

# Introduction to AI, machine learning and neural networks

Dr. Stefan Leijnen

Dr. Sieuwert van Otterloo

July 12-16 2021

<https://utrechtsummerschool.nl/courses/engineering-and-technology/introduction-to-artificial-intelligence-machine-learning-and-neural-networks>

# About the summer school team



This summer school is hosted by an all star team of the artificial intelligence research group of Utrecht University of Applied Sciences. Unlike more theoretical research groups, our research is not focused on more AI. Instead we focus on responsible AI, including:

- The role of AI in Media and financial sector
- Ethical AI
- AI and creativity
- Validation of AI systems

The research group supports all bachelor and master courses at the HU and is open to collaborations, thesis projects and internships

# About Sieuwert van Otterloo



**Sieuwert van Otterloo**  
Researcher

---

Research group  
[Artificial Intelligence](#)

E-mail  
[sieuwert.vanotterloo@hu.nl](mailto:sieuwert.vanotterloo@hu.nl)

Phone number  
+31 6 10 509 674

Follow Sieuwert

*Main lecturer, present all days*

## Background:

- Graduated in mathematics and Artificial Intelligence
- Doctorate from University of Liverpool in classical AI (logic and multi agent systems)
- IT auditor, privacy officer and expert for IT legal affairs
- Investor in several Dutch startups
- Lecturer (Vrije Universiteit) in software project management and Skills for AI.

## Notable research projects:

- Measuring explainability of AI
- Project risk management for crowdfunded and ICO-funded projects
- Testing user acceptance of AI related risks (autonomous vehicles, AI and recruitment)

More info on: [www.ictinstitute.nl](http://www.ictinstitute.nl)

# About Stefan Leijnen



**Stefan Leijnen**  
Professor

---

Research group  
[Artificial Intelligence](#)

E-mail  
[stefan.leijnen@hu.nl](mailto:stefan.leijnen@hu.nl)

Follow Stefan

## Background:

- Doctorate from University of Nijmegen on AI and creativity
- Former CTO for War Child

## Notable research projects:

- AI impact assessment
- AI in media sector field labs
- Ethics Inc – a game about responsible AI

# About Stan Meyberg



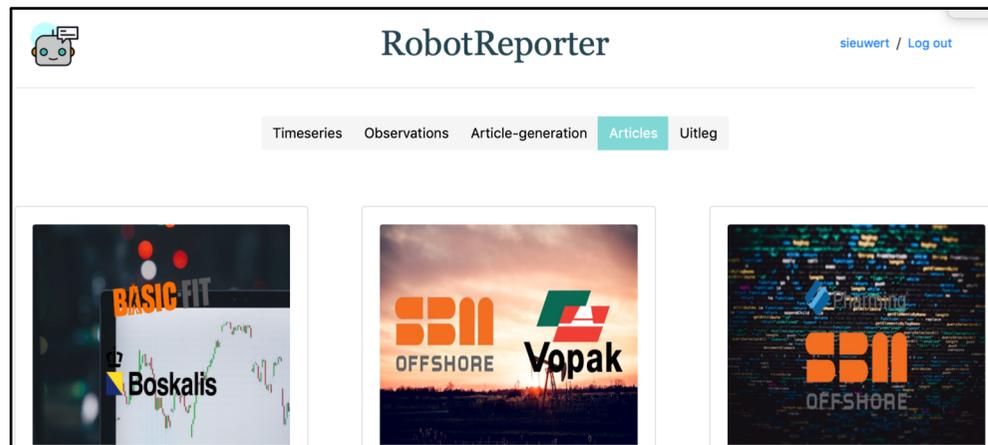
**Stan Meyberg**

Background:

- Bachelor student Applied Artificial Intelligence

Notable research projects:

- Robotreporter, an AI system for generating news articles



# Structure of each day

Time	Content	Remarks
8.45-9.05*	Walk-in and coffee	
9.05 – 9.30*	Recap and questions	Discuss previous day. On day 1: check if people have practical questions
9.30 – 10.30	Theory	Presentation by lecturer of key concepts
10.30 - 10.45	Coffee break	
10.30 – 11.45	Practical session	Working on assignments, individual or in groups
11.45 – 12.15	Discuss practice results, conclusion	
12.15 – 13.15	Lunch	
13.15 – 14.30	Theory	Presentation by lecturer of key concepts
14.30 – 14.45	Coffee break	
14.45 – 15.45	Practical session	Working on assignments, individual or in groups
15.45 – 16.15	Discuss practice results, conclusion	
16.15 – 16.30	Time for individual questions	Lecturer is available for individual questions

\* Day one will start later at 9.30

# Agenda

<b>Monday Jul12: Data science</b>	<b>Tue Jul13: machine learning</b>	<b>Wed Jul14: Standard neural networks</b>	<b>Thu Jul15: complicated neural networks</b>	<b>Fri Jul16: other AI algorithms</b>
Data exploration and visualisation	Decision trees and regression	Prediction with neural networks	Image recognition	Evolutionary algorithms
History of AI	AI and ethics (Ethics Inc)	AI validation / medical AI	Neural network types	Business process mining

# Monday Jul12: data science

Morning theory

Morning practical

Afternoon theory

Afternoon practical

## Programme:

- Your expectations for this week
- Data science basics
- Exploring data sets
  
- Exploring data sets with python
  
- History of AI
- Classic AI methods
  
- Understanding deepfakes

# Your expectations

Our 'backlog' for this week consists of your questions. Please list one learning goal and one or two most important questions. We will ask each lecturer to use their expertise to answer a few questions.

Why would also like to know what you already know and can help other people learn in this week. Our goal is to learn as much from you as you from us!

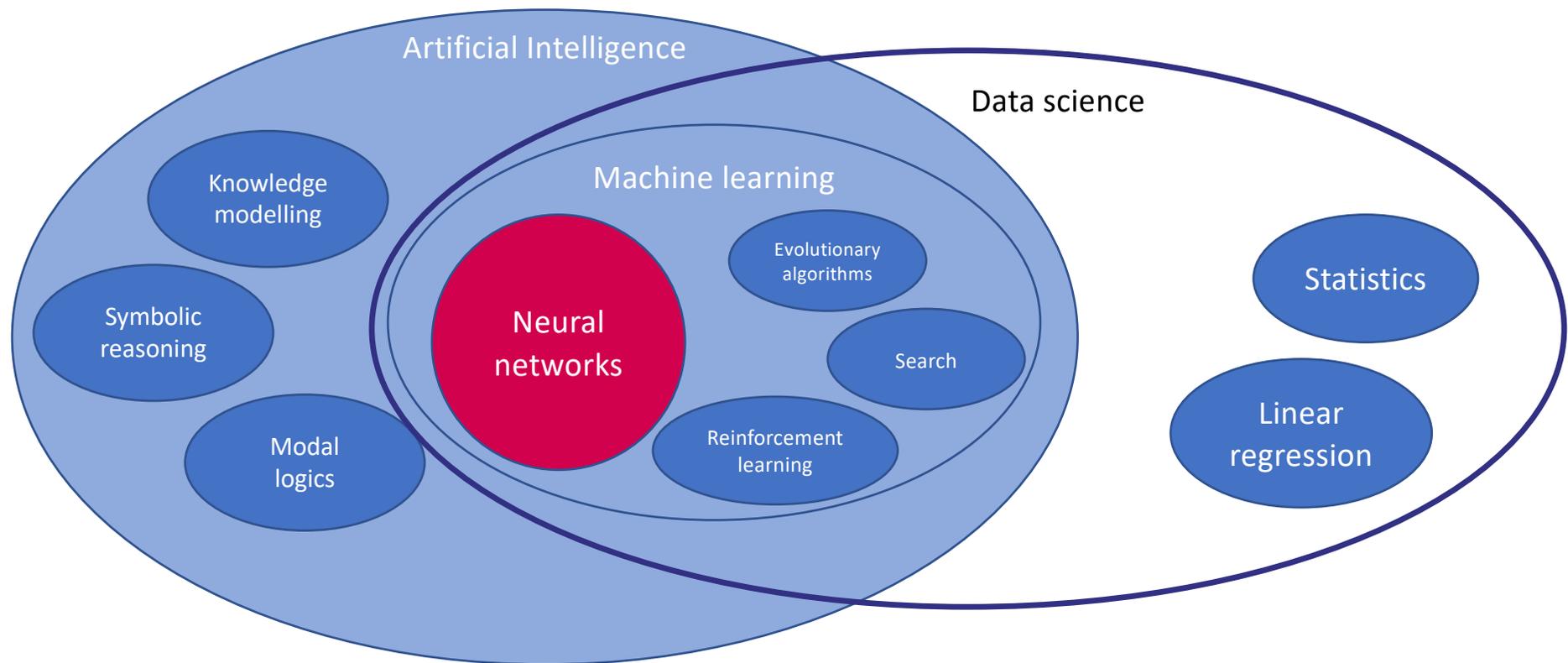
Todo



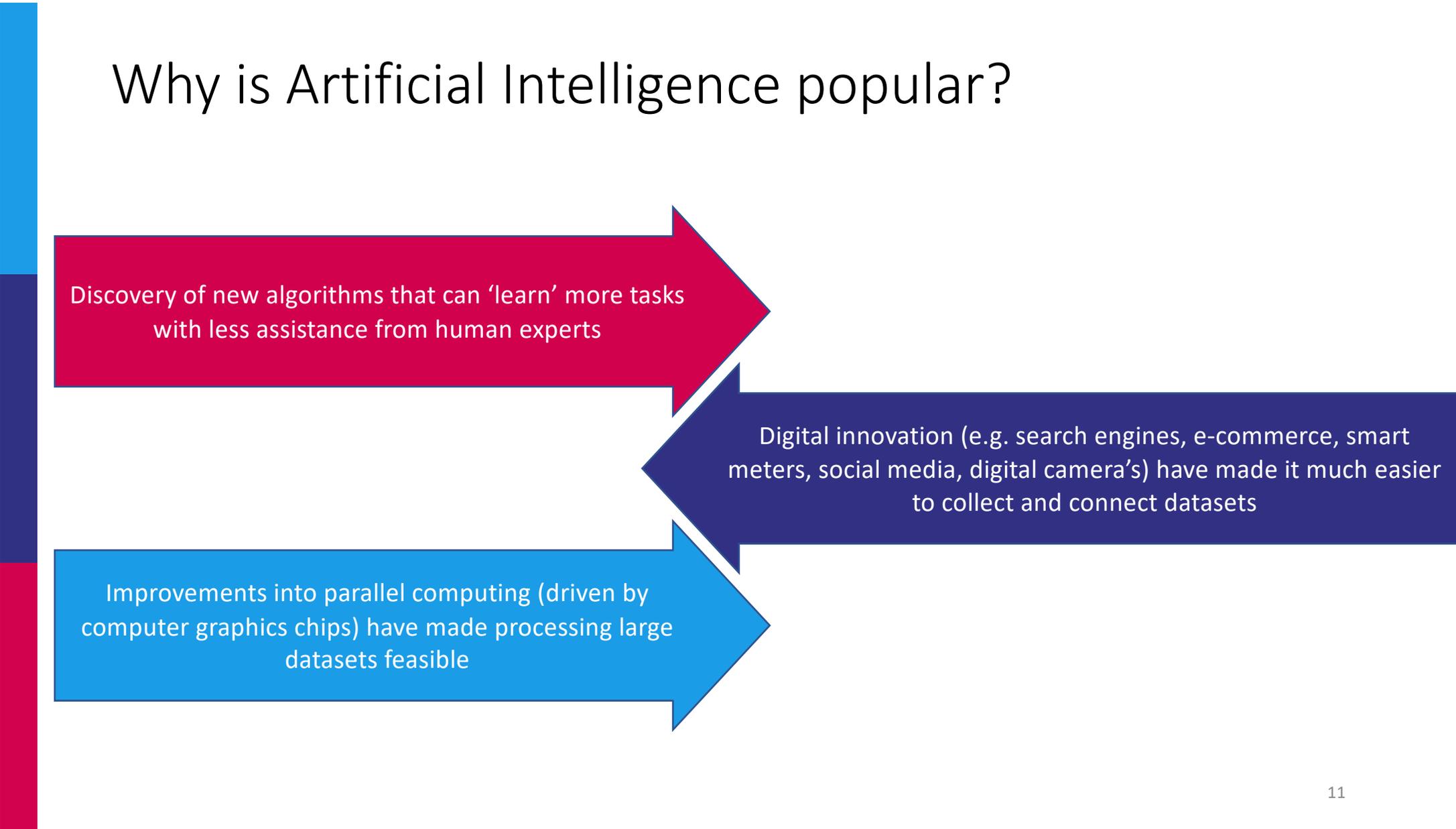
Doing

Done

# What is AI / machine learning?



# Why is Artificial Intelligence popular?



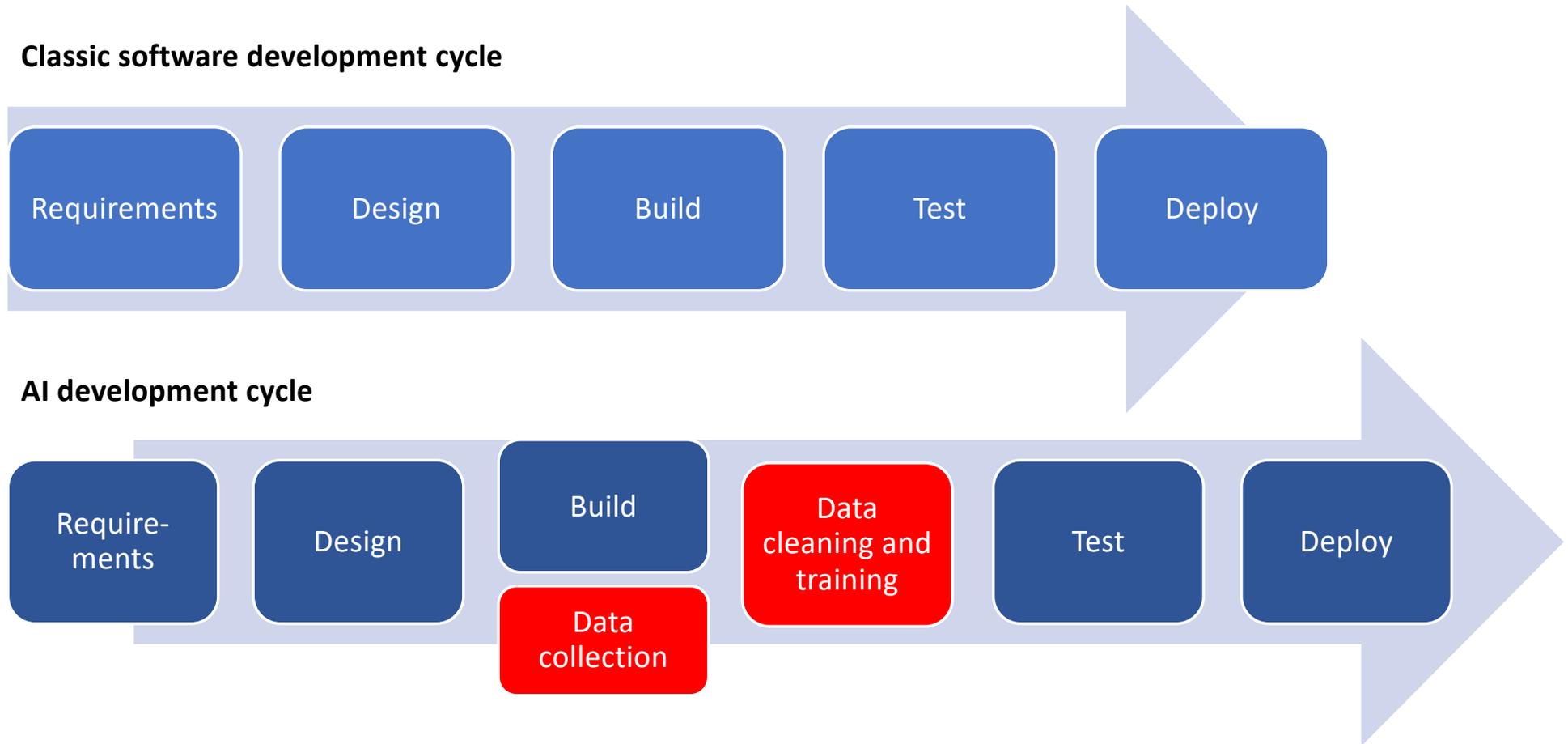
Discovery of new algorithms that can 'learn' more tasks with less assistance from human experts

Digital innovation (e.g. search engines, e-commerce, smart meters, social media, digital camera's) have made it much easier to collect and connect datasets

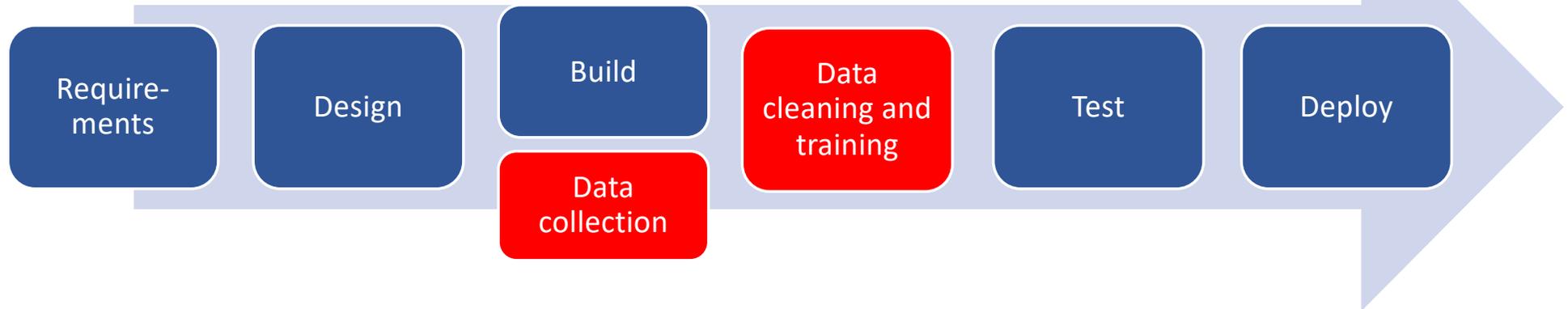
Improvements into parallel computing (driven by computer graphics chips) have made processing large datasets feasible

# AI will change the way AI systems are developed

## Classic software development cycle



## AI development cycle

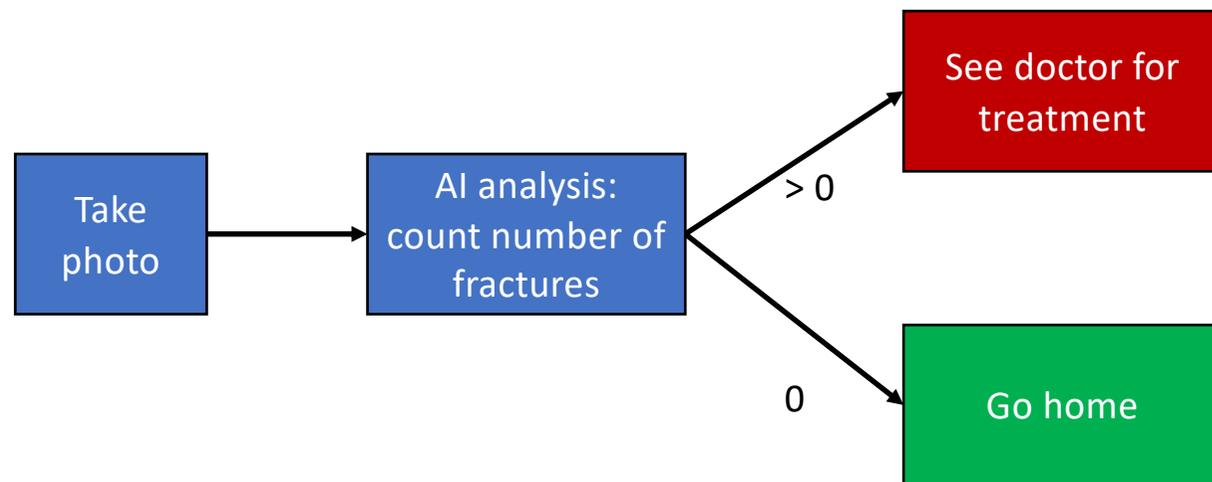


# AI in medicine



Img: Owen Beard via Unsplash

AI can be used to interpret medical image, for instance to count the number of fractures in a medical image.

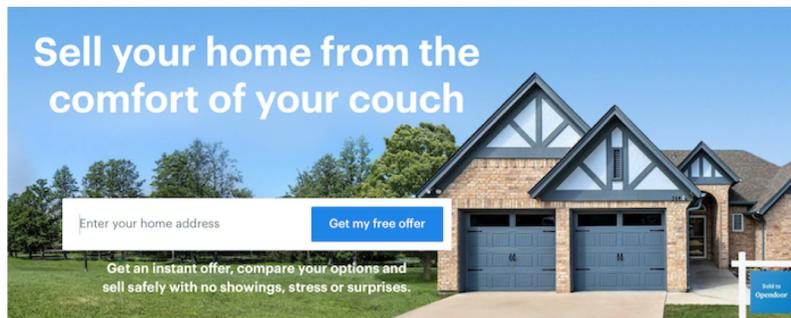


# AI in entertainment



- <https://www.youtube.com/watch?v=dxb2cvk246k>

# AI in finance - OpenDoor



## How it works



Many AI researchers and data scientists study house prices. Only a few people trust their own models enough to actually buy houses based on predicted values. OpenDoor (USA based) is such a company.

AI Explainability is a key concern. That is why many models are based on determining the value of house by adding the value of each feature

- Enter “instant buyers” such as Opendoor. These companies, known by the abbreviation “i-buyers”, try to do to property sales what Billy Beane did to baseball. Just as the manager of the Oakland A’s substituted software for conventional talent scouts, i-buyers replace estate agents with algorithms that crunch data on everything from the number of bedrooms to local crime rates, to estimate what a property should sell for. They then buy it at a discount to the computed price (as Mr Beane did with players), spruce it up and offload it. Opendoor says its average fee is 6-6.5%, about the same
- cut as conventional estate agents take on a sale.
- To be sure, there is a way to go before the promise of i- buyers is fulfilled. Their algorithms are good at appraising identikit single-family units, but struggle with idiosyncratic properties—flats in city centres, say, or luxury villas.
- <https://www.economist.com/business/2018/09/13/tech-firms-disrupt-the-property-market>

# House valuation model – simple matrix algebra



**Amsteldijk Zuid 195**  
 1188 VP Amstelveen  
 € 2.595.000 k.k.  
 350 m<sup>2</sup> / 1.445 m<sup>2</sup> • 7 kamers

Brockhoff Makelaars



**Bovenkerkerweg 79 B**  
 1187 XC Amstelveen  
 € 1.195.000 k.k.  
 198 m<sup>2</sup> / 397 m<sup>2</sup> • 5 kamers

Voorma & Millenaar makelaars - taxateurs o.g.



**Oostermeerweg 41**  
 1184 TS Amstelveen  
 € 1.450.000 k.k.  
 214 m<sup>2</sup> / 615 m<sup>2</sup> • 5 kamers

Voorma & Millenaar makelaars - taxateurs o.g.



- Linear algebra allows us to solve this

350	1445	2595	
198	397	1195	
1.000	4.129	7.414	rescale
1.000	2.005	6.035	rescale
1.000	4.129	7.414	
0.000	2.124	1.379	subtract
1.000	4.129	7.414	
0.000	1.000	0.649	rescale
1.000	0.000	4.733	subtract
0.000	1.000	0.649	

$$\text{Asking-price} = 4.733 * \text{floorspace} + 0.649 * \text{plotsize}$$

$$\text{Askingprice}_3 = ??$$

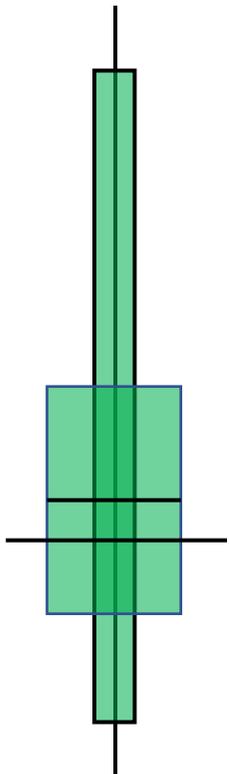
$$\text{Askingprice}_3 = 4.733 * 214 + 0.694 * 615 = 1412$$

Floorspace (m2)	Plot size (m2)	Asking price (€ x 1000)
350	1445	2595
198	397	1195
214	615	1450

# Example dataset

Column	Description
Id	Unique number for each house
Zipcode	4 digits that are used for mail delivery.
Lot-len	Length of the plot of land the house is on
Lot-width	Width of the plot of land the house is on
Lot-area	Area of the land the house is on in square meters. Defined as length x width
House-area	Floor area of the inside of the house in square meters
Garden-size	Size of the back garden in square meters
Balcony	Number of balconies the house has. Typically zero or one
X-coor	Horizontal position on the city map
Y-coor	Vertical position on the city map
Buildyear	Year the house was built
Bathrooms	Number of bathrooms in the house
taxvalue	Estimated value of the house as used by the city council for tax purposes
retailvalue	Estimated value that the house can be sold for if brought to the market

# Data exploration for each column



Before using any data column, you should understand:

- Minimum and maximum value ('range')
- Average and variation
- Most common values
- Shape of the distribution

Using AI before understanding the data will lead to wrong, impossible or dangerous predictions. You could even crash an AI system if you do not understand the valid values.

Many IT accidents are caused by confusing units, e.g. grams / kilograms. Square meters / foot, degrees / rad. Exploration will help you understand results and avoid mistakes.

# Average / median / modus

Number of bathrooms
11
3
3
1
1
1
0
0
0
0

Average : 2.0

*Sum divided by  
number of inputs*

Median : 1.0

*Middle element when  
sorted on value*

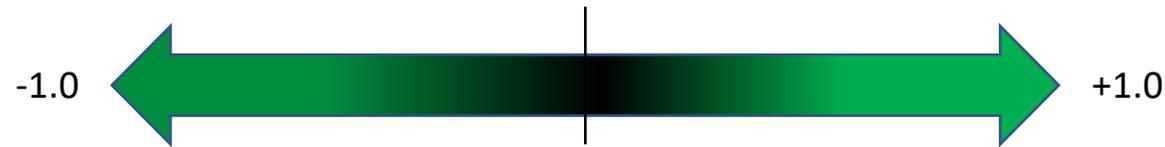
Modus: 0

*Most common  
element*

These three values coincide in a 'normal distribution'. In practice a lot of data is skewed and these values differ.

- The average value often does not occur and is not a good example value
- Sometimes cutoffs occur: values higher than 999 are replaced by 999
- Sometimes fields are left empty. Some computer programs takes this as meaning 0.0
- Sometimes data is inconsistent: len x width not equal to area
- Some variable refuse to vary: all datapoints have the value

# Correlation



Correlation expresses whether the values of two variables are related. If variables are correlated, one of the values can be used to predict the other.

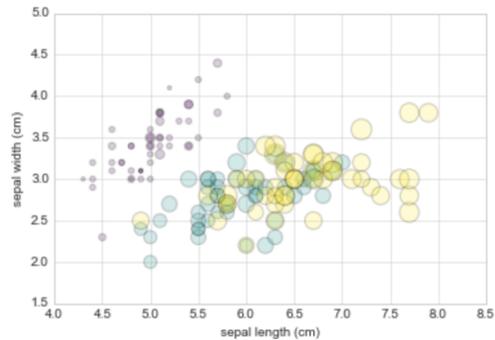
Correlation can be zero, positive or negative:

- Negative correlation means they move in opposite direction: e.g. house price and distance to city center
- Positive correlation means they move in the same direction: e.g. house-size and tax value
- Near-zero correlation means that there is no linear relation. There can be other relations

# Scatter plot

```
In [8]: from sklearn.datasets import load_iris
iris = load_iris()
features = iris.data.T

plt.scatter(features[0], features[1], alpha=0.2,
            s=100*features[3], c=iris.target, cmap='viridis')
plt.xlabel(iris.feature_names[0])
plt.ylabel(iris.feature_names[1]);
```



Scatterplots help you to see distribution and relation between two variables at once.

For small datasets, you can display all points at once. For large datasets, you can sample or use transparency.

It is possible to show 3, 4 or 5 variables at the same time using circle size, color or shape.

As explainability is more and more becoming a requirement for AI systems, delivering plots with any algorithm will become a requirement.

Example from: <https://jakevdp.github.io/PythonDataScienceHandbook/04.02-simple-scatter-plots.html>

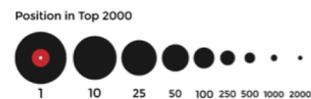
# Example scatter plot

## TOP 2000 ❤️ 70's & 80's

Since 1999 the 2000 most popular songs of all time, as voted by the show's audience, are played on Dutch national Radio 2 in a yearly marathon. The 2000 songs are on the air between noon on December 25th until New Year's Eve and over half of the Dutch population listens to the Top 2000 each year.

Each ● to the right represents a song in the Top 2000. It is placed according to its year of release. In the legend below you can see what the size and color of a song means.

The bulk of the songs and most of the top 10 are from the 70's & 80's...



### Golden oldie

The oldest song in the list, Billie Holiday's *Strange Fruit*, is from 1939. It's 17 years older than the second-oldest song. If it will make the 2017 edition remains to be seen, it's barely in now, on position 1989.

Year of release

### Newly discovered

Although already released in 1972, *Starman* from David Bowie is the highest new song in the list. It never appeared in the previous 17 editions of the Top 2000 and entered in 2016 on position 270.

### Prince

Another legend who passed away in 2016 (on April 21st). It seems that new people discovered his works, with all 9 songs that were in 2015's list rising significantly and 8 more songs joining in 2016.

### 1 Bohemian Rhapsody

Queen: 1975

### 3 Stairway to Heaven

Led Zeppelin: 1971

### The Beatles

No other artist or band has more songs in the Top 2000 as the Beatles. With 38 songs they are responsible for 1.4% of all titles before 1970. Nonetheless, only 5 years ago they still had 50 songs in the list.

### 4 Piano Man

Billy Joel: 1974

### 5 Child in Time

Deep Purple: 1972

### 6 Avond

Douloewijn de Groot: 1997

### 10 Black

Ruff Am: 1992

### 2 Hotel California

Eagles: 1977

### 7 Heroes

David Bowie: 1977

### 9 Wish you were here

Pink Floyd: 1975

### High riser

Adèle's *When we were young* from 2015 apparently needed some time to become fully appreciated. It is the song with the highest increase in the list, shooting 1599 places from position 1743 to 144.

### 2016's most popular

The swinging new song from Justin Timberlake, *Can't stop the feeling*, is the highest newcoming song that was released in 2016. It is part of the soundtrack of the animated movie *Trolls*.

### 8 Mag ik dan bij jou

Claudia de Breij: 2011

### Pokémon

Already in the list in 2015 due to a social media campaign, nobody can deny the impact that Pokémon had on many people's daily lives in 2016. *Colts catch 'em all* by Jason Paige rises 1434 spots to position 232!

### David Bowie

Passing away only days after the release of his new album *Blackstar* on January 10th 2016. His legend remains strong with 26 songs in the Top 2000. His most popular song *Heroes* jumps from 34 to position 7.

<https://top2000.visualcinnamon.com/img/Top2000PosterEnglish.png>

# Installing python and Jupyter

1. Install python (<https://www.python.org/downloads/>) and the Jupyter toolbox:
2. Download the data set and python notebook at <https://github.com/swzaken/cars-neuralnetwork>
3. Install python packages. You can use the following commands

Library	Description	Command to install
Updated version of Pip	Installing packages	<code>python -m pip install --upgrade pip</code>
Numpy	Arrays and numbers	<code>pip install numpy</code>
MatPlotLib	Data visualization	<code>pip install matplotlib</code>
Pillow	Image processing	<code>pip install pillow</code>
JuPyter Notebooks	Running the code	<code>pip install jupyterlab</code>
Tensorflow	Machine learning algorithms	<code>pip install tensorflow</code>

# Practical session question A

1. Show the distribution of the x-y coordinates
2. Show the distribution of the taxvalues
3. Show the distribution of the retailvalues
4. What is the average value for each zipcode?
5. What is the average value of houses with housesize $>100$ ? What is the average value of houses with housesize $<100$ ?

# Practical session question B

## Correlation:

- What is the correlation between lot-area and retailvalue?
- What is the correlation between house-area and retailvalue?
- What is the correlation between lot-width and lot-length?
- What is the correlation between bathrooms and retailvalue?



## Practical session question C

Can you make scatterplots for the following combinations, and argue whether you see a relation, and if this is linear correlation:

- Xcoor and retailvalue
- Different color for each postal code?

Can you think of other combinations that provide insight into retailvalue?



## Practical session question D

Can you make a scatterplot with colored circles with the following features:

- Xcoor and ycoor for location of each point
- Size of the circle corresponds to retailvalue
- Different color for each postal code?

Can you think of other combinations that provide insight into retailvalue?



## Bonus question E

Repeat all the work, but only on data from one zipcode.

- Do you get more structured charts?
- Would it be a good strategy for this dataset to treat each zipcode separately? What would be the pros and cons of this strategy?