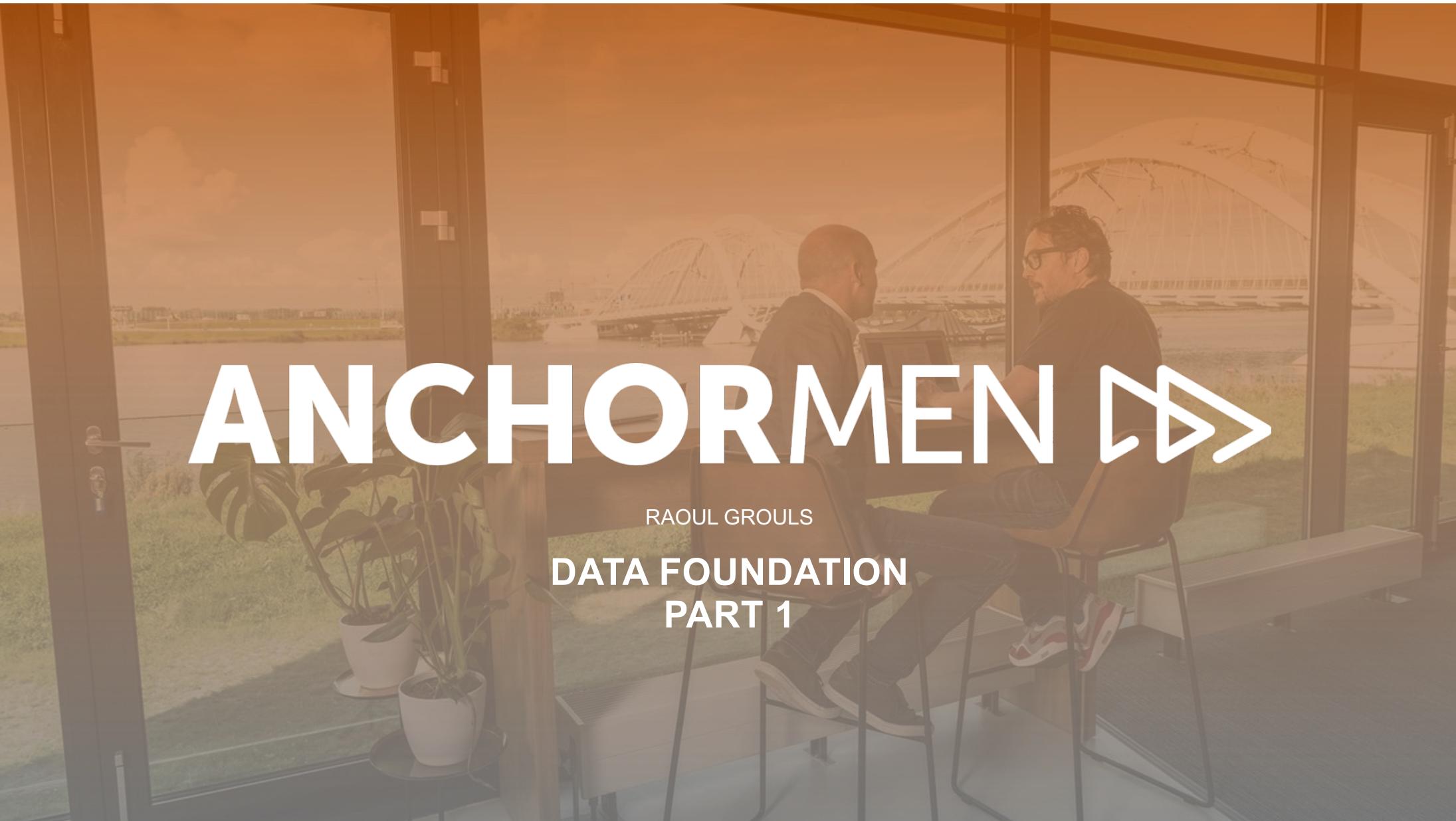


ANCHORMEN

RAOUL GROULS

DATA FOUNDATION
PART 1



WELKOM



Raoul Grouls

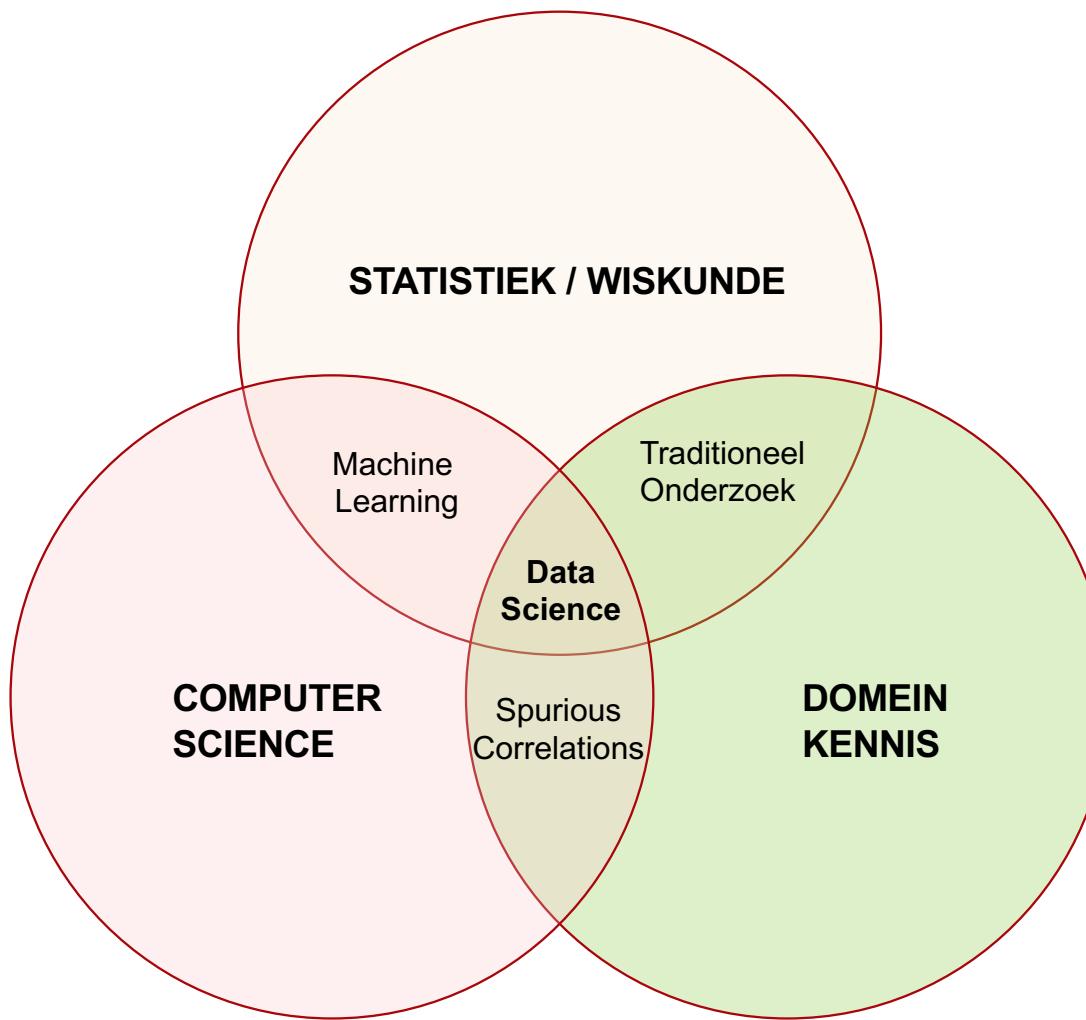
Data Scientist at Anchormen



INTRO



DATA SCIENCE



WAT IS A.I.



	Menselijk	Rationeel
Denken	Cognitieve wetenschap (modelleren)	Symbolische KI (logica)
Doen	Turing test	Maximaliseren van waarde

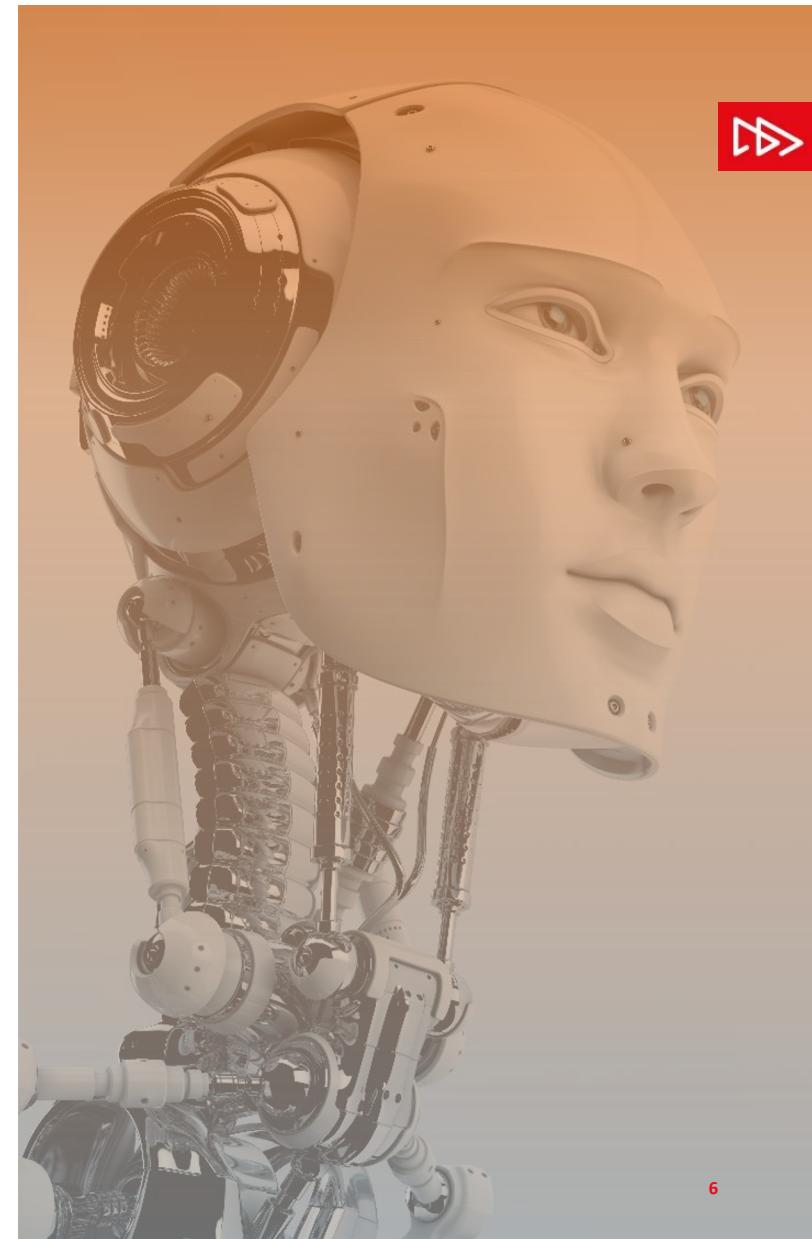
WAT IS A.I.



Agent: autonome computerprogramma's

Rationeel: de beste actie om een doel te bereiken, gegeven
overtuigingen

Rational agents : autonome computerprogramma's die de
verwachte waarde van hun prestatie maximaliseren gegeven hun
huidige kennis



MACHINE LEARNING



MACHINE LEARNING

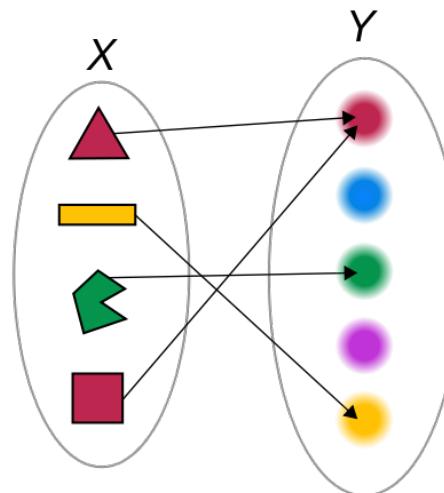


Leer een functie f die input X afbeeldt op Y .

Engels: f maps X to Y

Wiskundig: $f : X \mapsto Y$

Bijvoorbeeld: $f(x) = x^2$



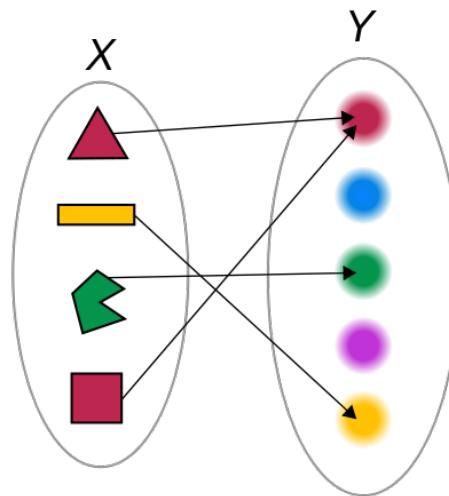
MACHINE LEARNING



Leer een functie f die input X afbeeldt op Y .

X zijn "features", oftewel kenmerken

Y zijn uitkomsten (labels of getallen).



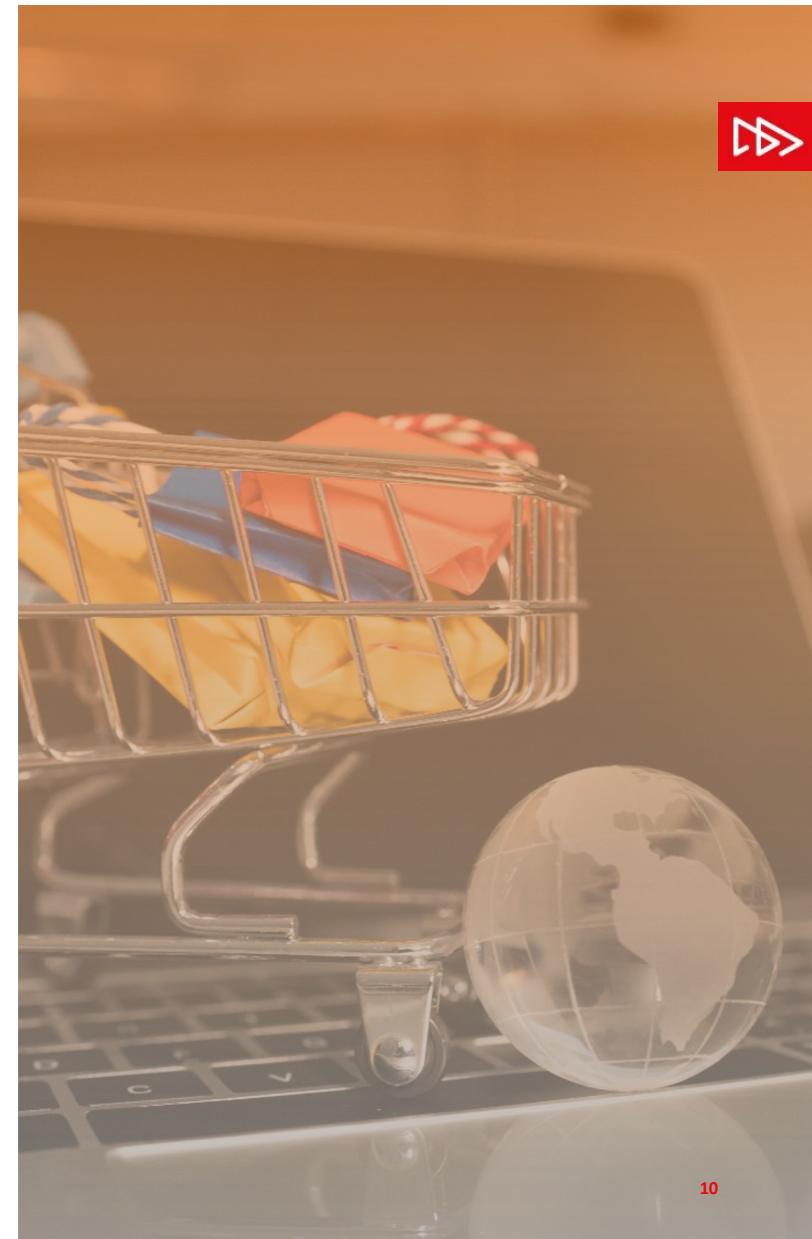
VOORBEELDEN



- Voorspellen van omzet in een supermarkt op basis van omzet vorige week, weer, feestdagen en acties;
- Het vinden van clusters van gebruikers op basis van gebruikspatronen (bv Netflix);
- Het plaatsen van laadpunten in een netwerk.

Wat zijn in deze voorbeelden de features (X)?

En wat zijn de labels (Y)?



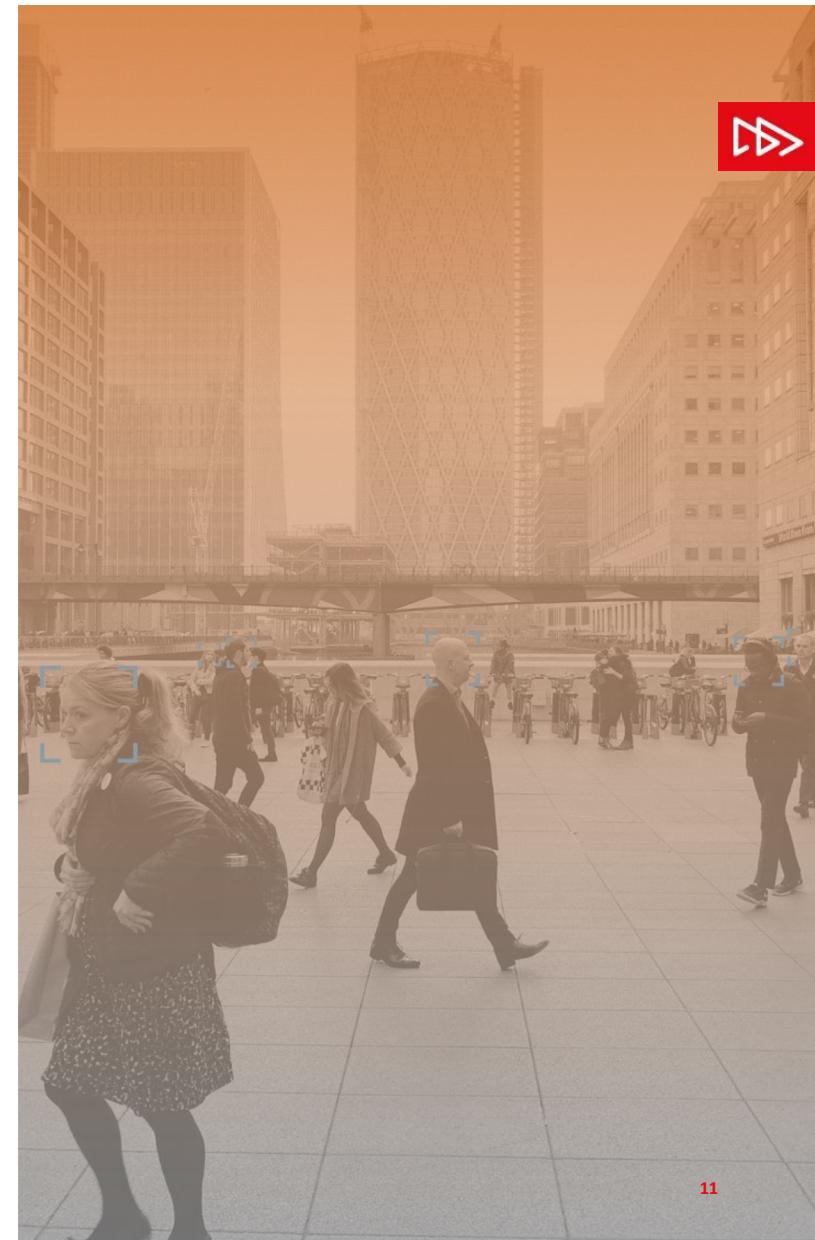
VOORBEELDEN



Maar hoe zit dat dan met:

- Spraakherkenning
- Sentiment-analyse (in spraak of in woord)
- Gezichtsherkenning

Wat zijn nu de X en Y?



SEMANTISCHE VECTOR



Overzetten van betekenis naar een getal.

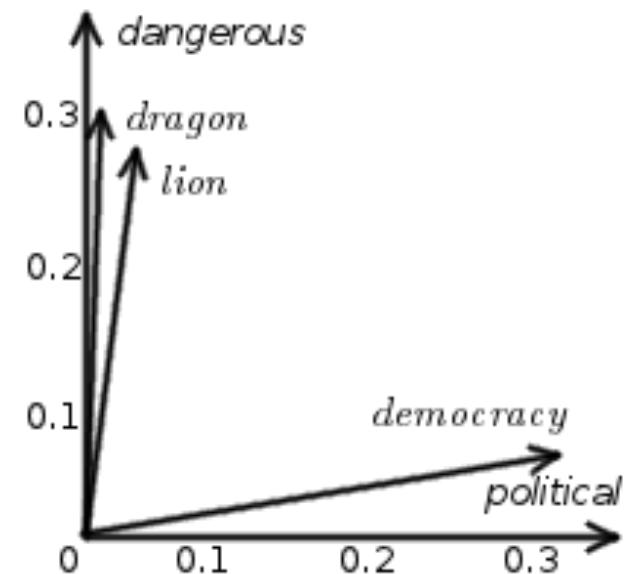
Intuïtie: dingen met dezelfde betekenis, liggen bij dichter bij elkaar in de ruimte.

Het voorbeeld heeft twee assen, en is dus tweedimensionaal:

Leeuw: [0.05, 0.27]

In de praktijk zijn vectoren bv 50 of 200 dimensionaal.

Leeuw: [0.01, 0.54, 0.45, ..., 0.56]



VOORBEELDEN

Onderstaande zijn (minstens) prototypes in een testomgeving in Amsterdam:

- Categorisering burgermeldingen op basis van woorden
- Kentekendetectie voor bepaling milieuzone
- Type afval herkenning vanaf vuilniswagens om afval te tellen
- Voorspellen parkeerdrukte
- Voorspellen stadsdrukte
- Voorspelling fraudekans wonen
- Tips op basis van persoonlijke voorkeuren voor culturele activiteiten via Facebook Messenger chatbot

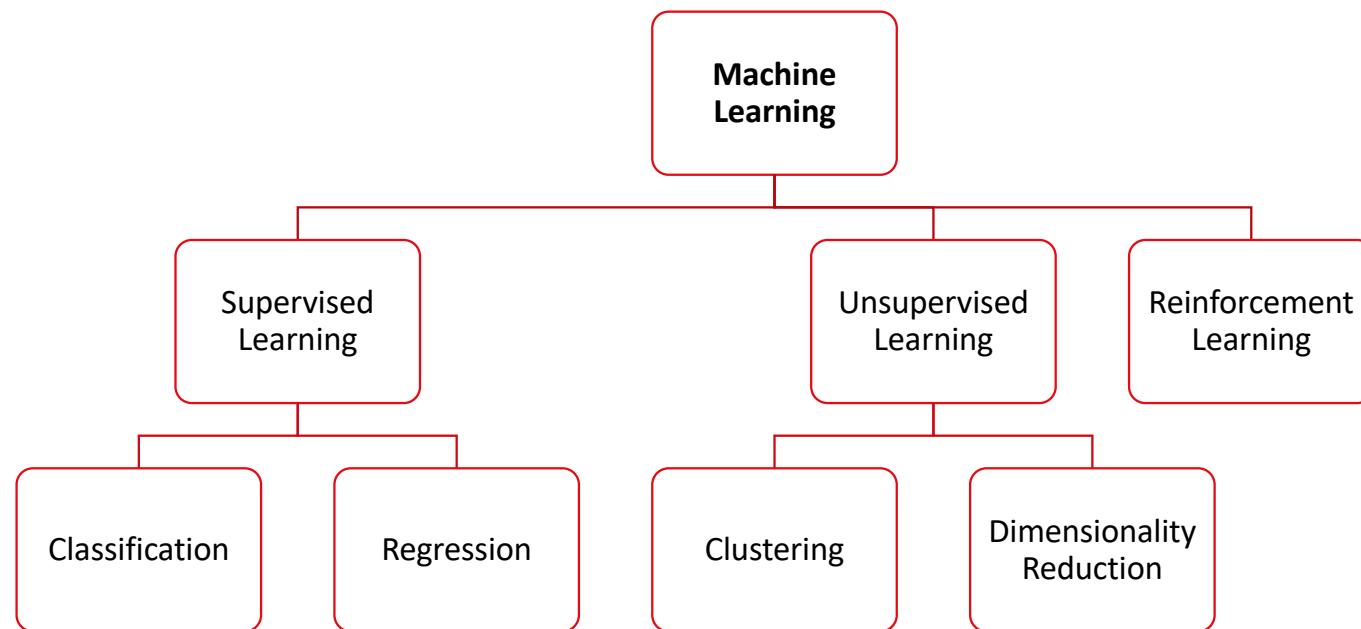
Bron: <https://www.amsterdam.nl/wonen-leefomgeving/innovatie/de-digitale-stad/amsterdamse-intelligentie/>



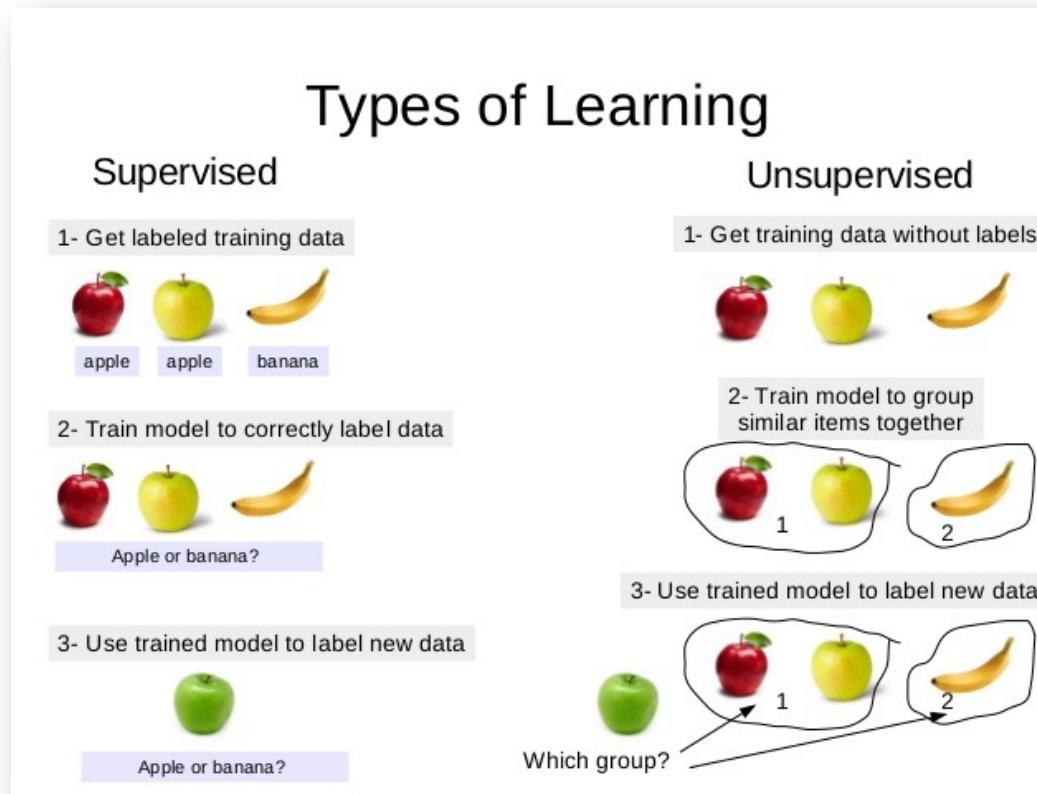
SUPERVISED VS UNSUPERVISED



DRIE SOORTEN MACHINE LEARNING



SUPERVISED VS UNSUPERVISED



SUPERVISED VS UNSUPERVISED



	Supervised	Unsupervised
Discreet	classificatie	clustering
Dontinu	regressie	dimensionality reduction

CLASSIFICATION & REGRESSION



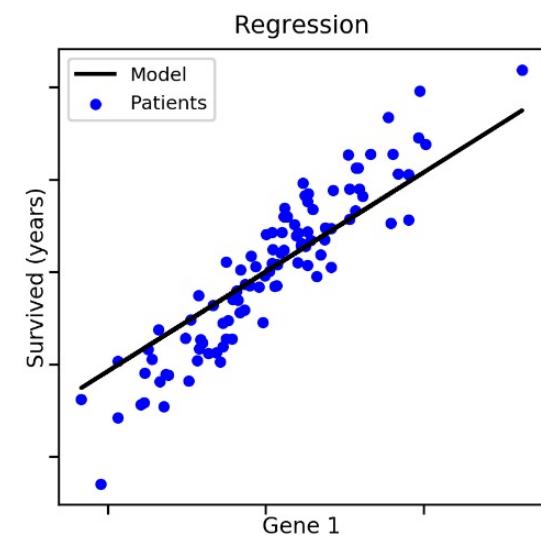
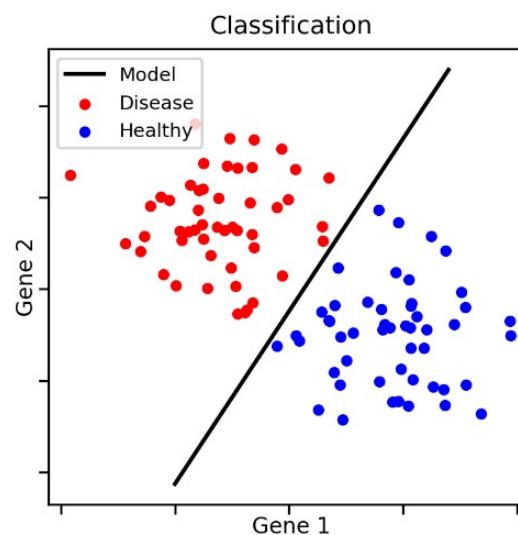
Leer een functie met behulp van gelabelde observaties en voorspel nieuwe, ongelabelde observaties.

Classificatie: voorspel een categorie

- Is deze patient ziek, of gezond?

Regressie: voorspel een waarde

- Hoeveel jaar zal deze patient nog overleven?

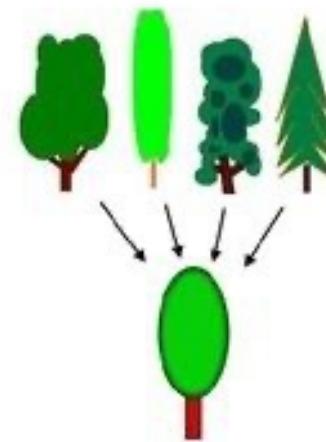


GENERALISATIE



- Het doel is om een y -waarde te voorspellen voor *nieuwe* observaties, *zonder* label (ook al zien de nieuwe voorbeelden er anders uit)
- Dit noemen we **generalisatie**
- Er zijn veel verschillende manieren om de performance van een algoritme te meten. Maar ze kijken allemaal naar het verschil tussen het voorspelde label en het feitelijke label. Dit verschil noemen we de **error**

Generalizing



REGRESSIE

Met regressive berekenen we geen categorie (bv “druk” of “rustig”) maar we berekenen een continue waarde (bv 500 mensen, of 850).

Bijvoorbeeld:

- Voorspellen parkeerdrukte
- Voorspellen stadsdrukte
- Voorspelling fraudekans wonen

Een regressie is altijd om te zetten in een classificatie, andersom gaat dat niet.



REGRESSIE



Voor regressive berekenen we die fout als:

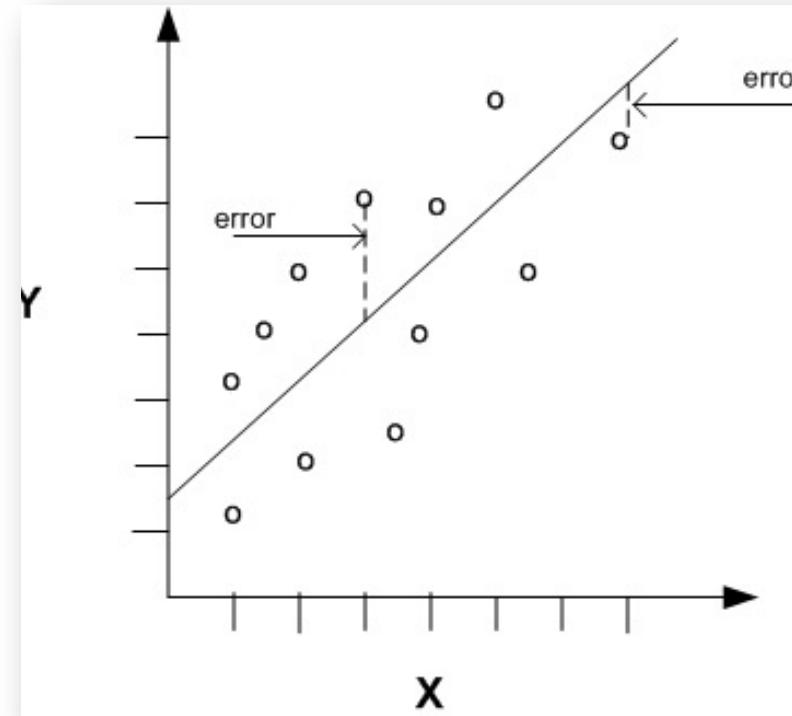
$$\text{error} = y - \hat{y}$$

Hierbij staat \hat{y} (y-hat) voor de voorspelde waarde.

Wat gaat er mogelijk fout als je simpelweg het gemiddelde neemt van alle fouten?

Vaak gebruikt:

- Mean squared error (MSE): $\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
- Mean absolute error (MAE): $\frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$



CLASSIFICATION



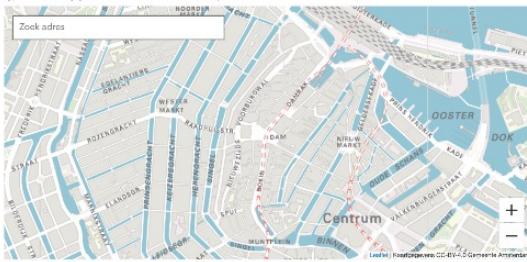
Gemeente
Amsterdam

We pakken op dit moment alleen urgente meldingen op. De afhandeling van uw melding kan daarom tijdelijk langer duren dan de standaard afhandeltermijn die in de bevestigingsmail van uw melding staat. Wij hopen op uw begrip.

Beschrijf uw melding

Waar is het?
Typ het dichtstbijzijnde adres of klik de locatie aan op de kaart.

Zoek adres



Waar gaat het om?
Typ geen persoonlijkegegevens in deze omschrijving, dit wordt apart gevraagd.

0/1000 tekens

Geef het tijdstip aan

Nu
 Eerder

Foto's toevoegen
Voeg een foto toe om de situatie te verduidelijken.

+

Volgende ➔

CLASSIFICATION



		diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
	id						
	8510653	B	13.08	15.71	85.63	520.0	0.10750
	84799002	M	14.54	27.54	96.73	658.8	0.11390
	853401	M	18.63	25.11	124.80	1088.0	0.10640
	84862001	M	16.13	20.68	108.10	798.8	0.11700
	85638502	M	13.17	21.81	85.42	531.5	0.09714

CLASSIFICATION

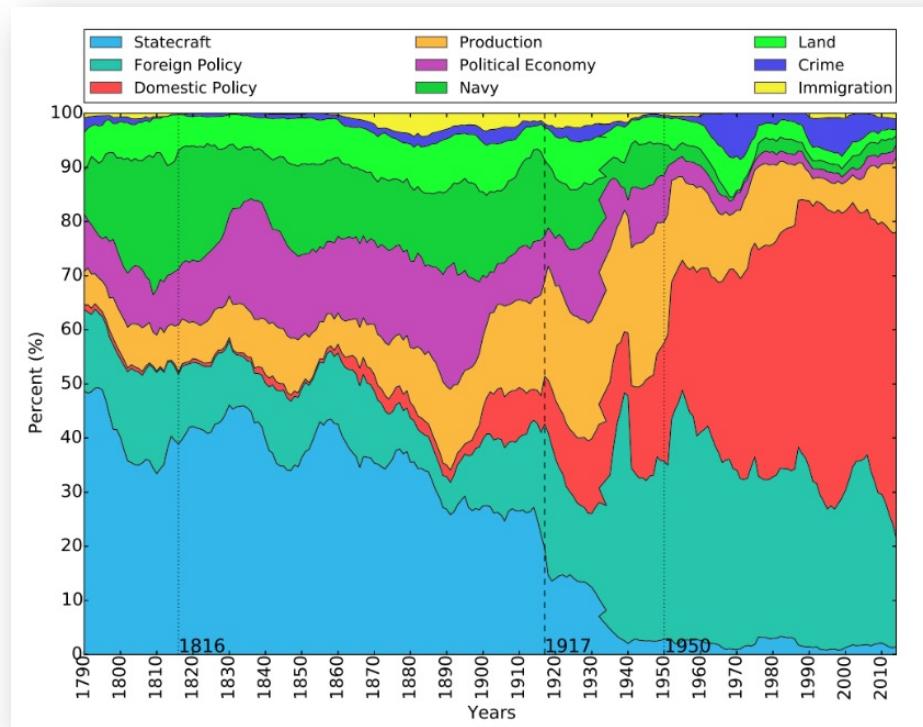
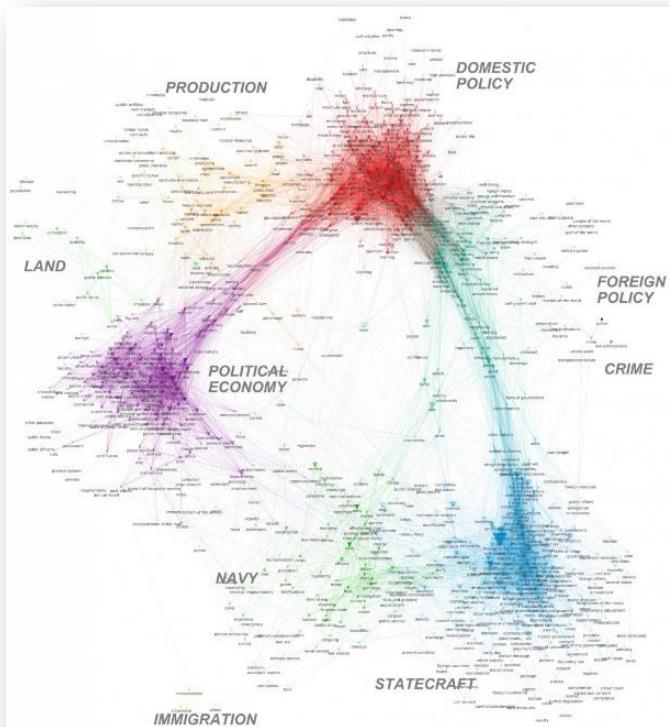


		diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
	id						
8510653		B	13.08	15.71	85.63	520.0	0.10750
84799002		M	14.54	27.54	96.73	658.8	0.11390
853401		M	18.63	25.11	124.80	1088.0	0.10640
84862001		M	16.13	20.68	108.10	798.8	0.11700
85638502		M	13.17	21.81	85.42	531.5	0.09714

voorspelling	correct
B	1
M	1
M	1
B	0
M	1

Accuracy
80%

CLUSTERING



Rule, A., Cointet, J.-P., Bearman, P. S., Breiger, R. L., & Mohr, J. (n.d.). *Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790-2014*. <https://doi.org/10.1073/pnas.1512221112>

CLUSTERING

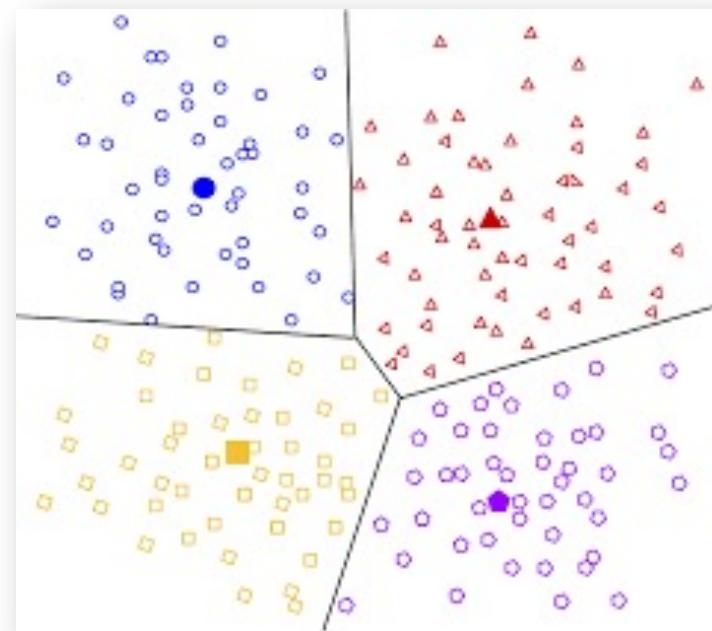


Zoek uit welke punten dicht bij elkaar liggen. Elk punt krijgt het label van zijn k dichtbijzijnste buren (met k een getal > 0, bijvoorbeeld 3).

Vandaar de naam: k-means nearest neighbor clustering.

De definitie van wat dichtbij is, hangt af van de manier waarop dat gemeten wordt:

- Euclidisch
- Manhattan
- Cosine similarity

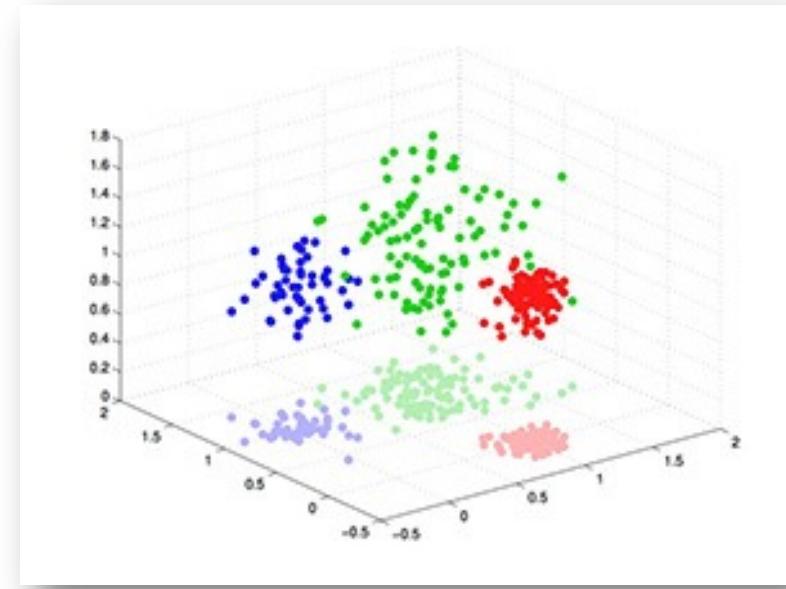


DIMENSIONALITY REDUCTION



Doel: verminder het aantal variabelen in X, maar behoudt de informatie.

- Als scheiden van kaf en koren niet goed lukt kun je de informatie 'platslaan'
- Bruikbaar voor visualisatie (bv 2-dimensionale weergave van complexere data)
- Vaak voorbereiding op supervised learning



AUTONOME ALGORITMES



PERSONALISATIE



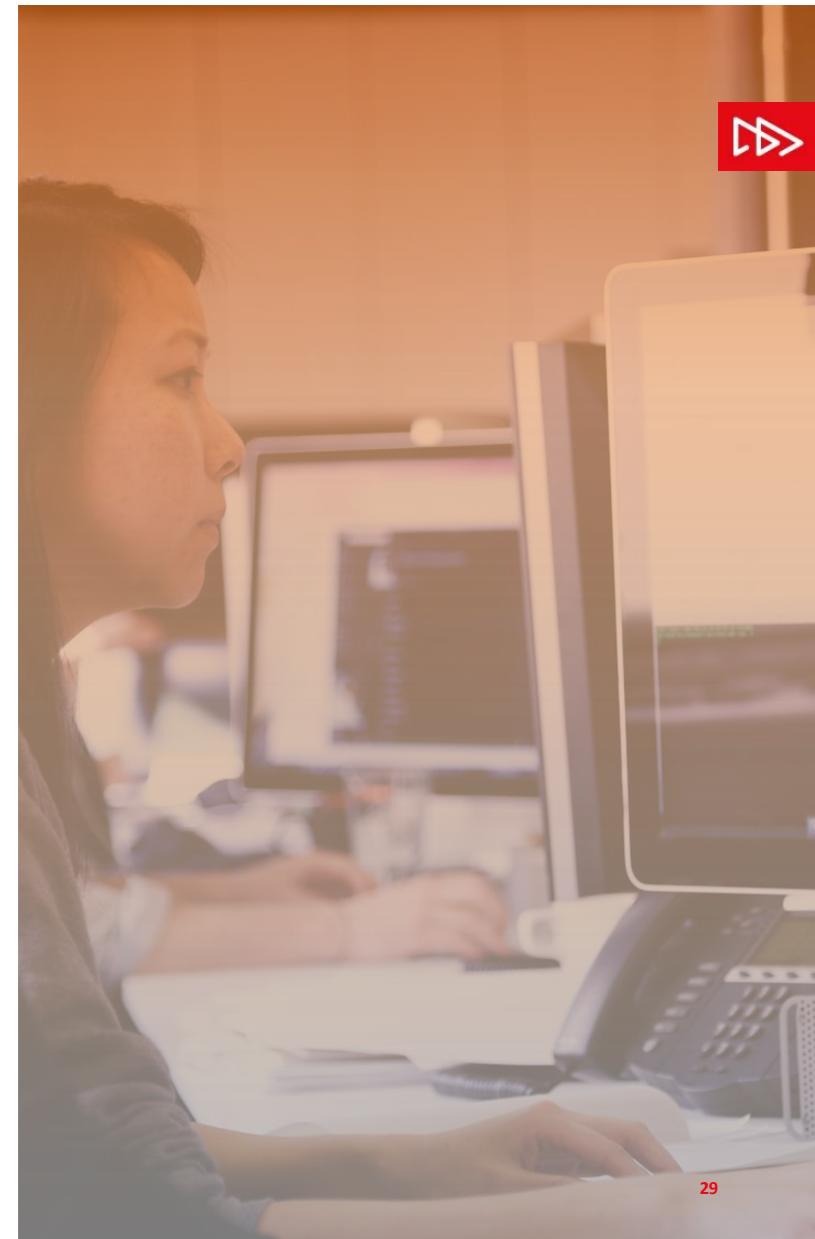
Er zijn algoritmes die speciaal op jou getraind zijn.

Algoritmes leren in de praktijk:

- Aanpassen van resultaten zoekmachine als het erop lijkt dat je niet vindt wat je zoekt

Maar ook:

- Beïnvloeden van verkiezingen met behulp van gepersonaliseerde advertenties



AUTONOME ALGORITMES

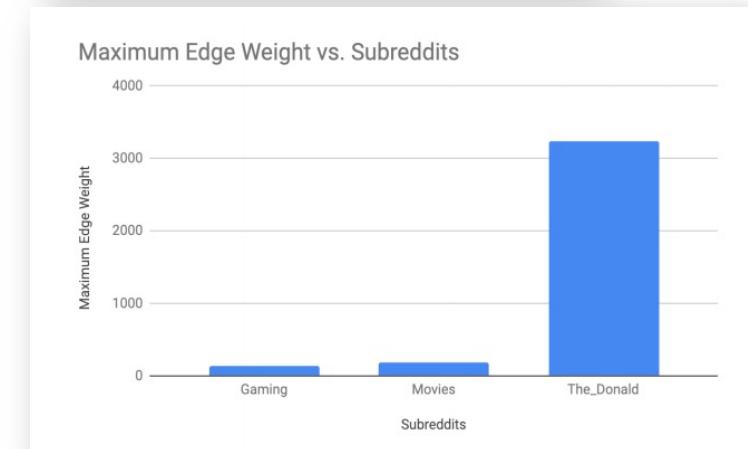
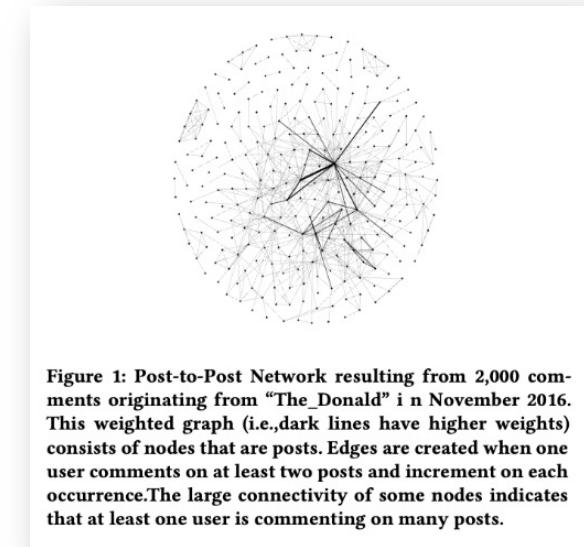


Geautomatiseerde social media bots die de politieke opinie proberen te beïnvloeden.

Bot detection:

Hurtado, S., Ray, P., & Marculescu, R. (2019, April). Bot Detection in Reddit Political Discussion. In *Proceedings of the Fourth International Workshop on Social Sensing* (pp. 30-35).

Badawy, A., Ferrara, E., & Lerman, K. (2018, August). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 258-265). IEEE.



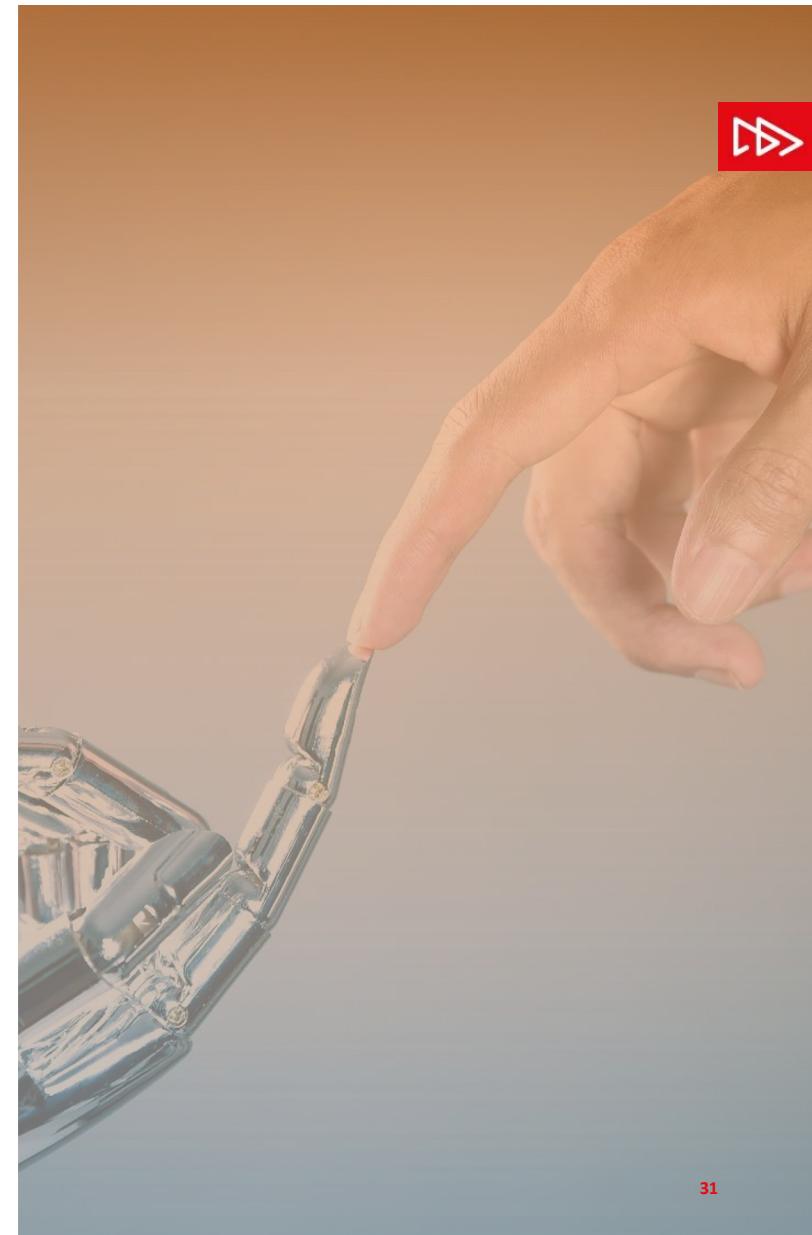
AUTONOME ALGORITMES



Nederland:

- De volgers van Nederlandse politici op sociale media zijn voor een groot deel nep-accounts, die soms plotseling in aantal toenemen
- Plotselinge stijging in 2015 van aantal volgers Wilders (van 400.000 naar 600.000) waarbij 75% bots blijken
- Tijdelijke, exponentiële toename van retweets en likes van DENK-voorman Kuzu op Twitter tijdens sociale en politieke onrust.

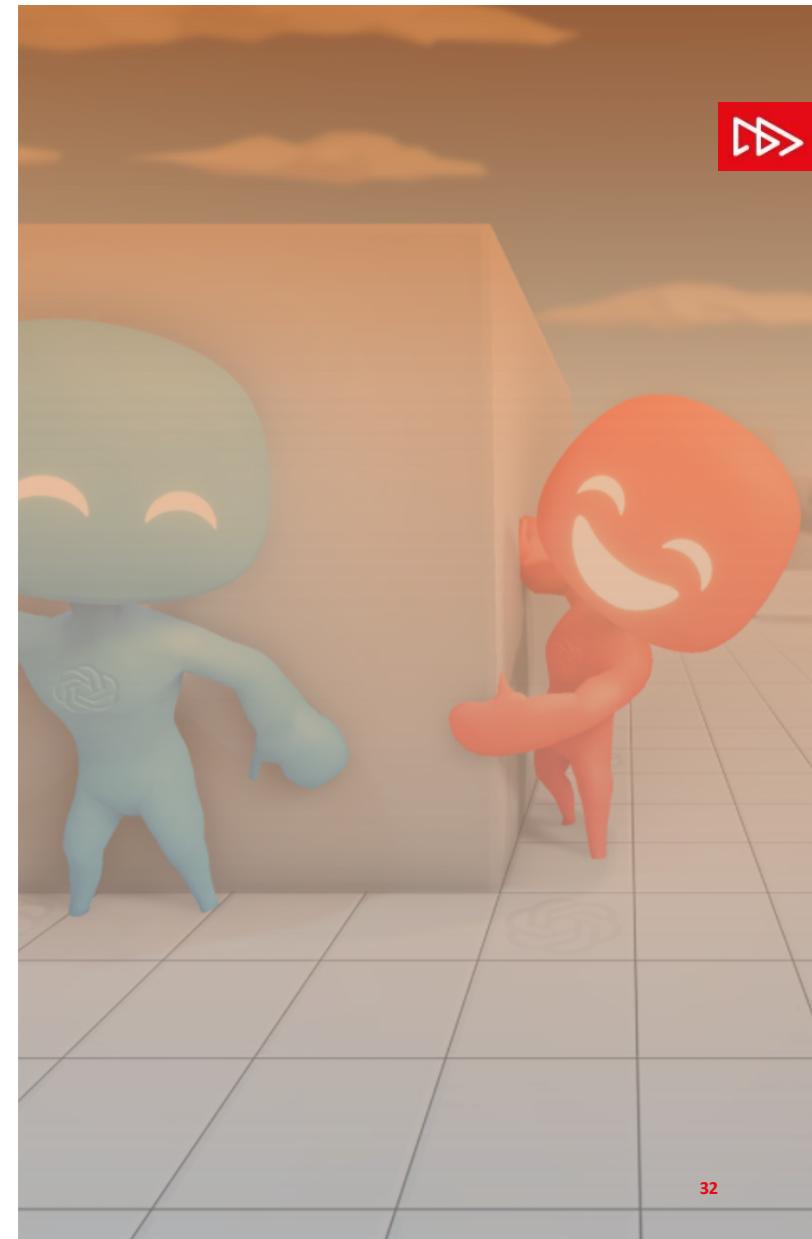
Bron: <https://www.rathenau.nl/nl/digitale-samenleving/digitalisering-van-het-nieuws/desinformatie-zo-werken-bots-nederland>



REINFORCEMENT LEARNING

- Wat valt je op aan hoe de twee teams zich gedragen?
- Wat verrast je?
- Wat maakt dat oplossingen worden gevonden?

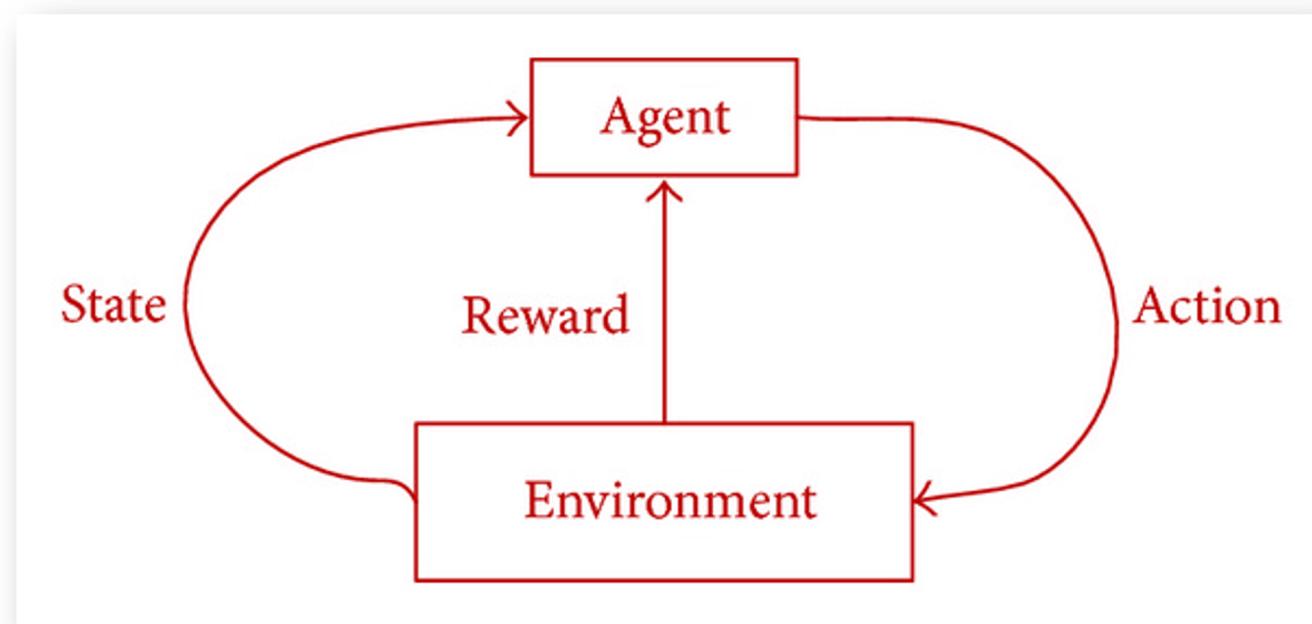
<https://youtu.be/Lu56xVIZ40M>



REINFORCEMENT LEARNING



- State: bijvoorbeeld een bepaalde combinatie van stukken op een schaakbord.
- Maar ook: de combinatie van verkeer en bewegwijzering voor een zelfrijdende auto.



REINFORCEMENT LEARNING



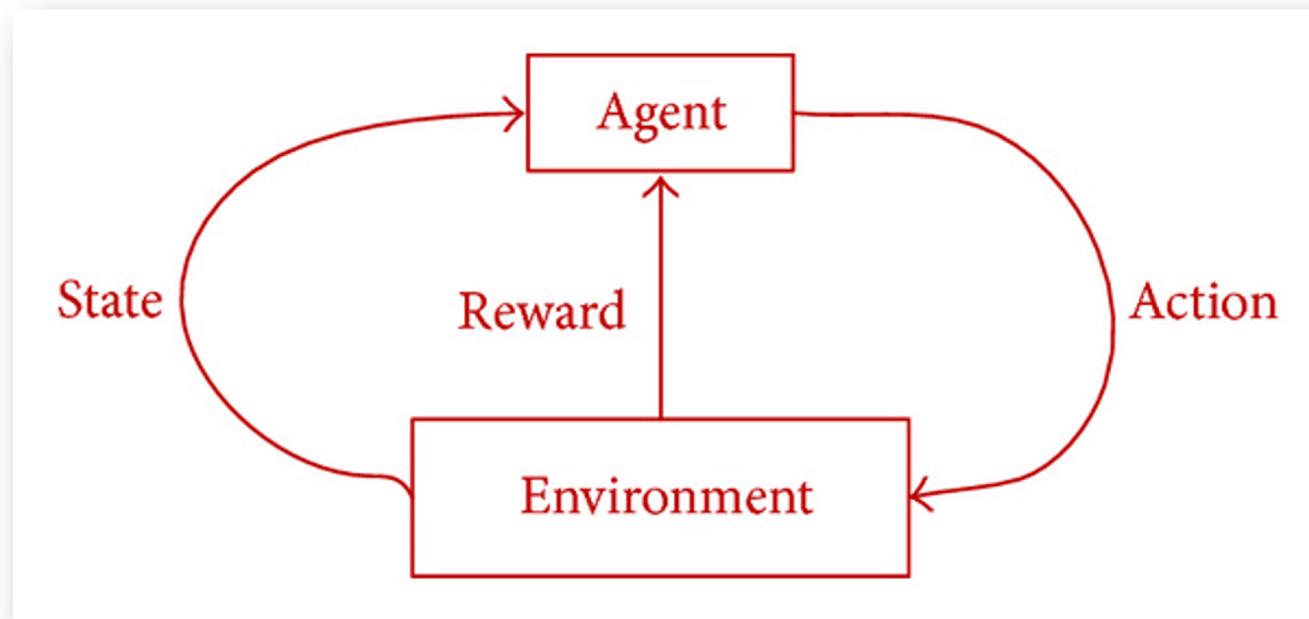
- Actie: doe iets, zodat je in een nieuwe state komt.



REINFORCEMENT LEARNING



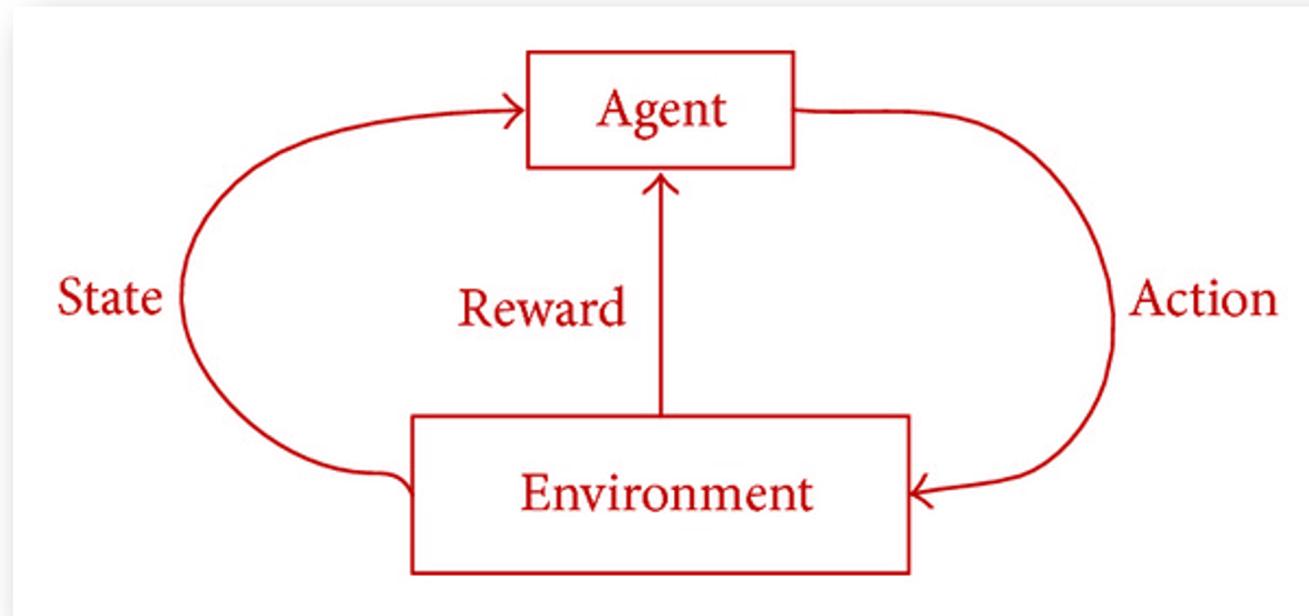
- Niet elke actie levert gegarandeerd een bepaalde transitie op naar een volgende state.
- Waarom niet?



REINFORCEMENT LEARNING



- Elke state levert een reward op.
- Dit is de 'reinforcement' voor een bepaalde actie in een bepaalde state.



REINFORCEMENT LEARNING

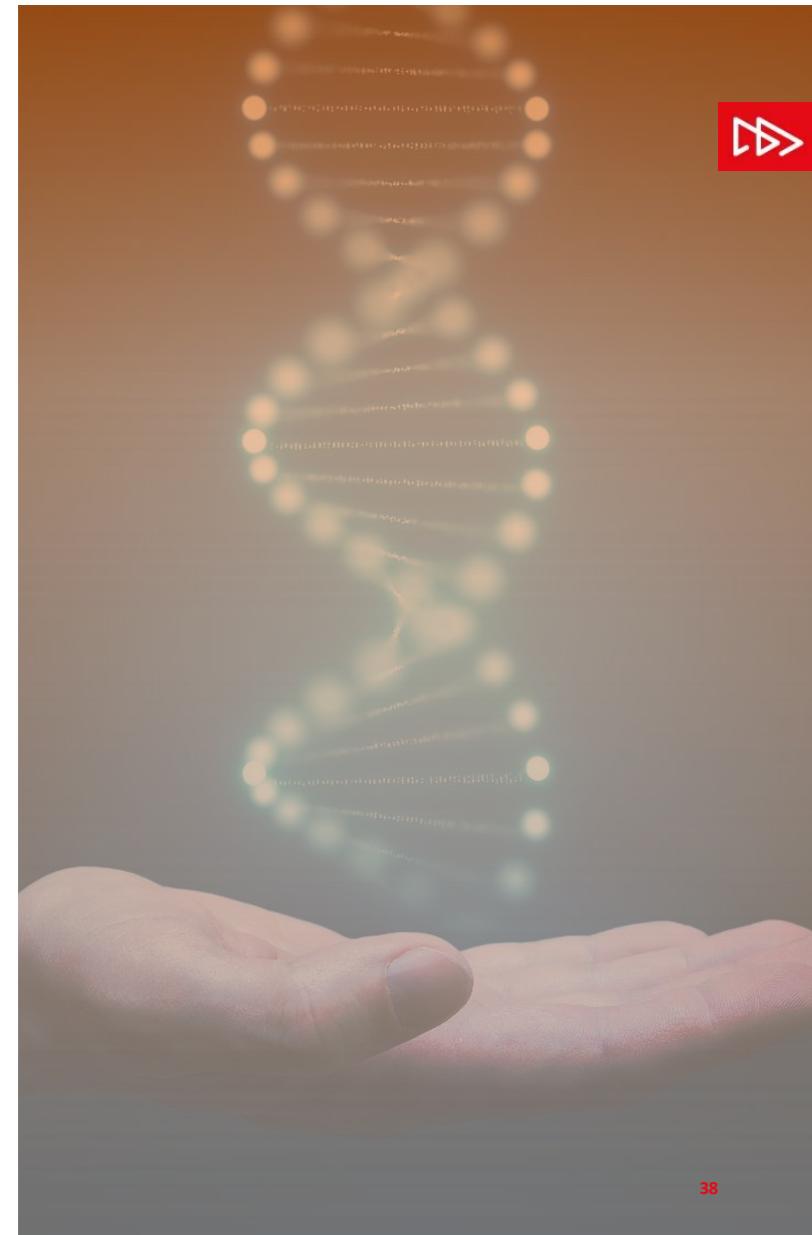


- Moet het algoritme zich richten op korte termijn rewards? Of lange termijn rewards?



REINFORCEMENT LEARNING

- Wanneer je een omgeving kunt simuleren, en je een reward kunt definieren, kan een algoritme heel veel tijd besteden aan uitzoeken wat werkt.
- De simulaties kunnen heel complex worden: bv het uitzoeken wat de effecten zijn van protheses en operaties:
- <https://www.youtube.com/watch?v=kie4wjB1MCw>



REINFORCEMENT LEARNING

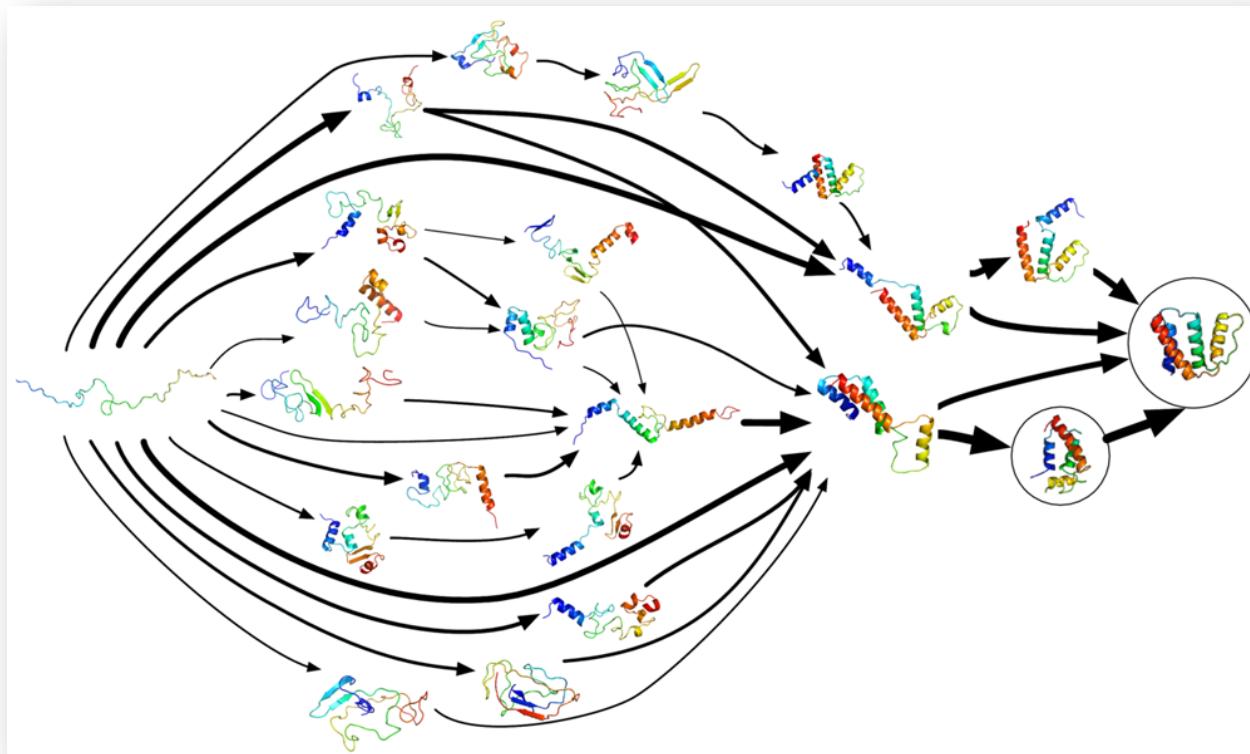
- Tip: AlphaGo documentaire
- <https://www.youtube.com/watch?v=WXuK6gekU1Y>



REINFORCEMENT LEARNING



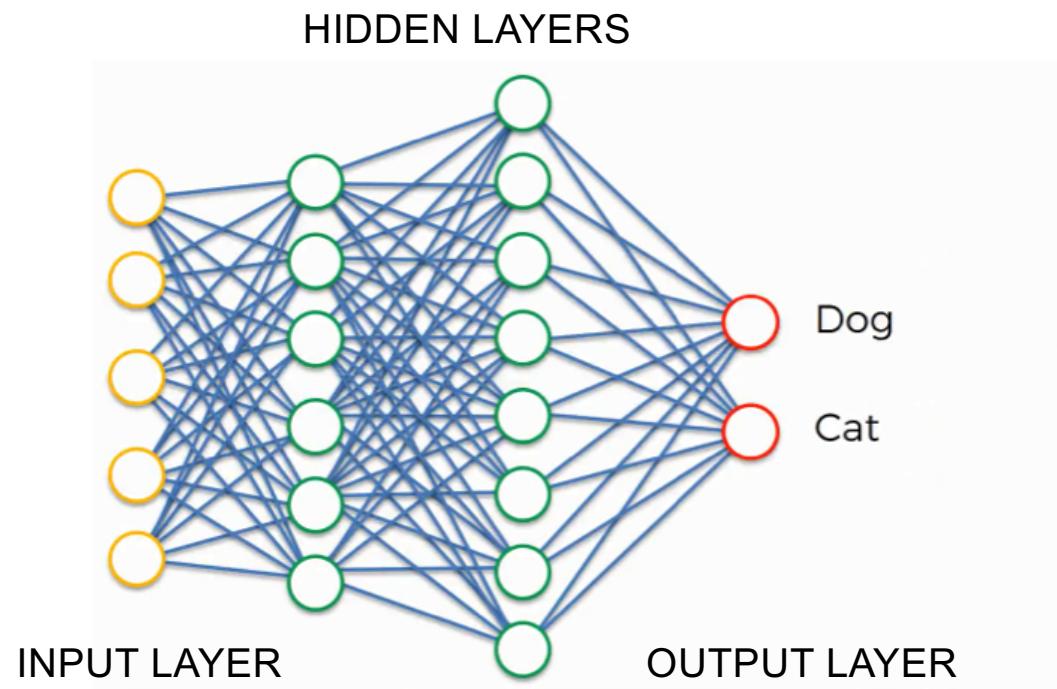
REINFORCEMENT LEARNING



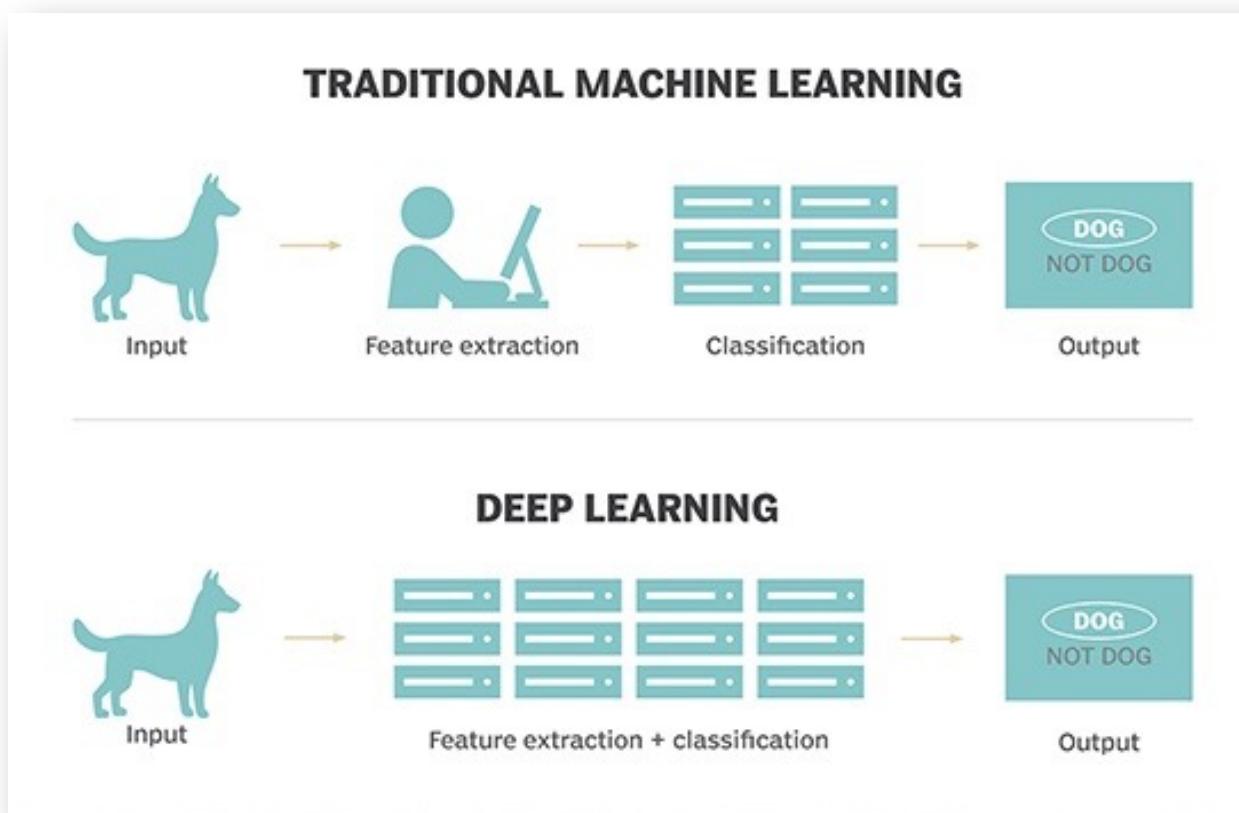
DEEP LEARNING



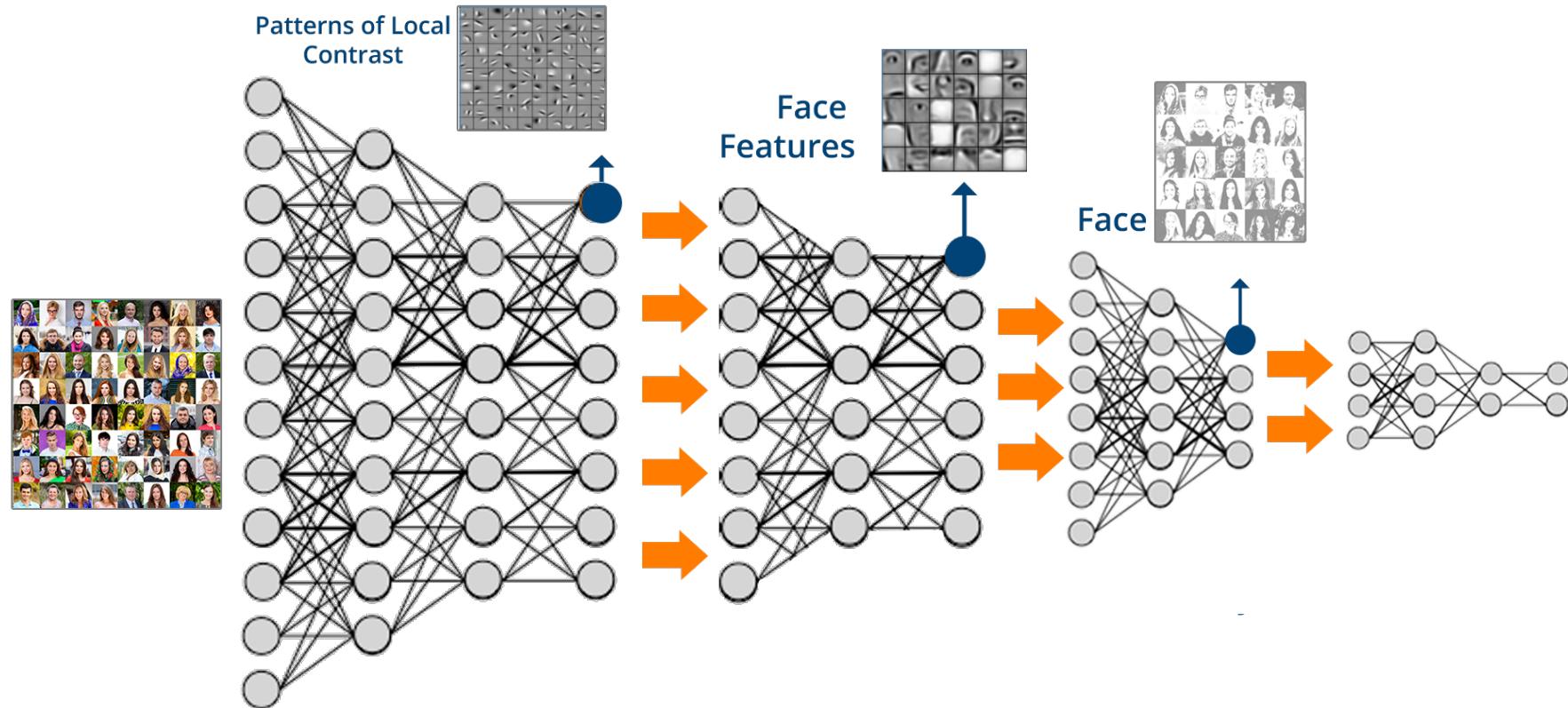
DEEP LEARNING - NEURAL NETWORK



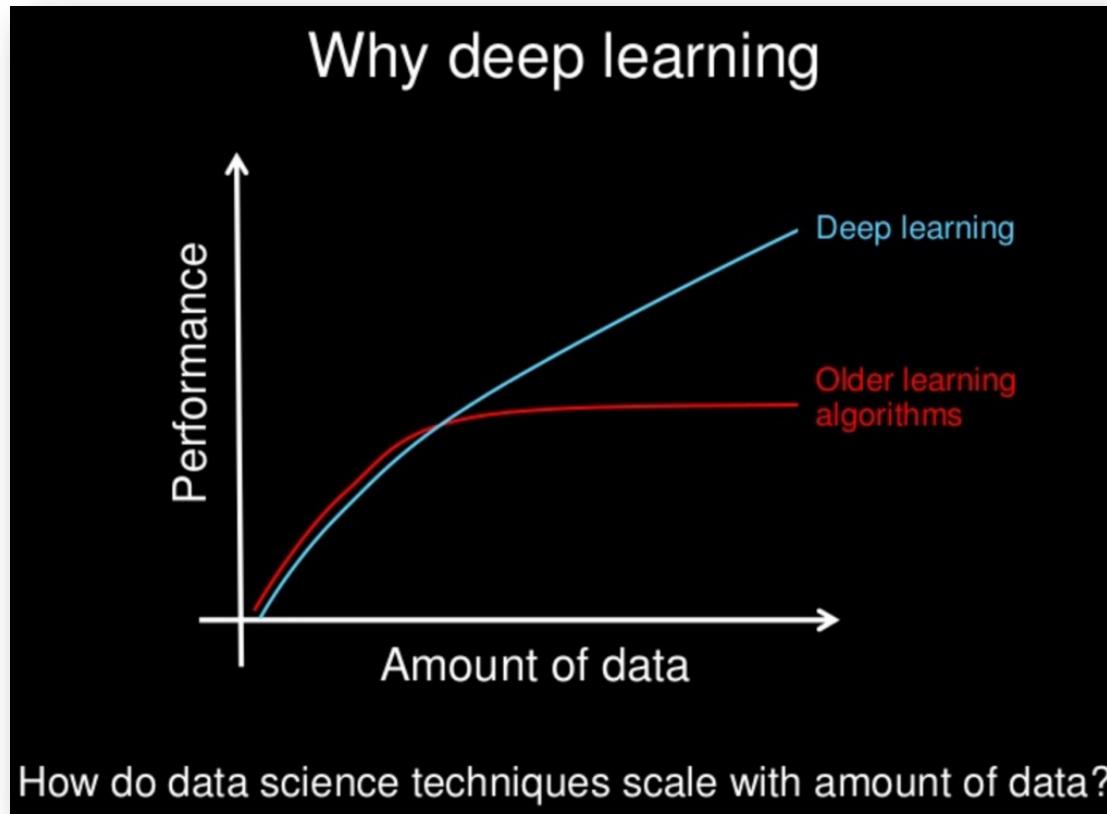
DEEP LEARNING



DEEP LEARNING



DEEP LEARNING





ANCHORMEN
data activators

THANK YOU FOR YOUR ATTENTION

- Pedro de Medinalaan 11,
1086 XK Amsterdam
- 020 - 773 1972
- www.anchormen.nl

