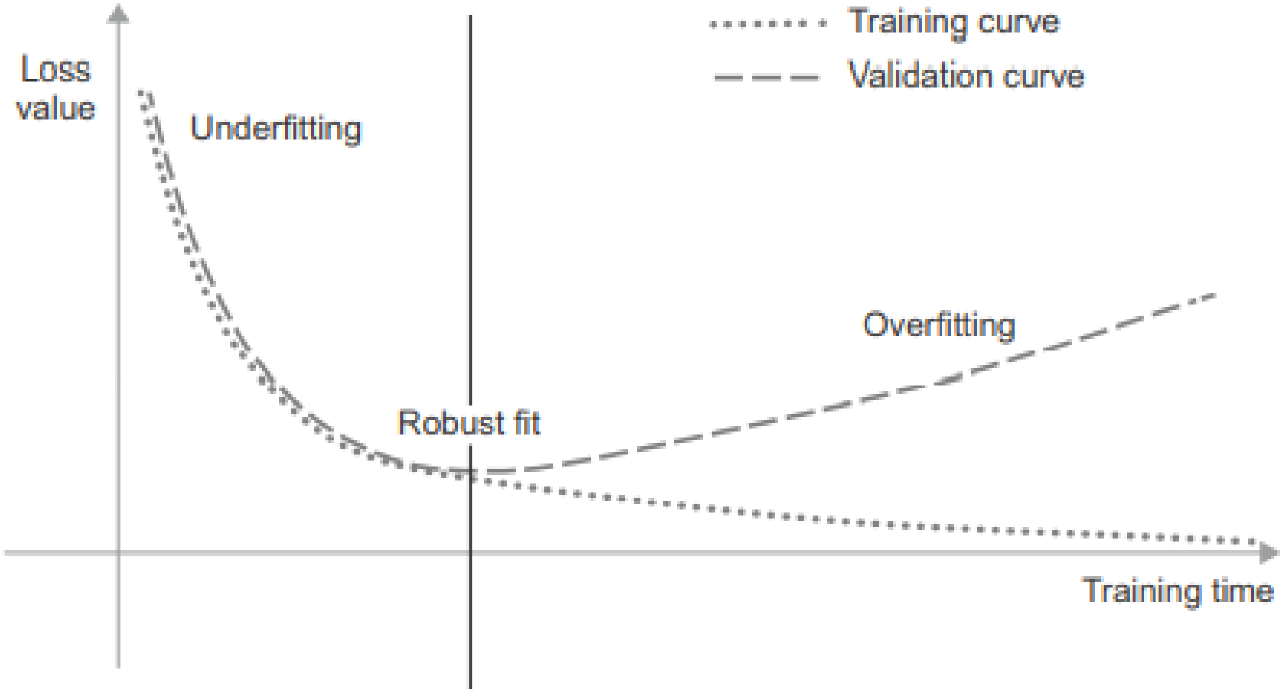
What about GenAI? Training vs Sampling



https://www.oreilly.com/library/view/generative-deep-learning/9781492041931/ch01.html

Optimization vs Generalization

- **<u>Optimization</u>:** fitting the data as best as possible during <u>training</u>
- **<u>Generalization</u>:** good model performance on new data during <u>inference</u>



DANIA ACADEMY

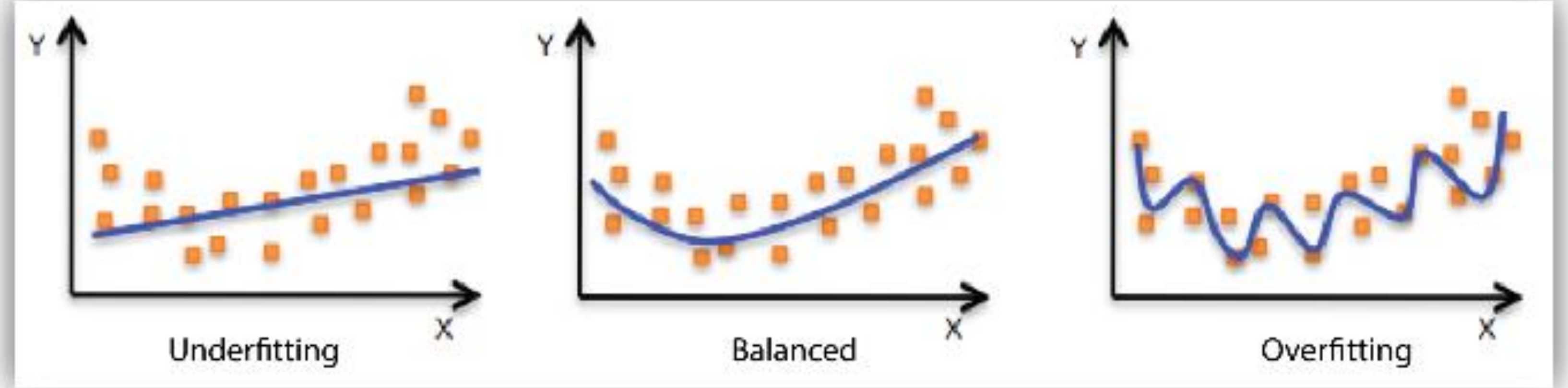
vives University of Applied Sciences

Underfitting vs Overfitting

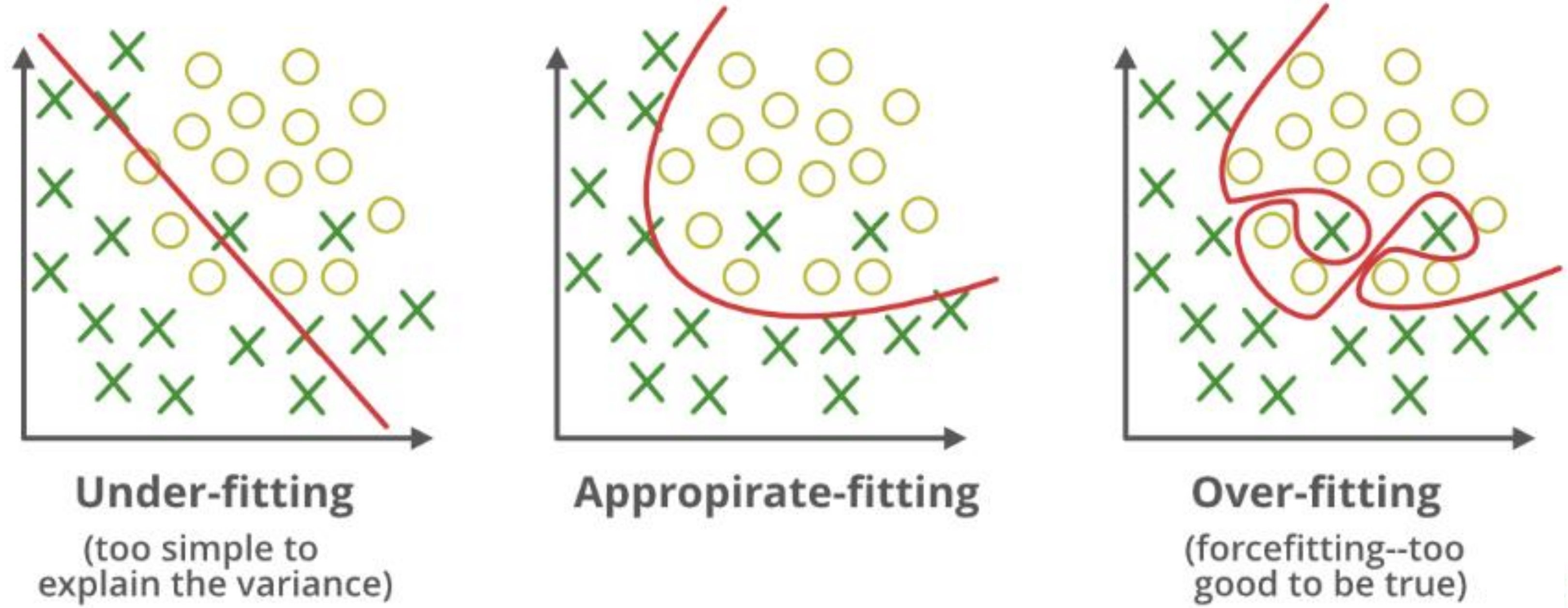
- **Underfitting:**
  model is too simple

- **Overfitting:**
  model is too specific

  Causes:
  - Noise
  - Uncertainty
  - Rare features
  - …



(source: https://towardsdatascience.com/underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6fe4a8a49dbf)

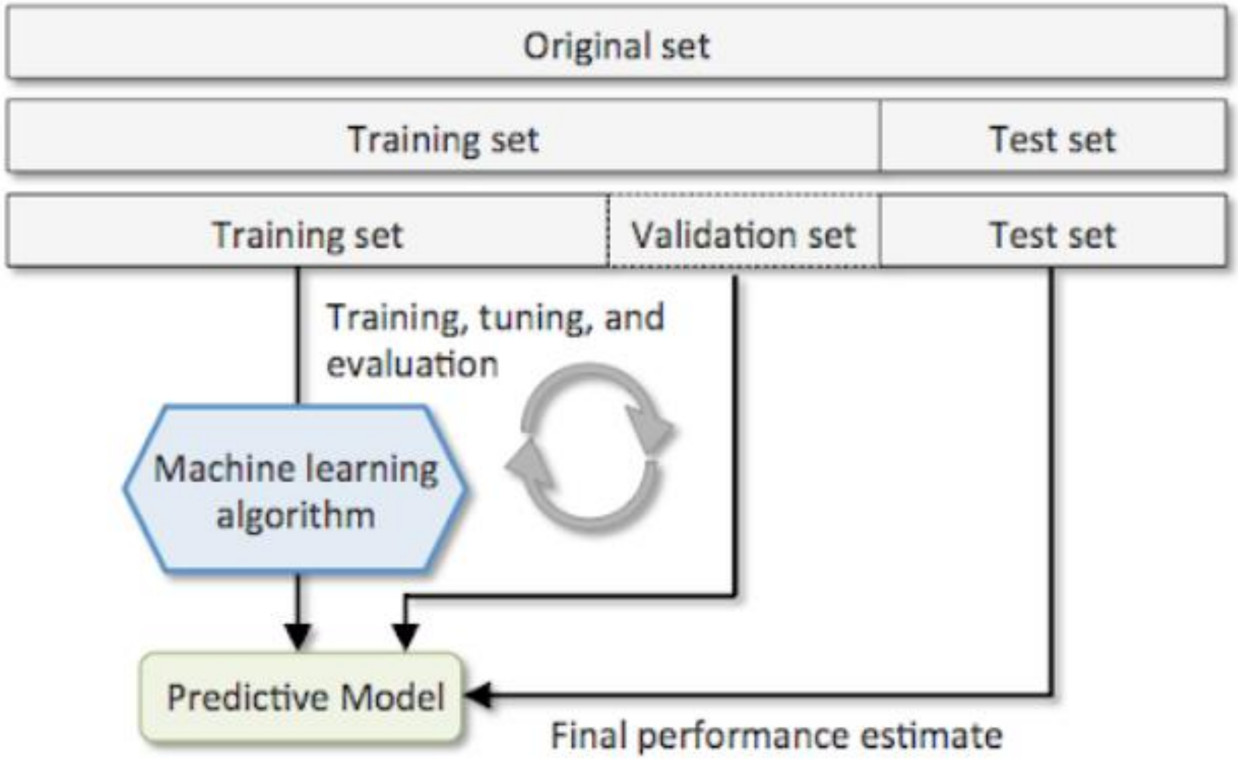


(source: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning)

Training, Validating and Testing

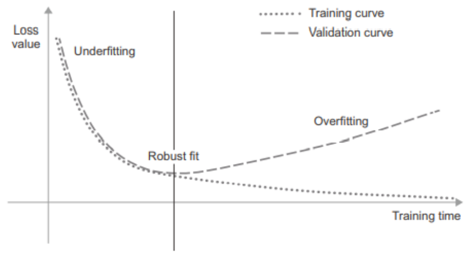
Splitting the dataset:

- **training**:
  - deriving the optimal model parameters
  - by the machine learning algorithm
- **validating**:
  - finding the optimal model configuration
  - fine-tuning the hyperparameters
  - to overcome overfitting
  - by the machine learning engineer
- **testing**:
  - final evaluation
  - by the machine learning engineer

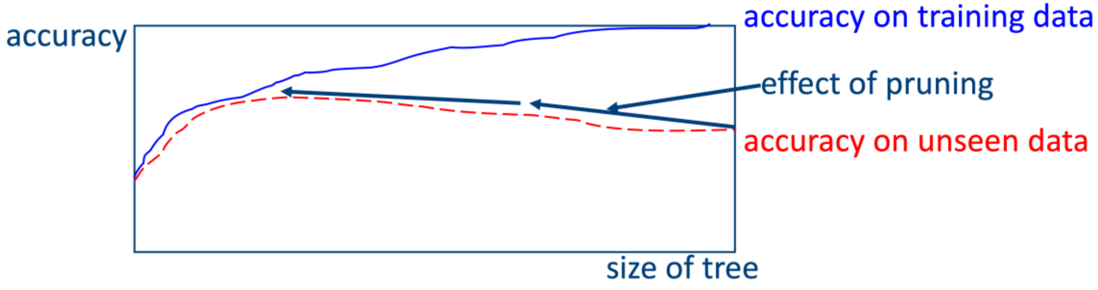


(source: https://vitalflux.com/hold-out-method-for-training-machine-learning-model)

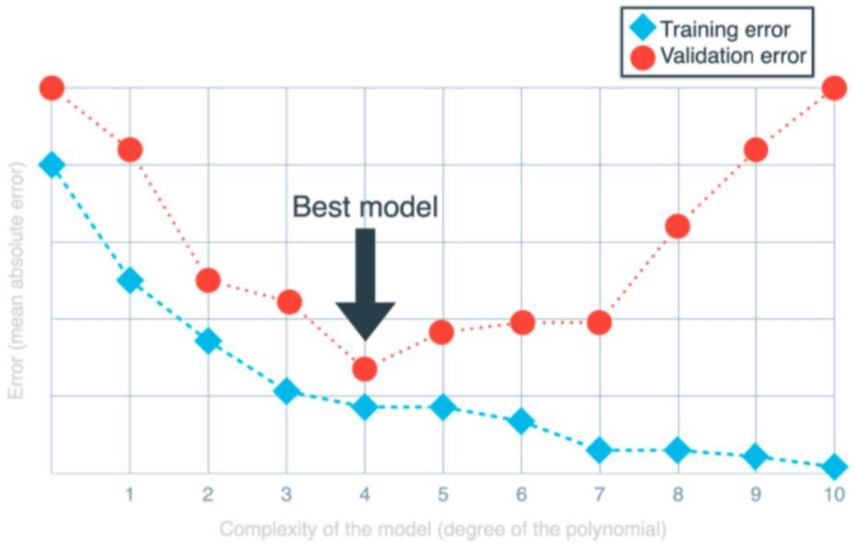
Training vs Validation Performance



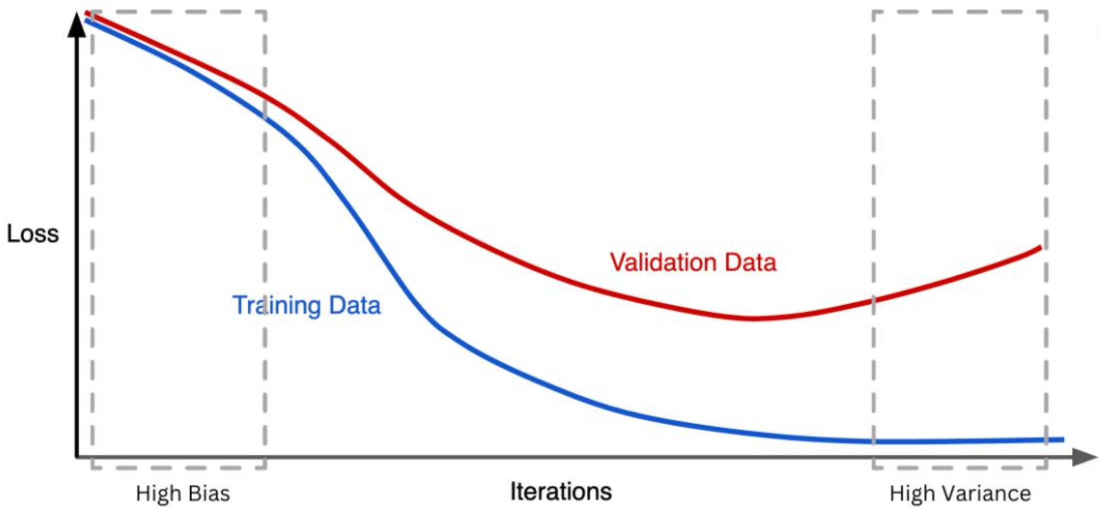
- **Decision Trees**



- **Polynomial Regression**



- **Artificial Neural Networks**



(source: https://www.dataquest.io/blog/regularization-in-machine-learning)

Model Evaluation

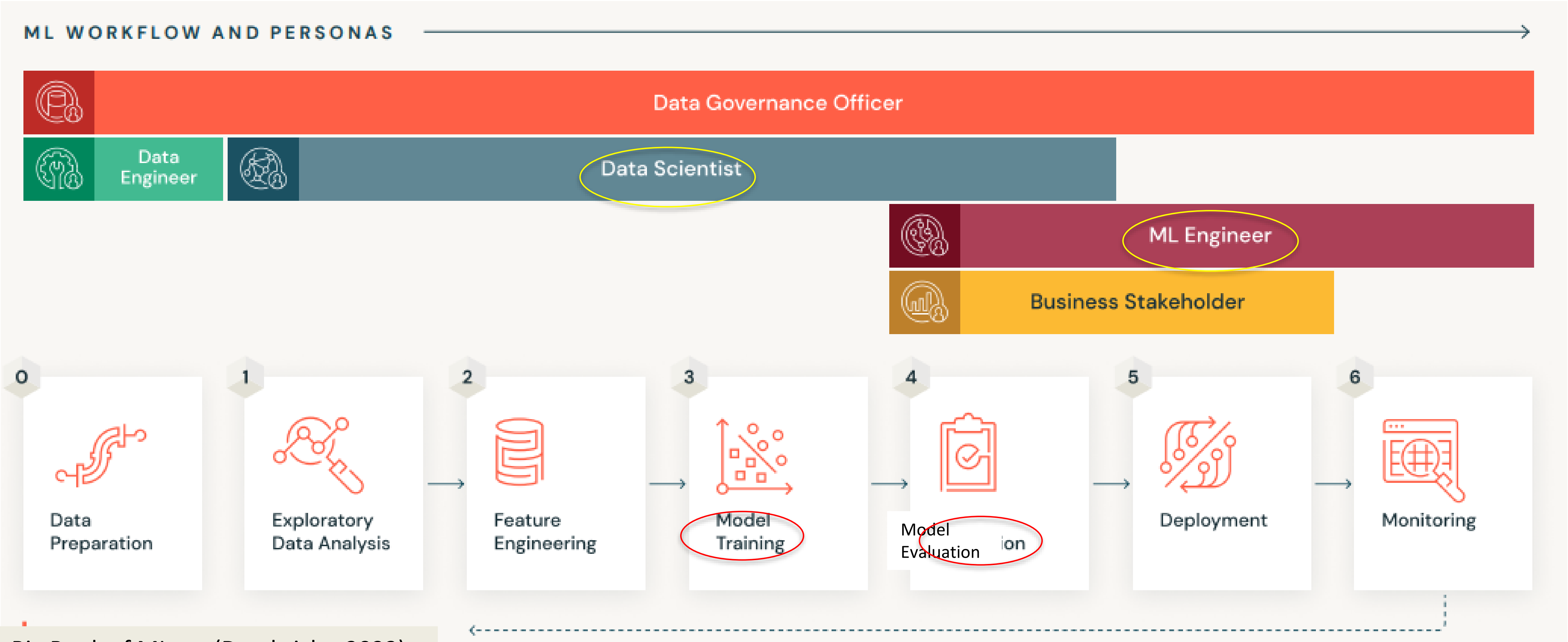
- **Loss functions:**
  - Compare predictions and true target values
  - Minimized by machine learning algorithms to obtain the best fit of the data
  - Should be mathematically convenient

- **Evaluation metrics:**
  - Also compare predictions and true target values
  - Used by machine learning engineers to evaluate the model performance
  - Easier to interpret by humans

Common Loss Functions and Metrics

Task	Loss	Metric
Regression	**Mean Squared Error (MSE)** = mean of the squared differences between predictions and targets	**Root Mean Squared Error (RMSE)** = square root of MSE
	Mean Absolute Error (MAE) = mean of the absolute differences between predictions and targets	**Coefficient of Determination (R^2)** = number between 0 and 1 expressing the goodness of fit where 1 indicates a perfect fit
Classification	**Cross-Entropy or Log Loss** quantifies the difference between the predicted probabilities and the true labels	**Accuracy** = the number of correct predictions divided by the total number of samples

Machine Learning Workflow



The Big Book of MLops (Databricks, 2023)

Sources

- Many slides are based on the book “Grokking Machine Learning” by Luis G Serrano (2021)
- Other slides are inspired by the book “Deep Learning with Python (2nd edition)” by François Chollet (2021)
- Some slides are adopted from the presentation on machine learning that was part of the course “Introduction to Artificial Intelligence” taught by Dr. Stefaan Haspeslagh at the Vives University of Applied Sciences during the academic year 2019-2020
- A few slides are taken from lectures given by Prof. Dr. Celine Vens and Prof. Dr. Hendrik Blockeel (Computer Sciences, KUL)
- Information was also obtained from Andrew Ng’s online course “AI for Everyone”: https://www.deeplearning.ai/courses/ai-for-everyone/
- Other sources are mentioned on the slides