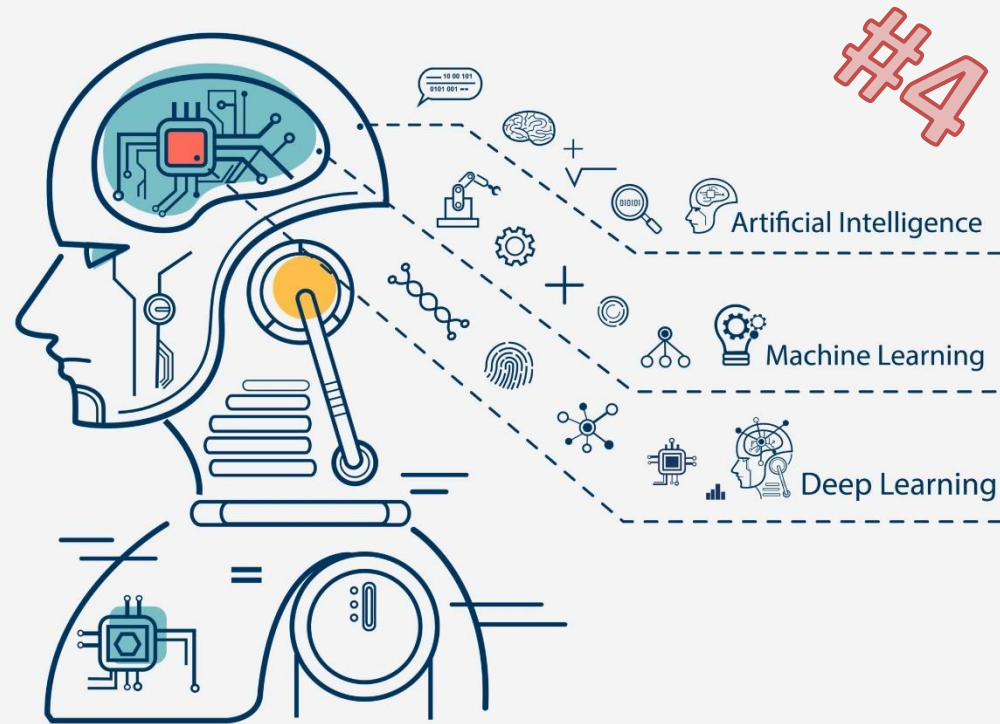




Machine Learning

Per Madsen

Faglærer

E-mail: pmad@techcollege.dk

Agenda

- Opfølgning på øvelser
 - #1 Indgang med Python
 - #2 Data cleaning and filtering

- Øvelse
 - #3 Visualisering af WordCloud

- Lidt teori om Machine Learning

- Øvelse (Start)
 - #4 Spam / Ham
 - #5 Spam / Ham / Målpinde

- Kort repetition
- Hjælp til Øvelse #4
 - Konkrete problemer??
 - TF-IDF Vectorizer
 - Confusion Matrix

Hvordan lærer mennesker?

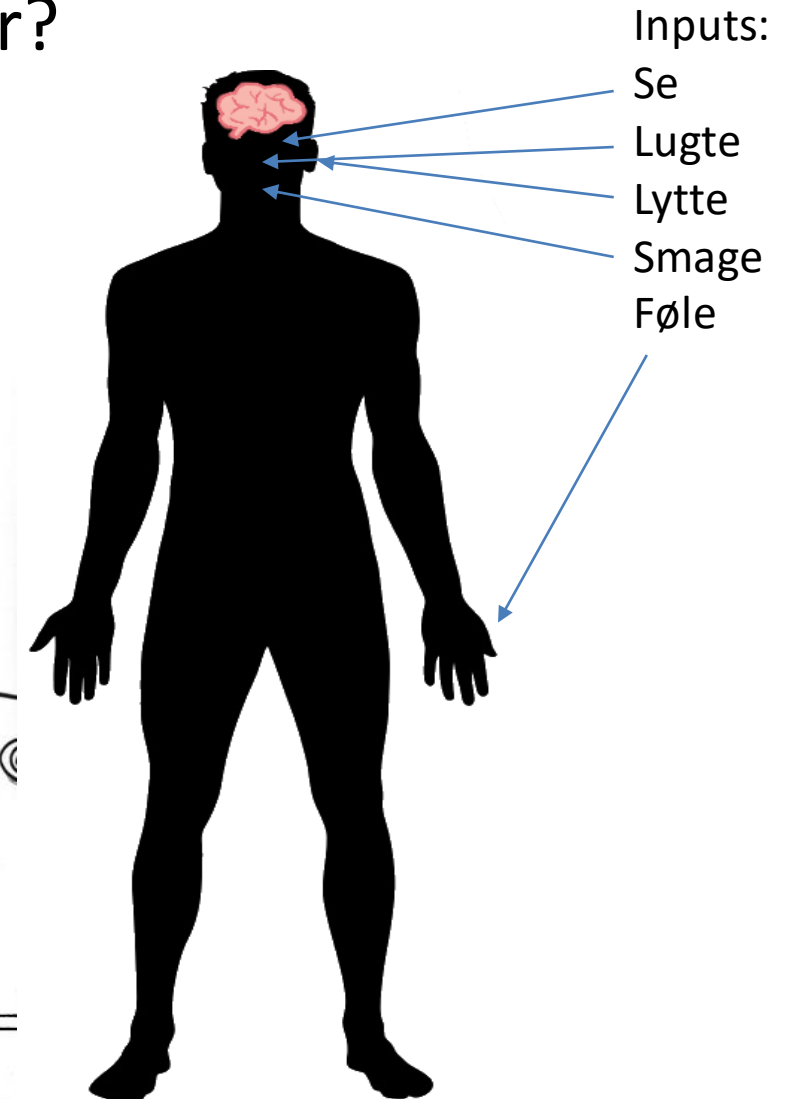
- Vi lærer baseret på vores sanser.
- Erfaring dannes ud fra oplevelser fra sanser.

Hvordan lærer vi af erfaring?

- Vi oplever gode og dårlige sanse inputs som former vores handlinger fremover.

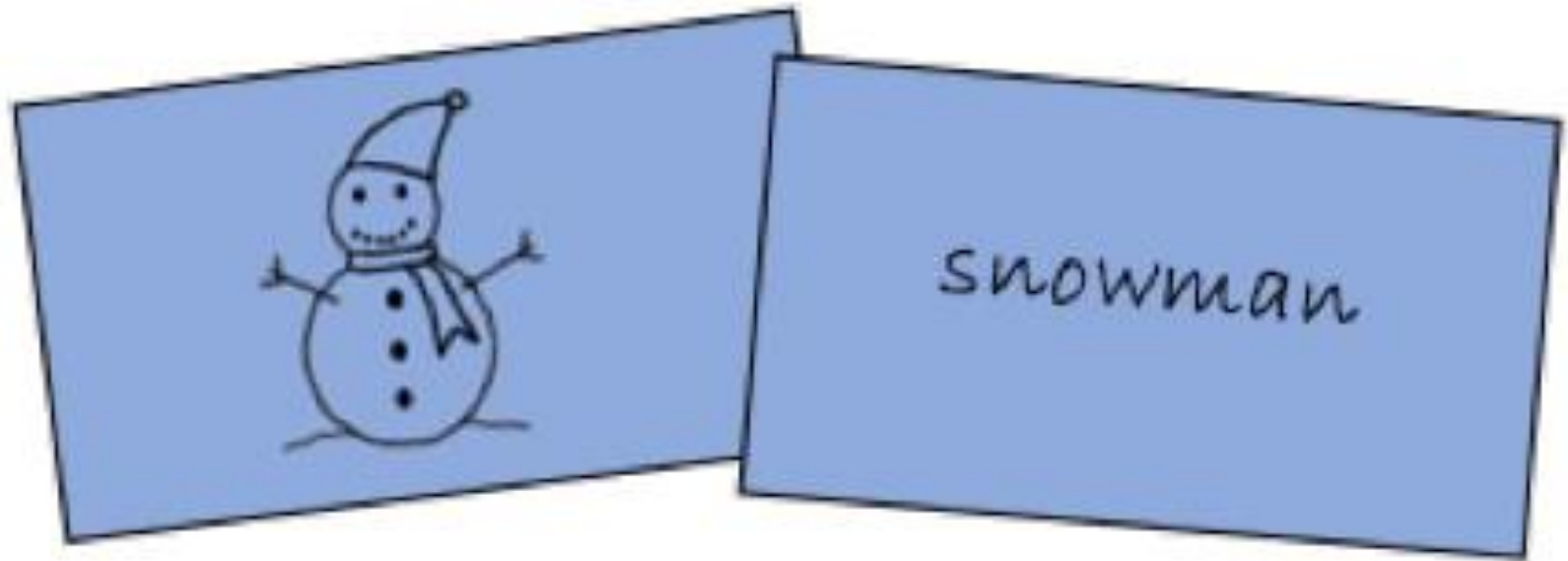
Hvordan lærer vi at læse?

- Vi ser mønstre i tekst og lærer at genkende mening bag ord og sætninger.



4 paradimer: Supervised

- Supervised
 - Når man træner modellen, ved at vise den svaret.



Flash card

4 paradimer: Unsupervised

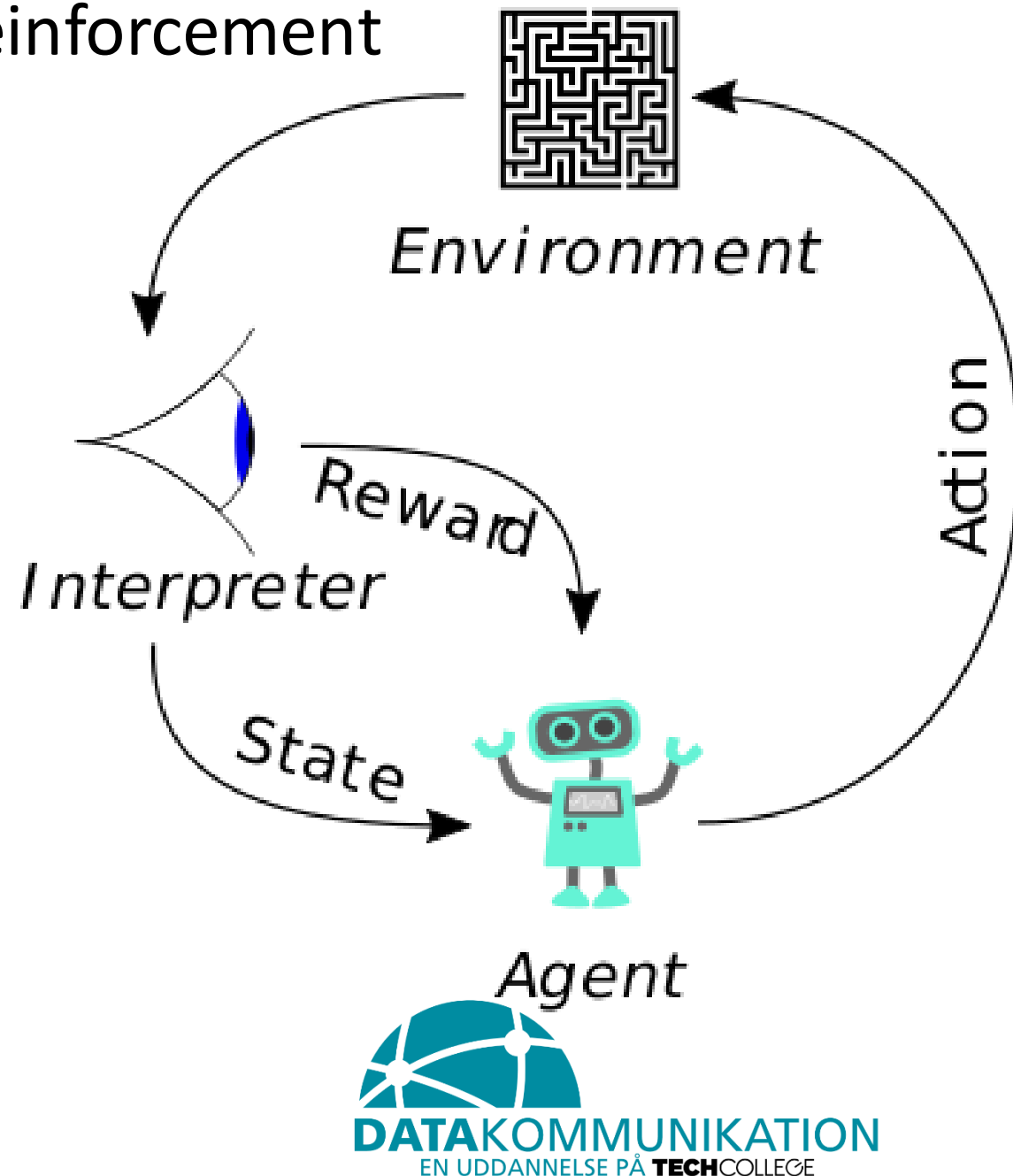
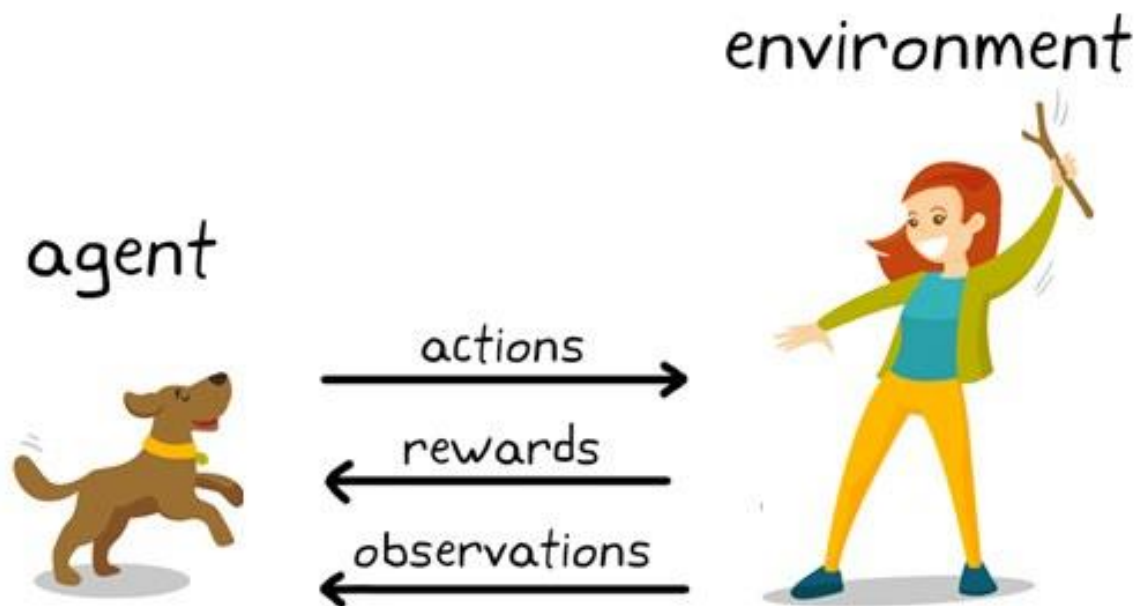
- Unsupervised
 - Når man træner modellen, hvor den ikke har svaret.

Repetition is the mother of all learning.
Repetition is the mother of all learning.
Repetition is the mother of all learning.
Repetition is the mother of all learning.
Repetition is the mother of all learning.
Repetition is the mother of all learning.
Repetition is the mother of all learning.

4 paradimer: Reinforcement

- Reinforcement

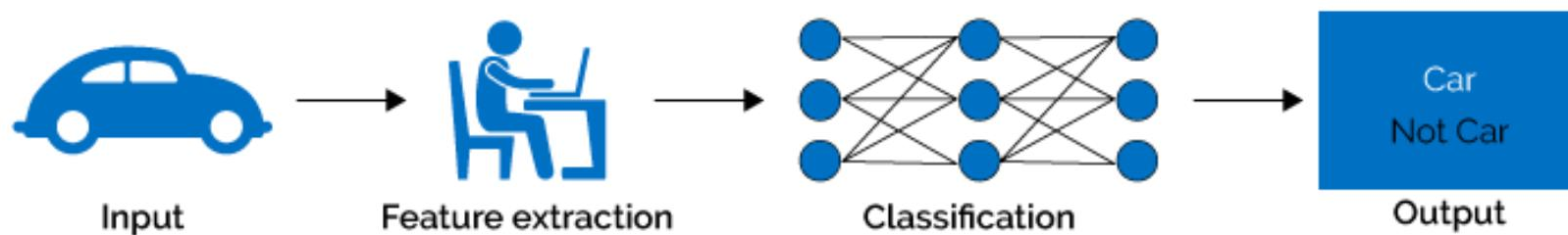
- Når man træner modellen, ved at sætte den i et miljø og giver den feedback baseret på dens handlinger



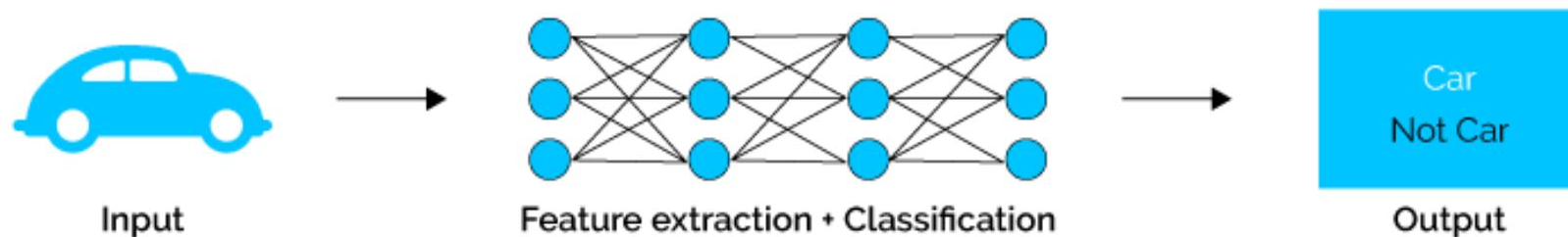
4 paradimer: Deep learning

- Deep learning algoritmer finder selv sammenhængen i dataen
- Det gør dem meget dårlige at bruge til at løse specifikke opgaver
- Der gode til at finde sammenhæng i kæmpe mængder af data, som vi ikke kan se en sammenhæng i.
- Neurale netværk: Layers, bias, weights, hidden layers

Machine Learning



Deep Learning



Learning by doing

Slide 8

- Øvelse #4: Spam / Ham
- Supervised Learning
- Tutorial med faldgruber
- Prøve at bruge **scikit-learn** uden at kende teorierne bag

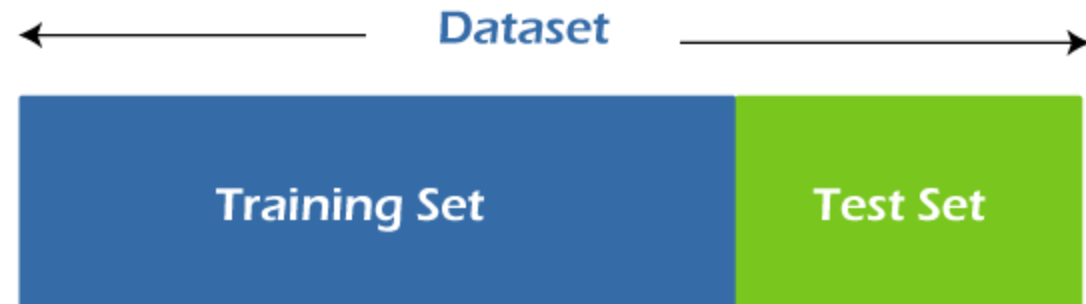
Data processing

- Import the required packages
- Loading the Dataset
- Remove the unwanted data columns
- Preprocessing and Exploring the Dataset
- Build word cloud to see which message is spam and which is not.
- Remove the stop words and punctuations
- Convert the text data into vectors

Building a sms spam classification model

Slide 10

- Split the data into train and test sets
- Use Sklearn built-in classifiers to build the models
- Train the data on the model
- Evaluate model(s)
- Make predictions on new data



Adobe Stock | #72757953

Term Frequency - Inverse Document Frequency (TF-IDF)

Term frequency [\[edit \]](#)

Term frequency, $\text{tf}(t, d)$, is the relative frequency of term t within document d ,

$$\text{tf}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

Inverse document frequency [\[edit \]](#)

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

- Fra lister af ord til en vector som kan bruges af vores algoritme
- Søg på TD-IDF for en fuld forklaring

Confusion Matrix

Slide 12

		Actual Values	
		1	0
Predicted Values	1		
	0		

GO,
Again!

Plan fra nu ..

- Opsamling: Spam / Ham / Goal
- Øvelse #5:
 - Billede genkendelse med et neuralt netværk
 - Datasæt og passende algoritme
- Case opgave
 - Find selv passende datasæt / algoritme
 - Det er ok at blive inspireret
 - Så længe man er ærlig 😊
 - Skal afleveres med minimal beskrivelse
 - Gruppevis eller solo
 - Fokus på at arbejde eksperimentielt

Out[1]: <matplotlib.image.AxesImage at 0xb3ded33c8>

