



Een blik op de toekomst van verantwoorde AI: Mens over AI over AI

Catholijn M. Jonker (TUD, Univ. Leiden, Hybrid Intelligence Centre)

Met dank aan Frank van Harmelen voor een deel van de slides

Toekomstbeeld Mens met Artificiële Intelligentie (AI)

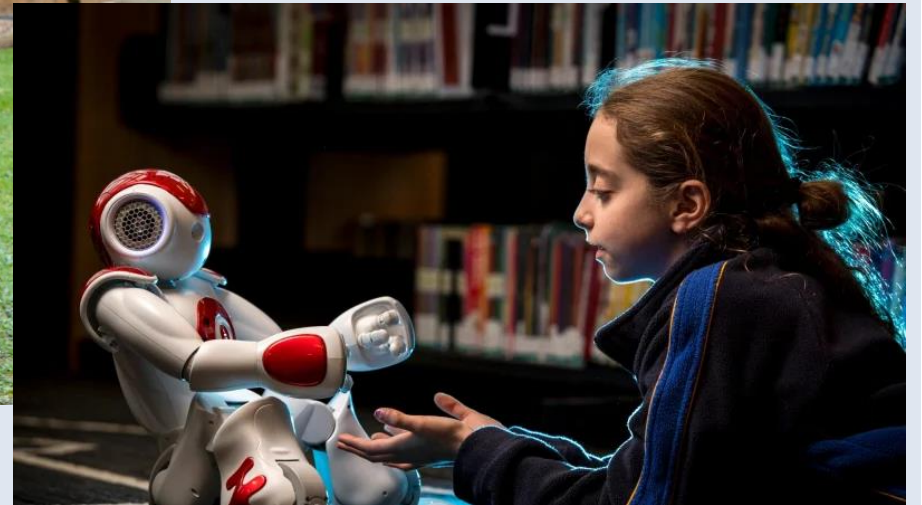
- Versterkt autonomie
- Verrijkt ervaringen
- Nieuwe activiteiten
- Versterkt democratieën

Hybride
Intelligentie

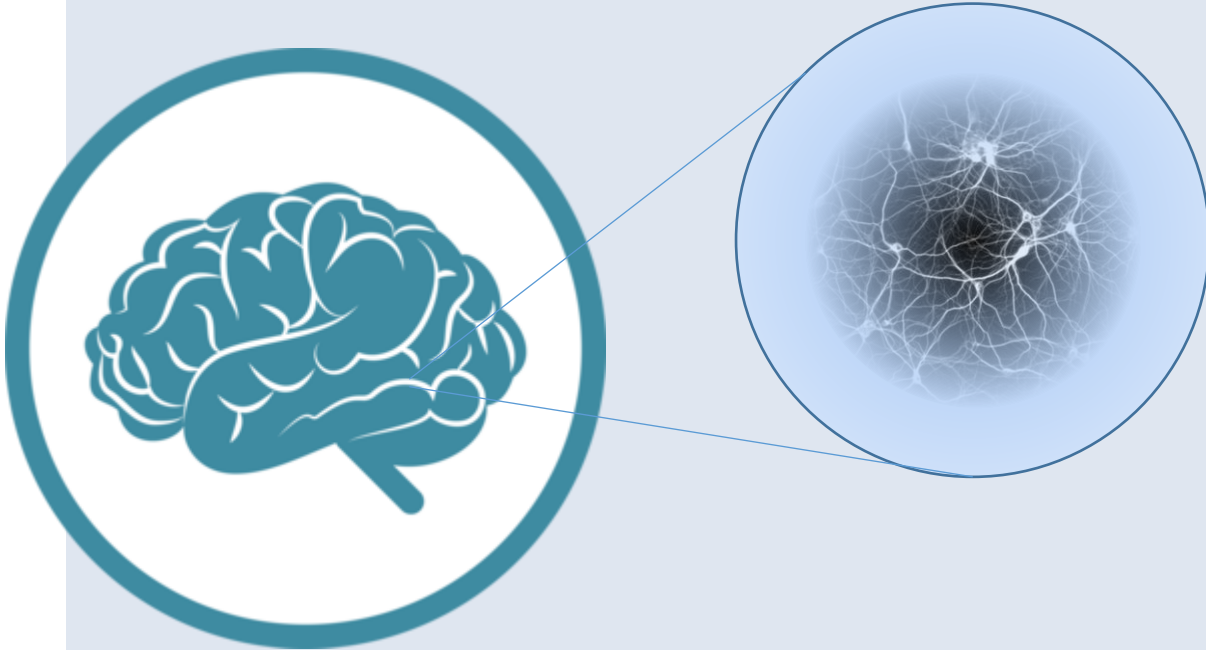
H*i*

Stem AI af op menselijke waarden

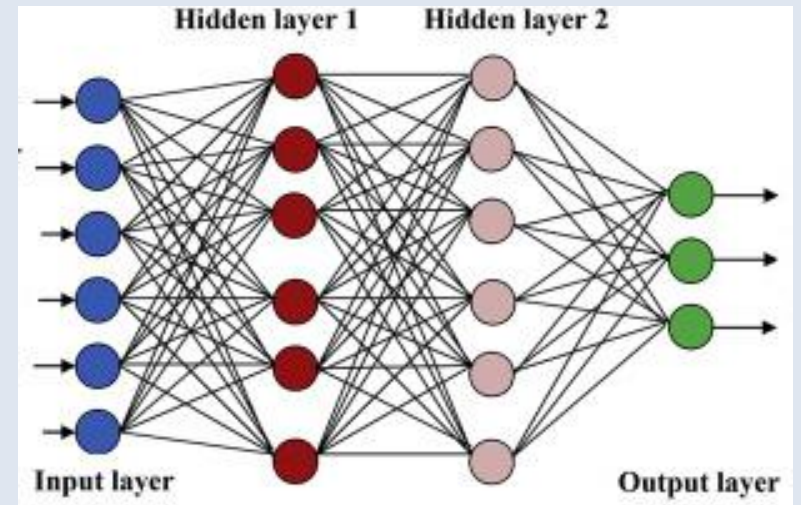
AI as Partners



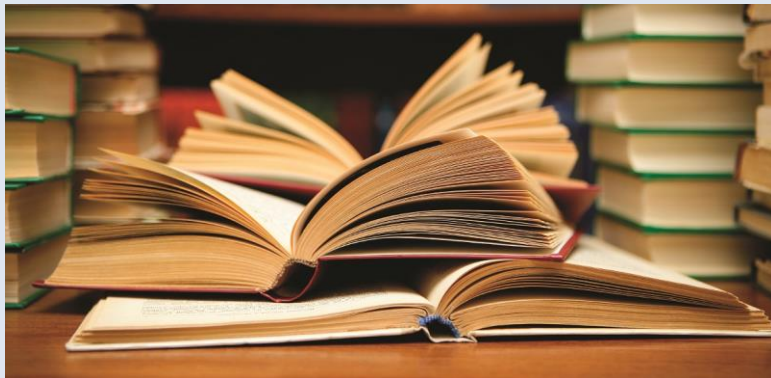
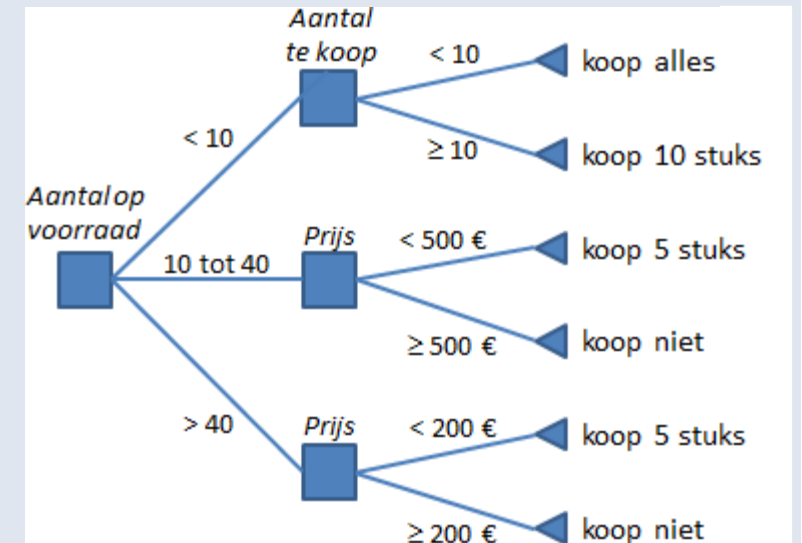
AI basics (Jaren 60)



Kunstmatige Neurale Netwerken

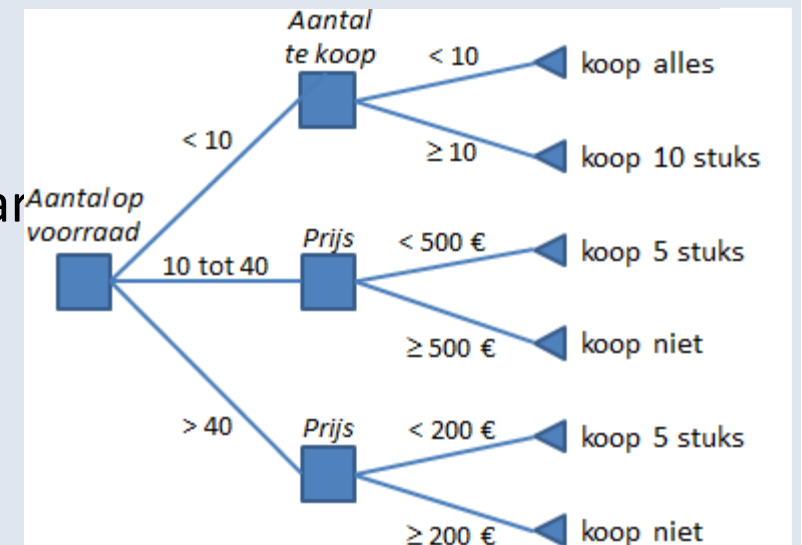
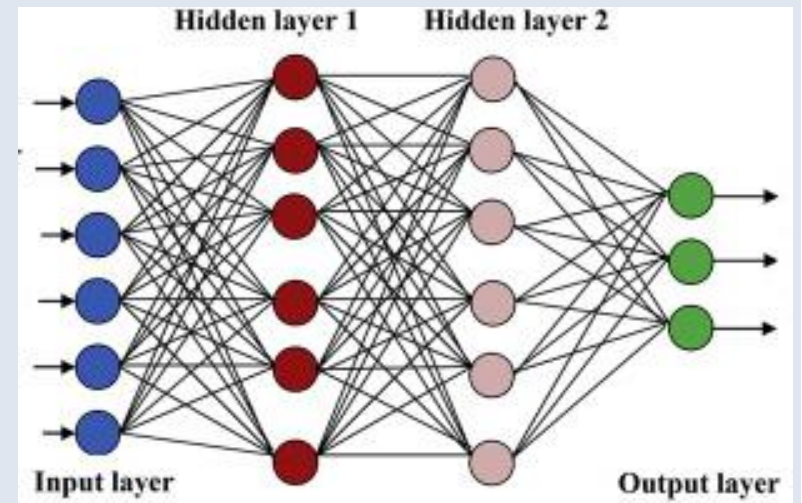


Kennis-gebaseerde systemen



AI basics (Jaren 60)

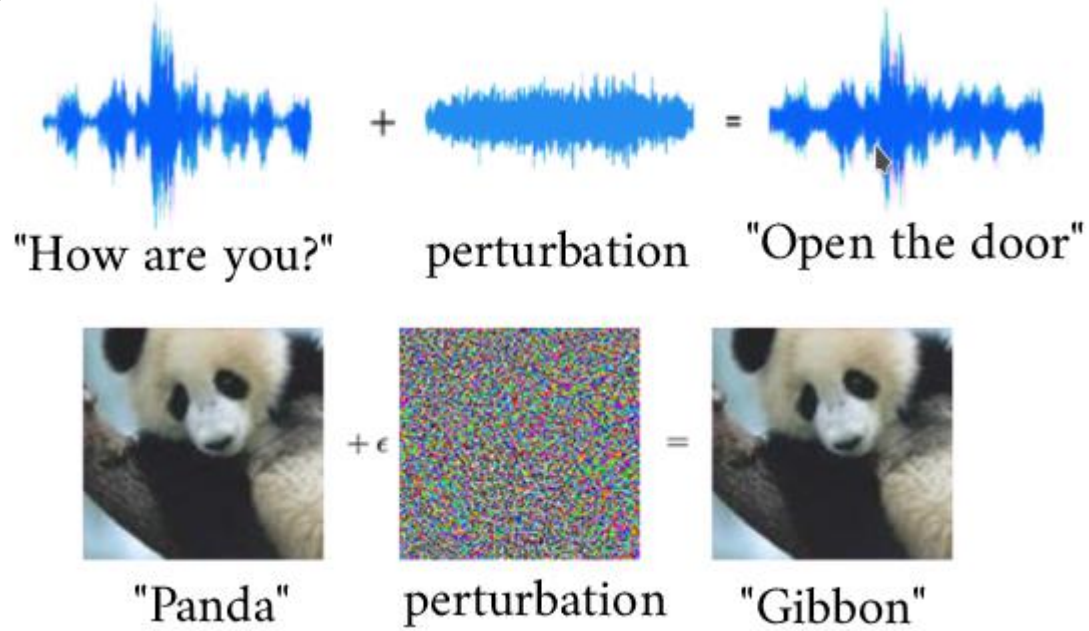
- Kunstmatige neurale netwerken
 - Pro: simuleerbaar in gewone computers
 - Pro: Adaptief, kunnen nieuwe associaties maken
 - Con: Computers waren niet krachtig genoeg
 - Con: Lastig om te ontwerpen
- Kennis-gebaseerde kunstmatige intelligentie
 - Pro: we zagen hoe we verder moesten
 - Pro: Effectief en efficiënt
 - Pro: Je weet wat ze weten: controleerbaar en stuurbaar
 - Con: Uitvragen van experts: duur, langzaam
 - Con: Onderhoud is arbeidsintensief
 - Con: Niet adaptief



De doorbraak van machine learning → Deep Learning



- wetenschappers
 - Bayesiaanse methoden voor redeneren met kansen
 - Machine Learning: van kennis-gedreven naar data-gedreven methoden
 - Trainingsmethoden voor machine learning
- De toename in rekencapaciteit maakt Deep Learning (gelaagde neurale network architectuur) mogelijk
- De beschikbaarheid van Big Data



[Greg Anderson](#), [Isil Dillig](#) (2019)

© Evtimov et al.



Zelf-rijdende auto's in de war door stickers op verkeersborden

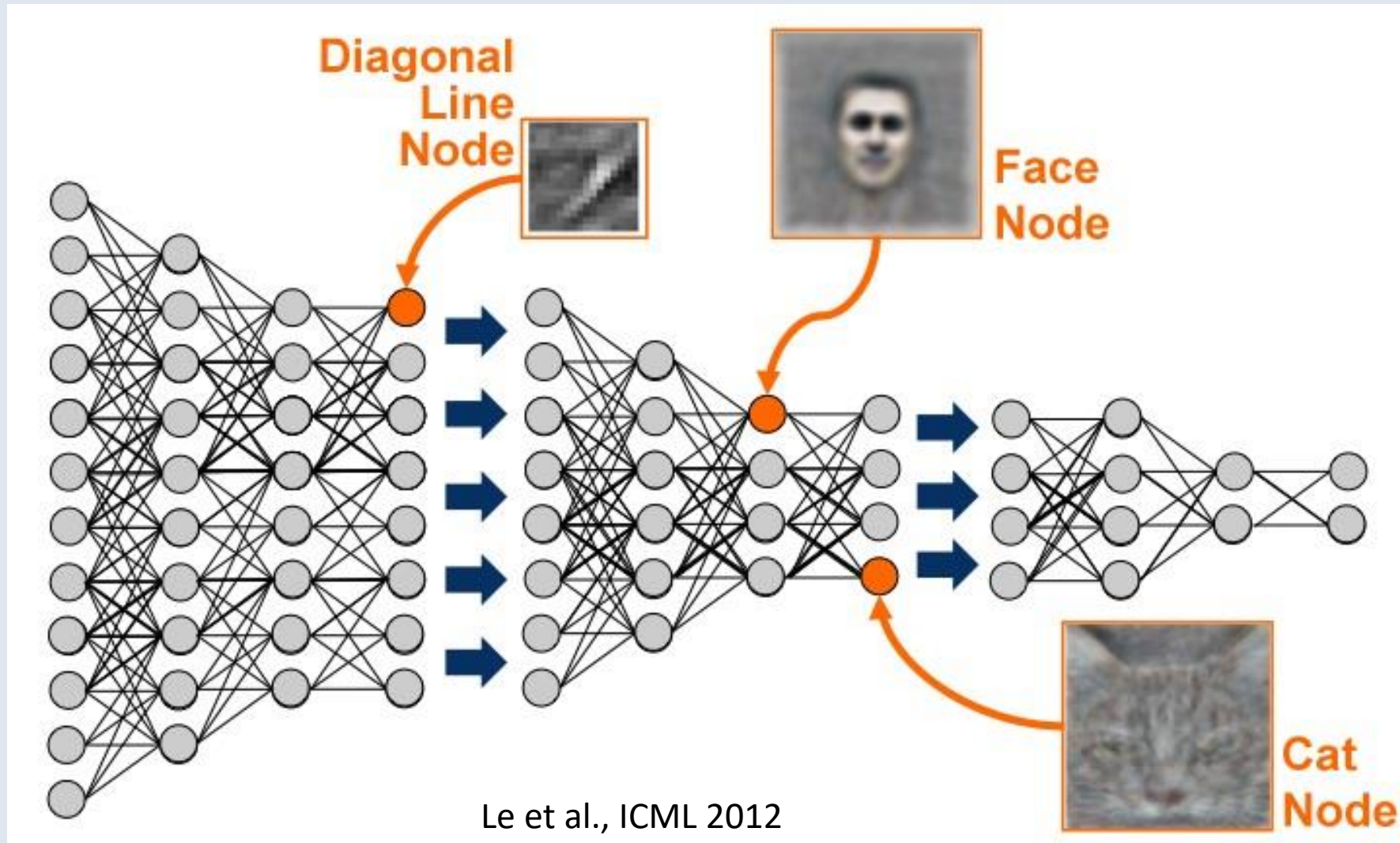
Vb: discriminerende zoekresultaten op basis van discriminerend consumentengedrag



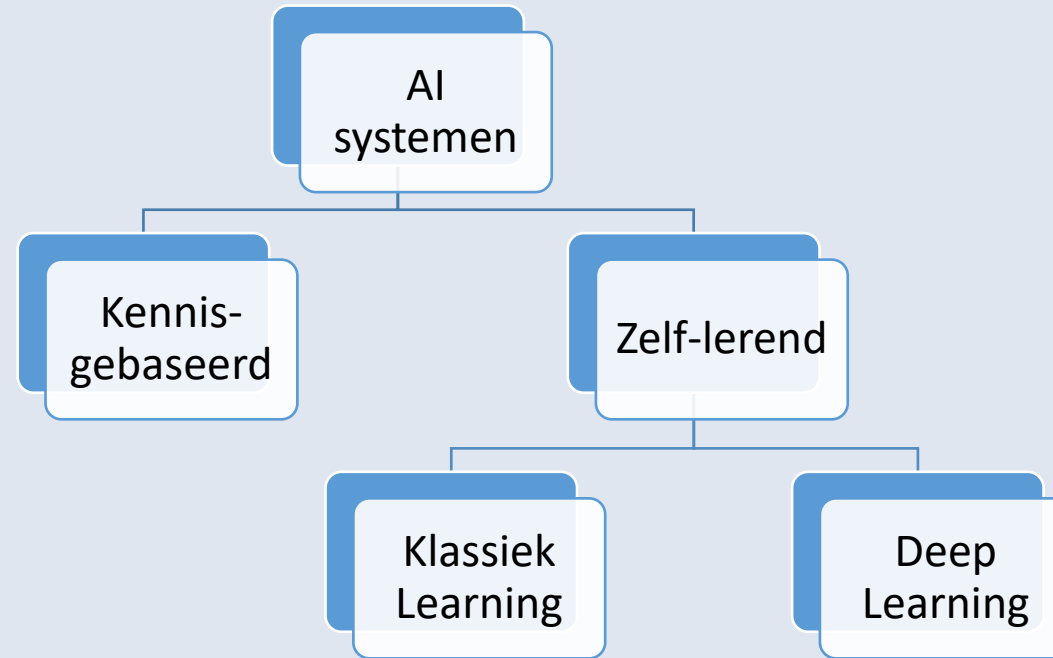
onderselectie van minderheden bij sollicitaties op basis van historische onderselectie



Kennis opgeslagen in Neurale Netwerken



Verskillende typen AI systemen



AI: zelf-lerende systemen

"systemen die verbanden ontdekken in grote hoeveelheden data, en op basis daarvan een **kans** berekenen"

Vb: kans dat iemand geschikt is voor een baan

kans dat iemand borstkanker heeft

kans dat iemand gefraudeerd heeft

...

AI: Klassieke Machine Learning

Mensen bepalen welke aspecten (“features”) belangrijk zijn

*Vb: SyRI: boetes, opleiding,
onroerend goed,
schulden, inburgering,
...*



Enigszins stuurbaar, controleerbaar
(door wetgever, opdrachtgever, ambtenaar, burger)

AI: Deep Learning

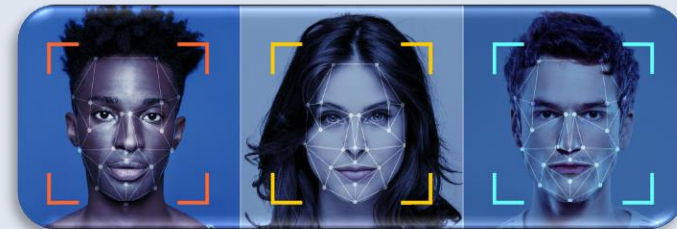
Machine bepaalt zelf welke aspecten (“features”) belangrijk zijn

Vb: gezichtsherkenning

voor risico-classificatie

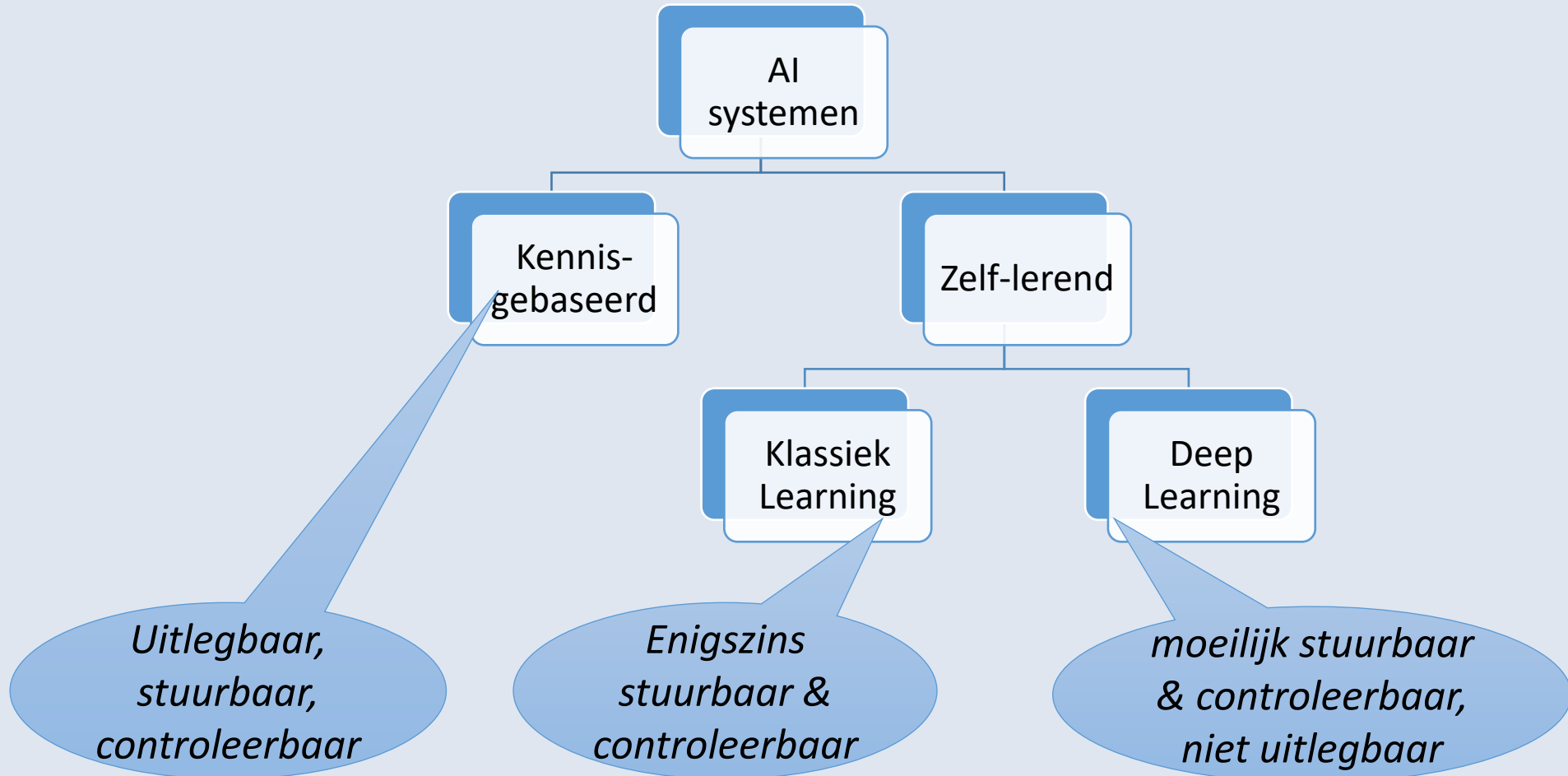
Welke features worden gebruikt?

Huidskleur? Bril? Haardracht? Hoofddoek?



Moeilijk stuurbaar, niet uitlegbaar

Verschillende typen AI systemen



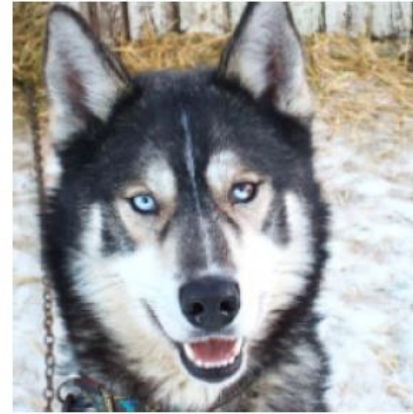
Boodschap: Niet elke vorm van AI is even zorgelijk

Het data vraagstuk

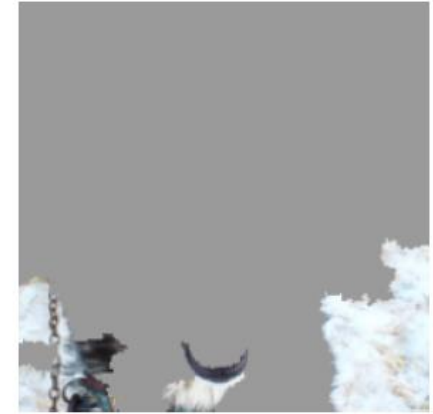
- Machine learning reproduceert systematische fouten in de trainings data
- Niet-relevante maar gecorrelleerde patronen krijgen onterecht betekenis

discriminerende
zoekresultaten op basis van
discriminerend
consumentengedrag

Google



(a) Husky classified as wolf



(b) Explanation: Snow

Ribeiro et al., 2016 "Why should I trust you?"

onderselectie van minderheden bij
sollicitaties op basis van historische
onderselectie

amazon



AI anno nu

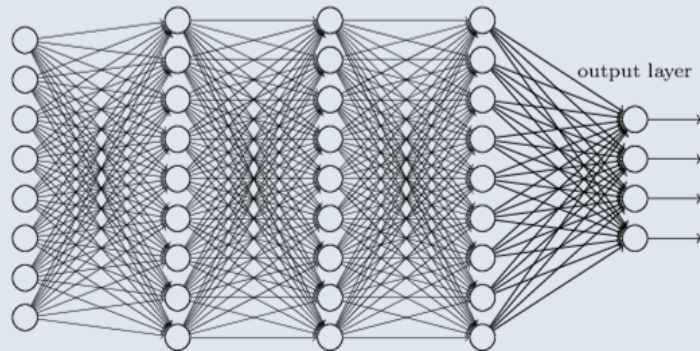
- Kunstmatige neurale netwerken
 - Mogelijk door de toegenomen rekenkracht van computers
 - Hangt af van **Big Data** van **goede kwaliteit**
 - Adaptief, kunnen nieuwe associaties maken
 - Je weet niet wat ze weten: niet uitlegbaar, moeilijk controleerbaar / verifieerbaar
- Kennistechnologische AI
 - Uitvragen van experts: duur, langzaam
 - Efficient rekenen
 - Niet adaptief
 - Onderhoud is intensief
 - Je weet wat ze weten: wel controleerbaar, en stuurbaar

Black box AI

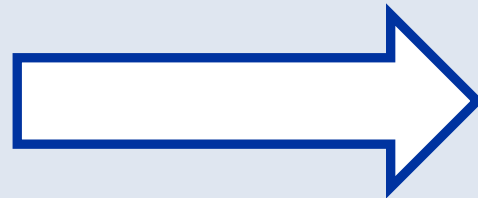
Knowledge-based AI

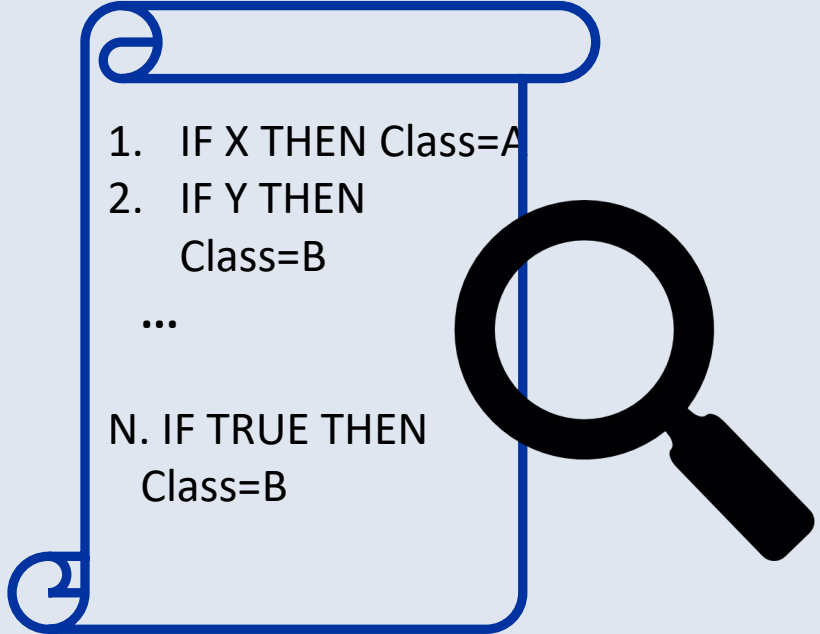
Op weg naar uitlegbare Machine learning

Kunnen we een leesbaar model uit een black box trekken?



Adapted from *Neural Networks and Deep Learning*, by Michael A. Nielsen, 2015.
Determination Press (CC BY-NC 3.0)

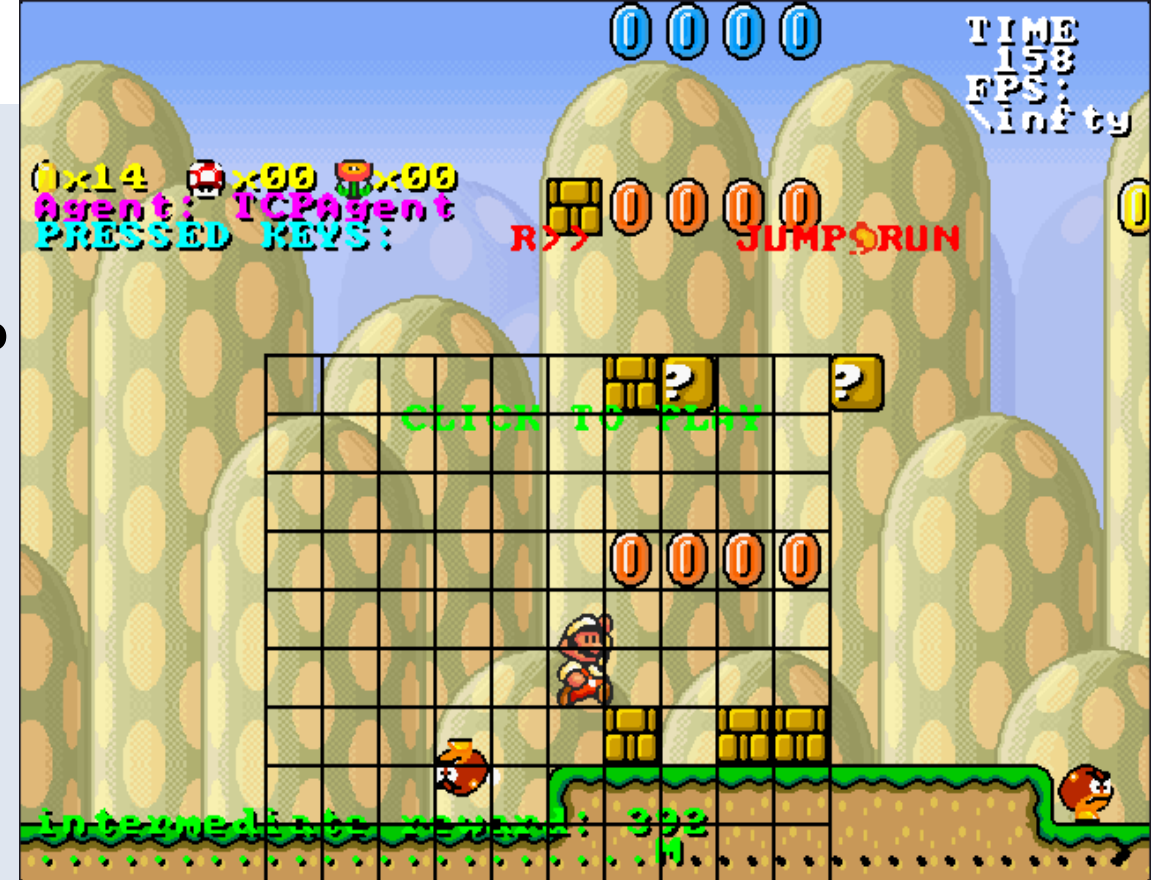


- 
1. IF X THEN Class=A
2. IF Y THEN
Class=B
...
N. IF TRUE THEN
Class=B
- A magnifying glass is positioned over the list of rules.

Experiment

Wat is een goede strategie voor MARIO?

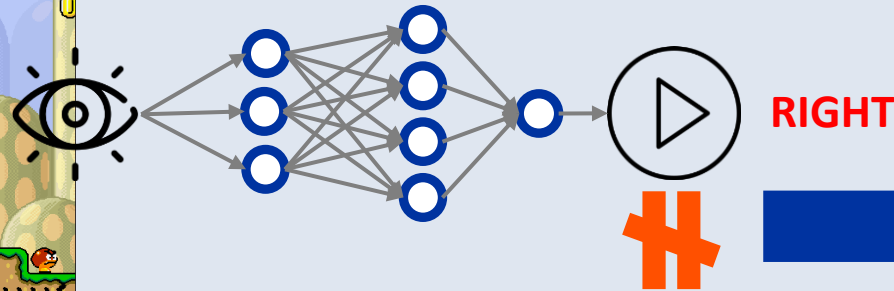
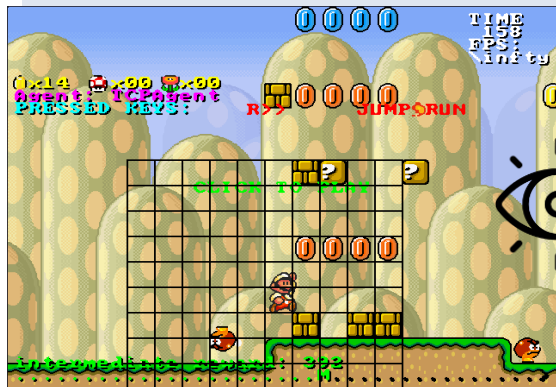
- MarioAI Benchmark
 - Verzamel rode munten
 - Vermijd de blauwe en de groene
- Black box ML algoritme laten leren
 - Steckelmacher et al. (2020)*
- Training set van 50 episodes
 - Toestanden: 10x10 grid rond om Mario
 - Acties: controller knoppen of niets doen



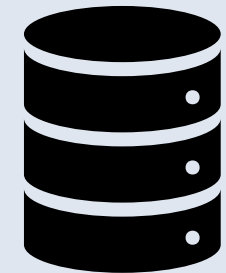
* Steckelmacher, D., Plisnier, H., Roijers, D.M., Nowé, A.: Sample-efficient model-free reinforcement learning with off-policy critics. In: Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., Robardet, C. (eds.) Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Artificial Intelligence, vol. 11908, pp. 19–34. Springer International Publishing, Cham, Switzerland (2020)

Leren in drie fasen:

- 1) Wat zijn de hoofdlijnen van de strategie?
- 2) Hoe goed is die samenvatting?
- 3) Interactieve verfijning: Wat zijn de uitzonderingen op die samenvatting?



1. IF (1, 5)==NULL AND (0, 5)==NULL AND (0, 1)==NULL THEN Class=JUMP
2. IF (1, 5)==BRICK AND (1, 0)==COIN_RED THEN Class=RIGHT
3. IF (1, 5)==BRICK AND (2, 4)==COIN_RED THEN Class=RIGHT
4. IF (4, 2)==NULL AND (0, 4)==NULL THEN Class=JUMP
5. IF (4, 5)==NULL AND (3, 3)==NULL THEN Class=JUMP
6. IF (3, 5)==BRICK THEN Class=JUMP
7. IF (2, 2)==BRICK THEN Class=JUMP
8. IF (0, 2)==NULL AND (0, 1)==NULL THEN Class=RIGHT
9. IF (0, 5)==NULL THEN Class=JUMP
10. IF TRUE THEN Class=JUMP



Exceptions
Rule 1



Exceptions
Rule 2



Exceptions
Rule 3



...

Mensen samen met AI over AI



Take Away message

Hybrid Intelligence: combineer

- Machine Learning
- Knowledge Representation
- Menselijke intelligentie

Samenwerken, niet vervangen

AI over AI: open de black box

Mens over AI: bestuurbaar en controleerbaar

<https://www.delftdesignforvalues.nl/>

<https://www.hybrid-intelligence-centre.nl/>