Meepraten over betrouwbare en verantwoorde AI

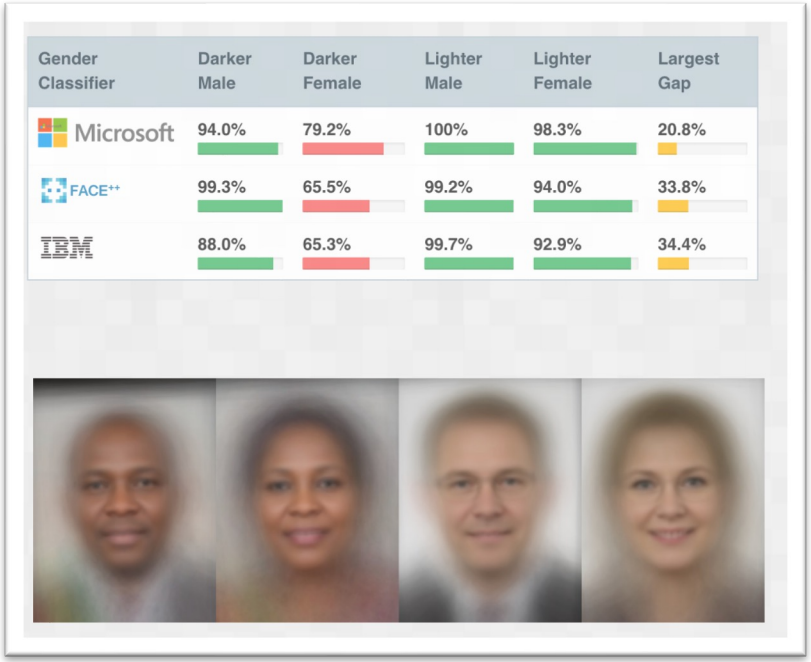
Wat moet ik weten?

Cynthia C. S. Liem
Multimedia Computing Group, TU Delft
@cynthiacsliem@akademienl.social.nl
c.c.s.liem@tudelft.nl

Spraakverwarring!

- Imprecies taalgebruik
  - AI, algoritme, systeem, robot bijna synoniem in publiek debat
  - Heel verschillende jargontermen voor informatici
  - ‘Wat is een algoritme dan exact?’ / ‘Wat is het exacte algoritme?’

Schandalen



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

http://gendershades.org/overview.html

deVerdieping Trouw

Het is de hoogste tijd dat al die onmisbare werknemers een eerlijk loon krijgen

Belastingdienst ging vooral achter lage inkomens aan

Fraudejacht • Om toeslagen te controleren op fouten en fraude gebruikte de Belastingdienst een zelflerend algoritme. Dat selecteerde vooral mensen met lage inkomens voor controle.

Jan Kleinnijenhuis
redactie binnenland

In het kort

De Belastingdienst gebruikte bij de controle op fraude een zelflerend algoritme

In de praktijk weer dat model lagere inkomens aan als risicovol

Er was geen proces om deze selectiebias tegen te gaan

Als de uitkomst zo inzoomt op een specifieke groep dan moeten er alarmbellen afgaan

Cynthia Liem
universitair hoofddocent kunstmatige intelligentie

82,3
procent van de groep met een hoge risicoscore had een inkomen van minder dan 20.000 euro

Selectiebias

Verzoek om openbaarheid

de Verdieping 2|3
Hoe de Belastingdienst lage inkomens profileerde

Plannen voor code zwart
Ziekenhuizen in de knel
vandaag 4|5

Hoe gaat een agent om met rellen?
Geestelijke zorg bij hulpdiensten
vandaag 7

Criminelen werken op feestdagen
Online oplichters schakelen tandje bij
nederland 9

Nieuw huis voor De Schreeuw
Oslo heeft een Munchmuseum
de Verdieping 12|13

Ontdek Trouw in optima forma

https://www.trouw.nl/politiek/hoe-de-belastingdienst-lage-inkomens-profileerde-in-de-jacht-op-fraude~bbb66add/

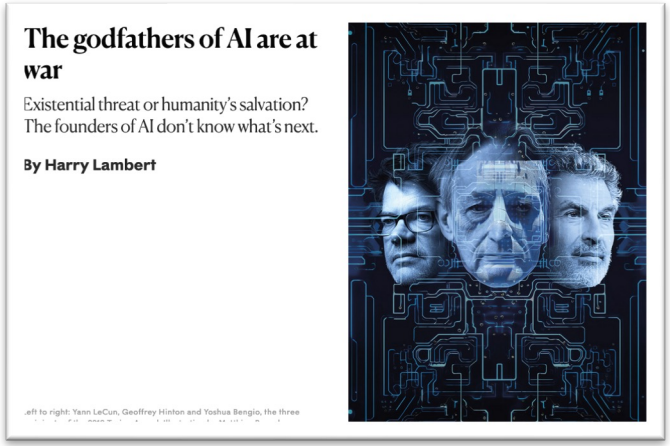
Amazon scraps secret AI recruiting tool that showed bias against women

Wat nu?

- We spreken momenteel veel over algoritmen, systemen en verantwoordelijkheid.
- Volgens mij gaan deze discussies eigenlijk niet over computersystemen, maar over **hoe wij als mensen met beslisvorming omgaan**.
- Wij (informatici en niet-informatici) moeten heel goed opletten hoe deze discussies gevoerd worden.
- Niet-informatici durven vaak niet mee te doen aan technische/wiskundige discussies. Dat moet anders.

En…

- De AI-discussie neemt momenteel sensationele en bangmakende vormen aan
- Hier lopen veel belangen door elkaar heen
  https://www.nrc.nl/nieuws/2023/06/24/sterft-de-mensheid-uit-door-ai-dat-is-sciencefiction-a4168053
- Schadelijke beeldvorming


The godfathers of AI are at war

Existential threat or humanity's salvation? The founders of AI don't know what's next.

By Harry Lambert

eft to right: Yann LeCun, Geoffrey Hinton and Yoshua Bengio, the three

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Opnieuw: de kernvragen gaan om **hoe wij als mensen (en instanties) met beslisvorming omgaan**.

Algoritmes 101

Klassiek probleem: sorteren

- Kan op slimme en domme manieren worden gedaan
- Wat is efficiënt bij grote hoeveelheden?
- Wat is toepasbaar op alle problemen waarbij je een volgorde kunt vaststellen?

Bijbehorende waarden

- Efficiëntie
- Generalisatie
- Schaalvergroting
- Exact dezelfde aanpak hanteren

Zie ook Birhane et al. (2022), The Values Encoded in Machine Learning Research, https://dl.acm.org/doi/10.1145/3531146.3533083

Bijbehorende waarden

- Efficiëntie
- Generalisatie
- Schaalvergroting
- Exact dezelfde aanpak hanteren

- Dit kan mensen ‘tot een nummer’ maken

Zie ook Birhane et al. (2022), The Values Encoded in Machine Learning Research, https://dl.acm.org/doi/10.1145/3531146.3533083

Bijbehorende waarden

- Efficiëntie
- Generalisatie
- Schaalvergroting
- Exact dezelfde aanpak hanteren

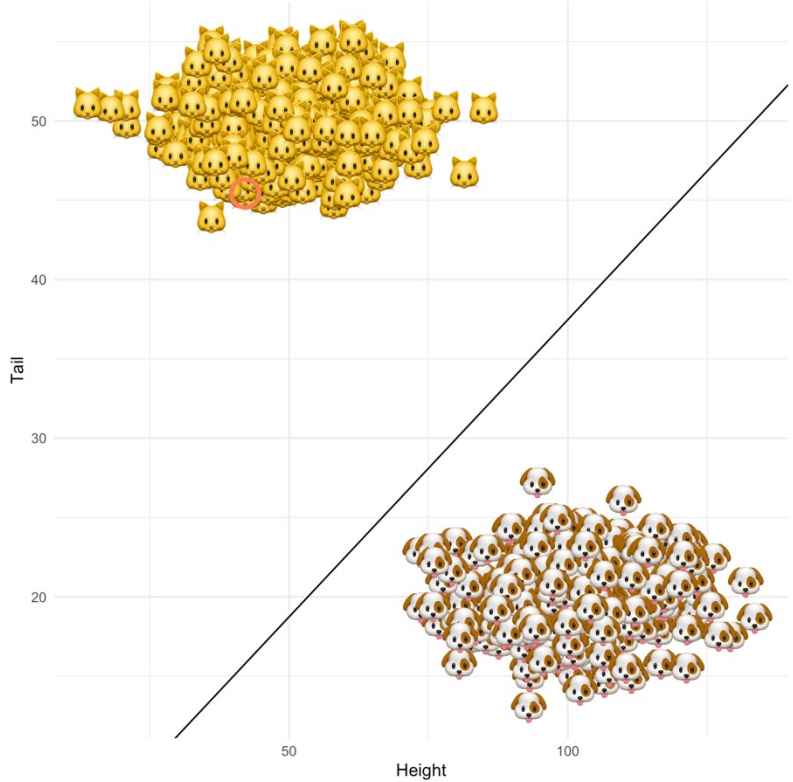
- Tegelijkertijd kan het ons een spiegel voorhouden

Precies zijn over stappen is moeilijk!

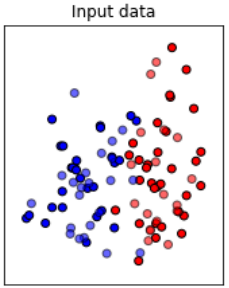
https://www.youtube.com/watch?v=cDA3_5982h8

Machine learning 101

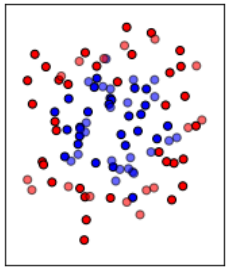
Patronen leren uit data en labels

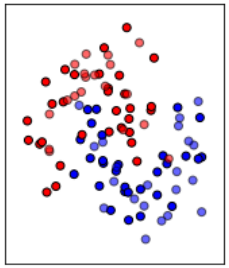


Patronen leren uit data en labels

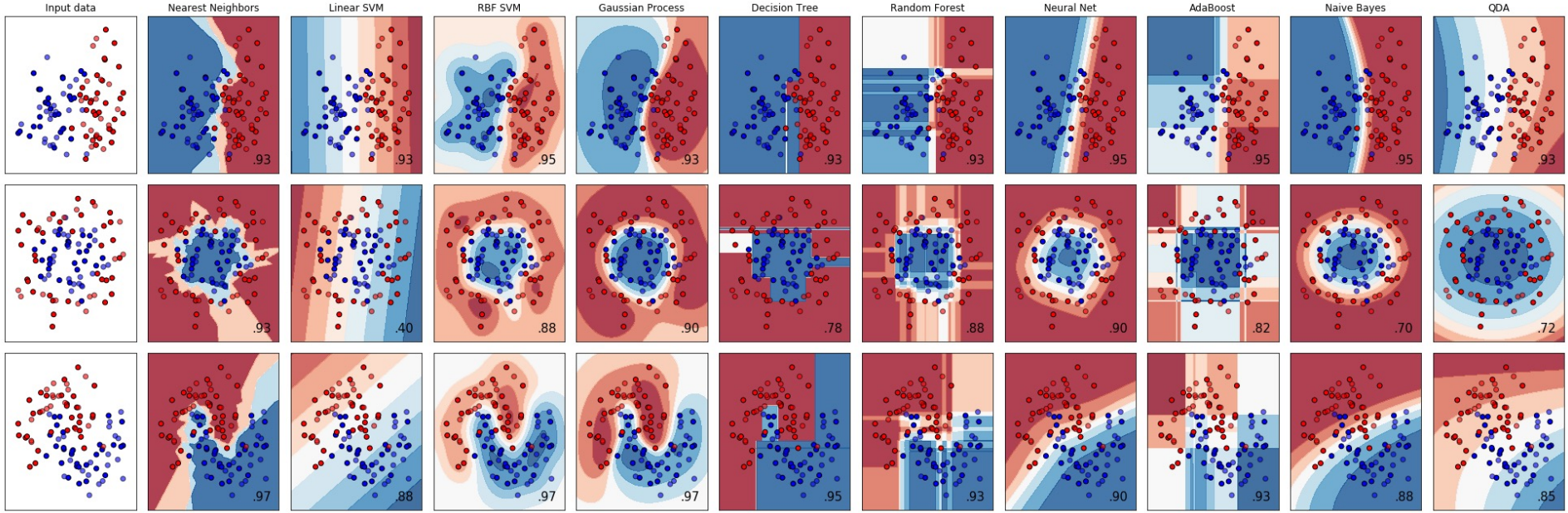


Patronen leren uit data en labels

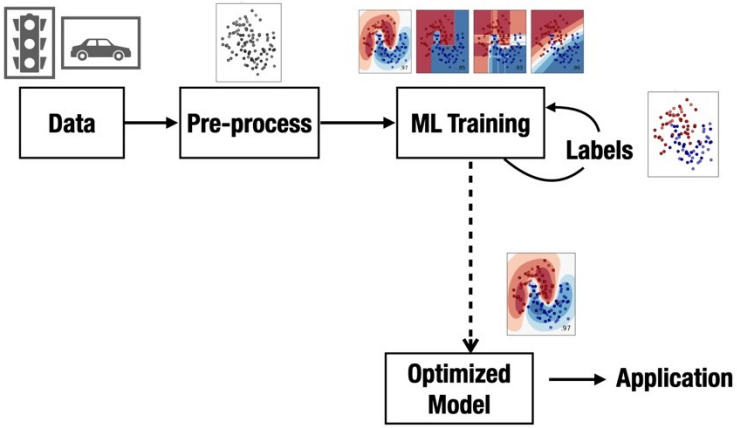




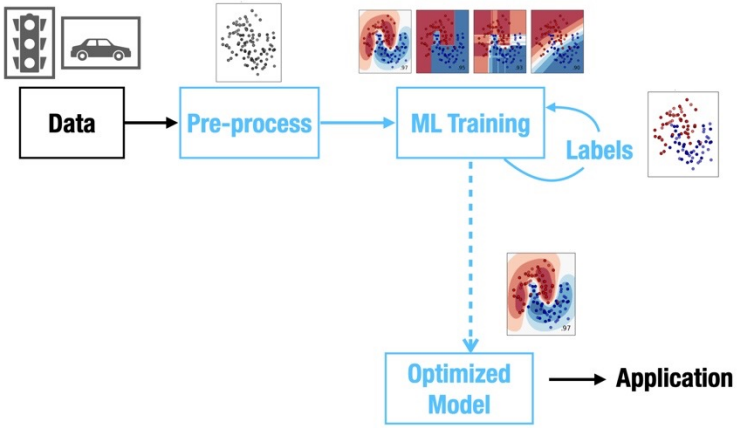
Patronen leren uit data en labels



De opzet

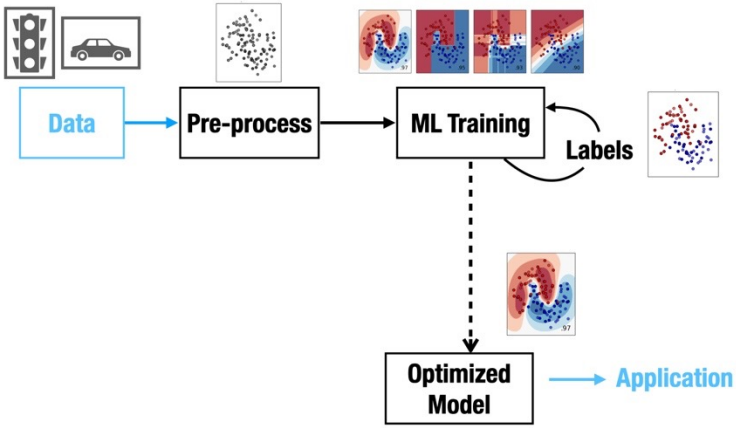

Data
Pre-process
ML Training
Labels
Optimized Model
Application

Focus van machine learning-expert



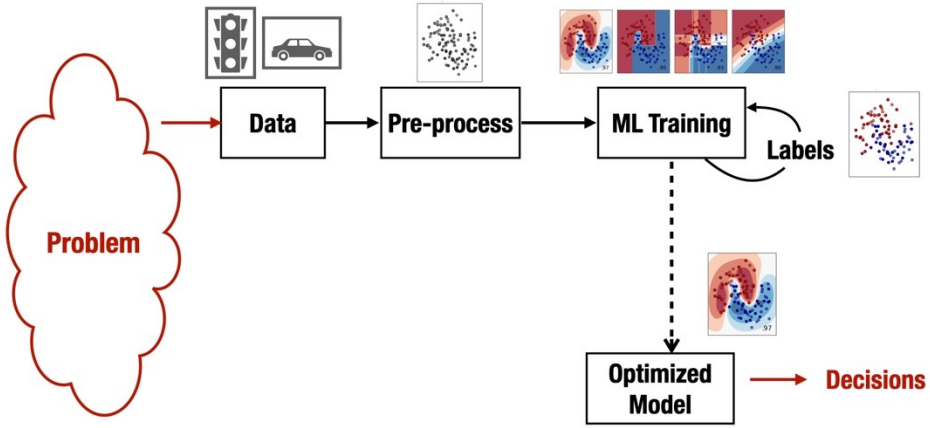
- Wiskundig optimale oplossingen
- We vertrouwen erop dat de data en toepassing geldig zijn

Focus van domeinexpert



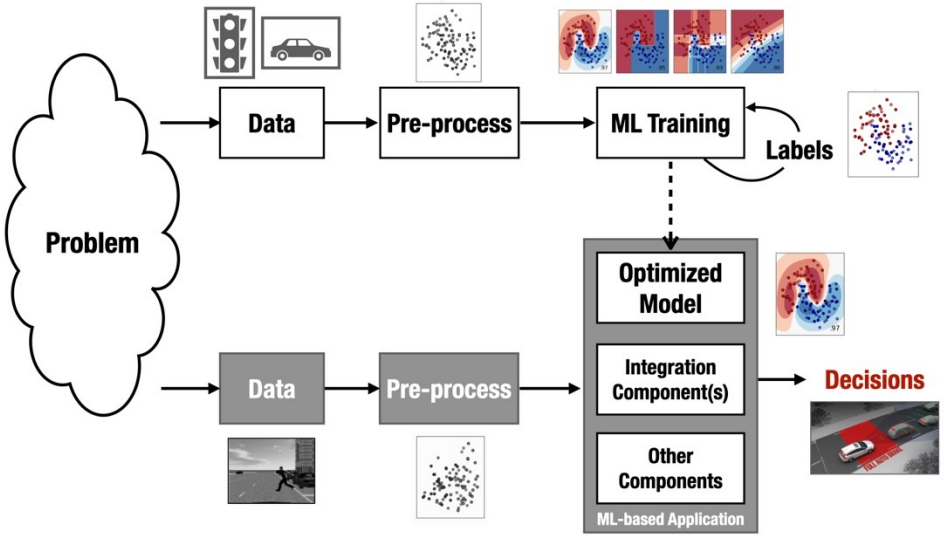
- Zijn de juiste modaliteiten, metingen, kenmerken meegenomen?
- Welk wereldbeeld optimaliseren we?

Waar het eigenlijk om gaat



- Welk probleem willen we werkelijk oplossen?
- Is dit probleem goed ‘vertaald’?
- Trekken we de juiste conclusies?

…of eigenlijk gaat het hierom



- Socio-technische systemen met veel componenten
- Moeilijk een enkele verantwoordelijke aan te wijzen

Patronen als pijnlijke, maar realistische spiegel

Amazon scraps secret AI recruiting tool that showed bias against women

av_nationaliteit	1, _MISSING_, _UNKNOWN_	8
	0	-1

- Patronen worden in context problematisch
- Let hierop: niet meten kan ook ruimte geven voor wegduiken

Hoe ‘ziet’ een computer de wereld?

Ruwe multimediadata: heel veel getallen

- 44.1 kHz audio:
  44,100 metingen per seconde
- Voor spraak is 8 kHz voldoende. Dat gaat nog steeds om 8000 metingen per seconde. In 45 minuten heb je dan 21.6 miljoen metingen…

← Dit portretje: 224 x 224 x 3 pixels = 150,528 intensiteitswaarden

Wat is het dichtst bij?

Accuraat kunnen voorspellen ≠ objectieve natuurwetenschappelijke waarheid

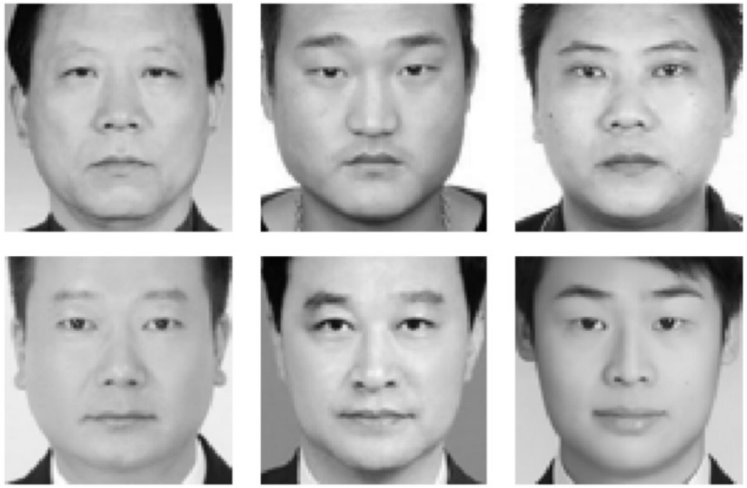


Figure 3. Wu and Zhang's "criminal" images (top) and "non-criminal" images (bottom). In the top images, the people are frowning. In the bottom, they are not. These types of superficial differences can be picked up by a deep learning system.

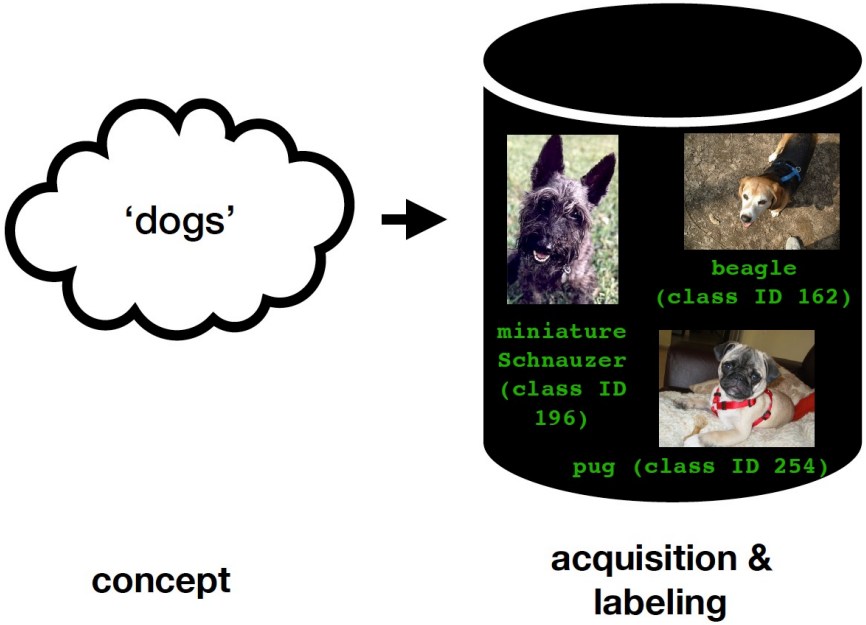
Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.

Article | Open Access | Published: 31 March 2023

Using deep learning to predict ideology from facial photographs: expressions, beauty, and extra-facial information

https://arxiv.org/abs/1611.04135

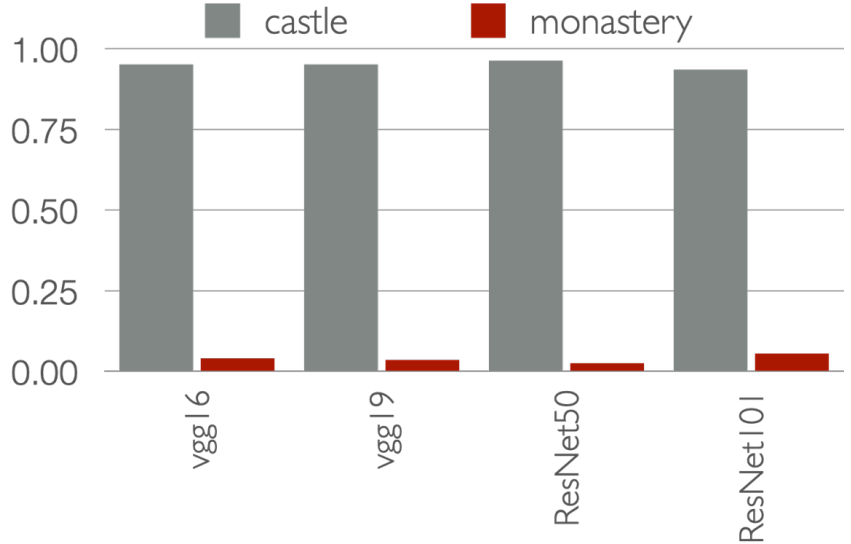
Classificeren en categoriseren



‘Ground truth’: het correcte label

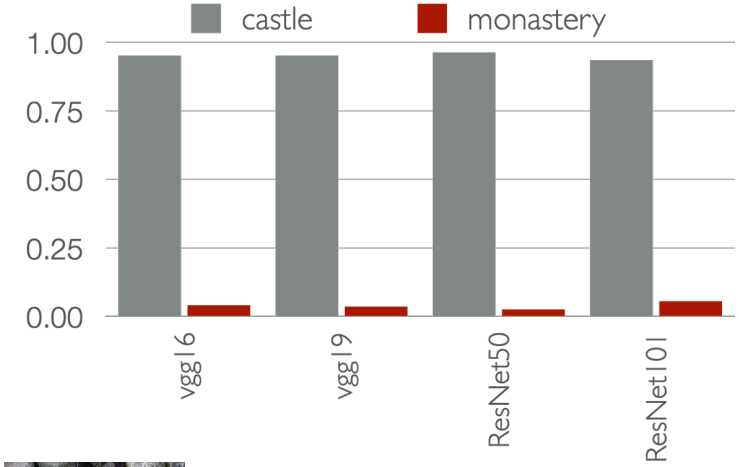
- Is er een kasteel in dit plaatje?

Is er een kasteel in dit plaatje?



Cynthia C. S. Liem and Annibale Panichella, “Oracle Issues in Machine Learning and Where to Find Them,” Proc. RAISE 2020.

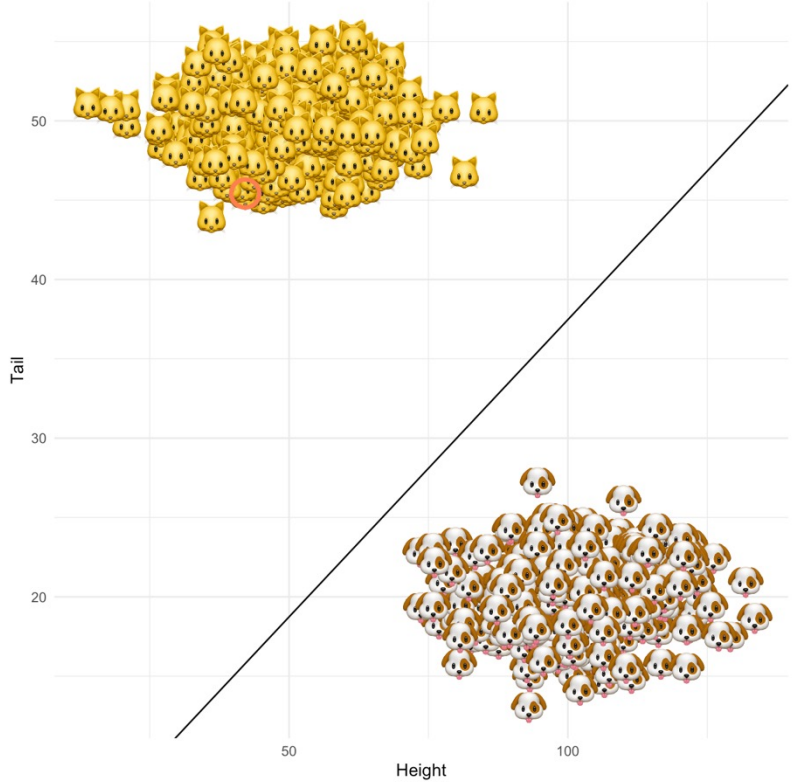
Implicaties van de representatie



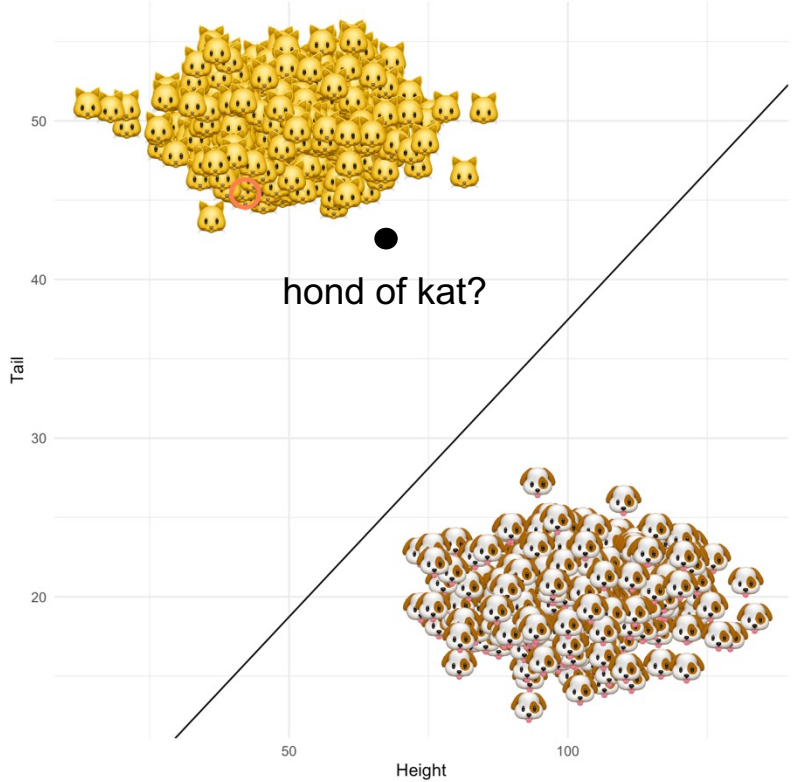
Cynthia C. S. Liem and Annibale Panichella, “Oracle Issues in Machine Learning and Where to Find Them,” Proc. RAISE 2020.

Wat doen we met patronen die we vinden?

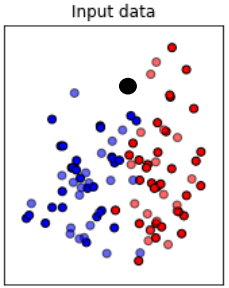
Terug naar de honden en katten



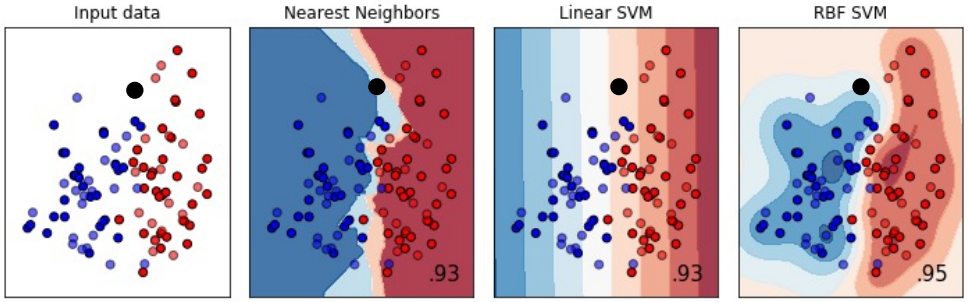
Ongeziene datapunten voorspellen



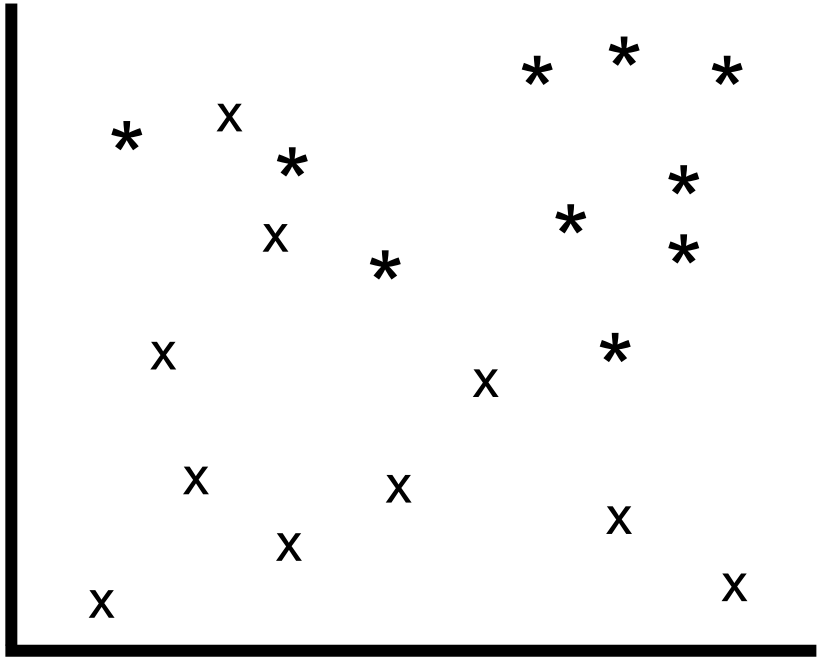
Patronen leren uit data en labels



Patronen leren uit data en labels

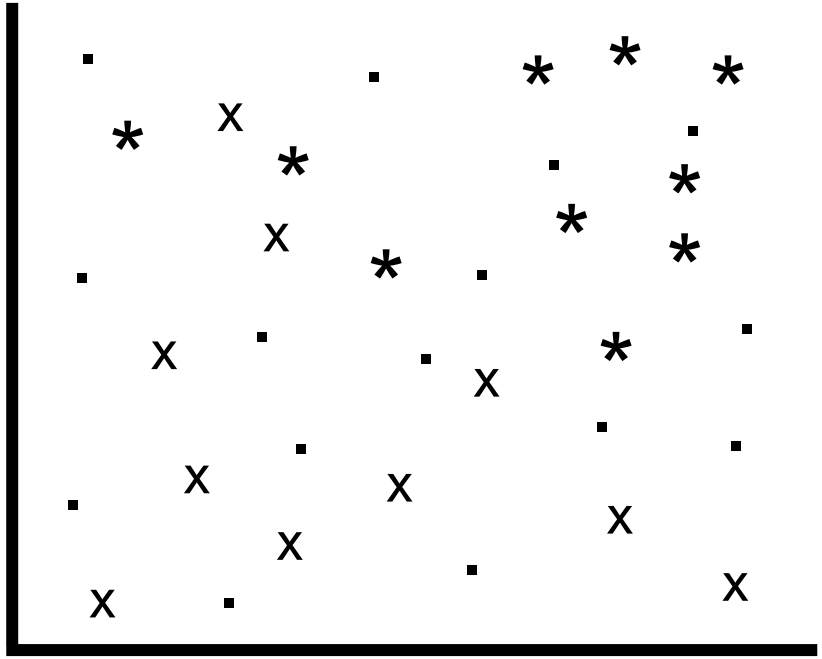


Beslissingen nemen en prioriteren



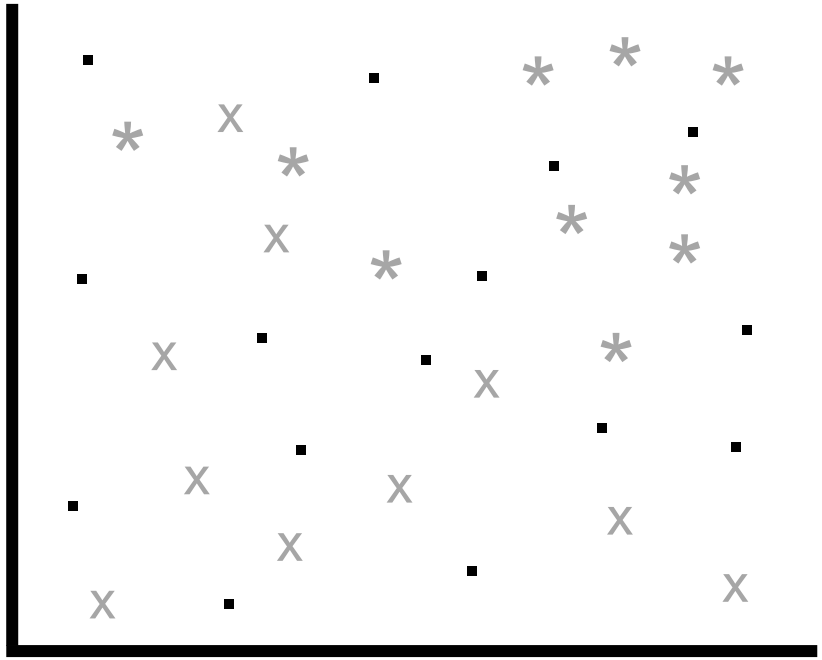
Beslissingen nemen en prioriteren

- Kies 5 datapunten



Beslissingen nemen en prioriteren

- Wat als die datapunten over mensen gaan?



Beslissingen nemen en prioriteren

- Wat als die datapunten over mensen gaan?

https://tudelft.fra1.qualtrics.com/jfe/form/SV_6sqrujNkBAT8N2m

Drie keer hetzelfde?

Drie keer hetzelfde?

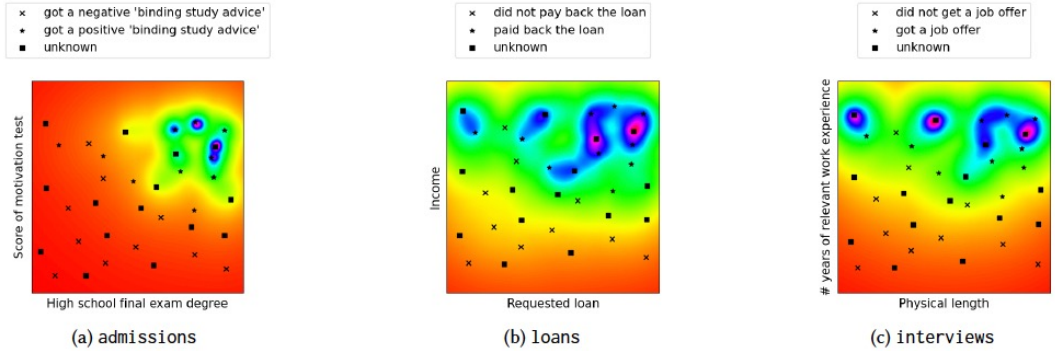


Fig. 6. Aggregated click heatmaps for computer science/data science audiences (python, dataScience and students combined).

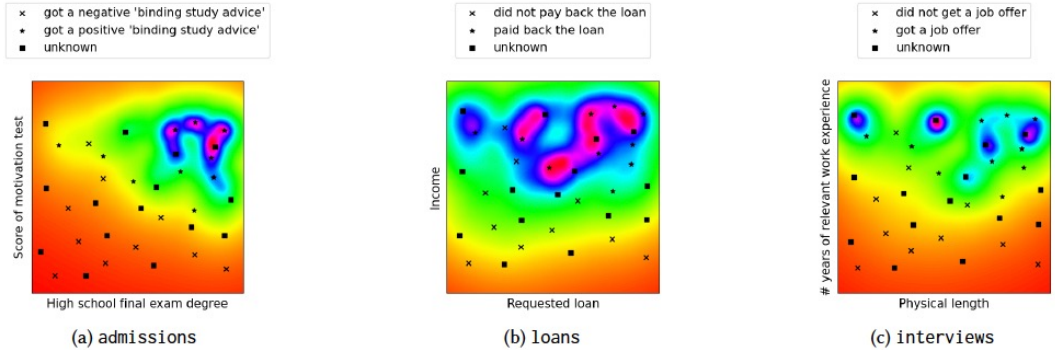
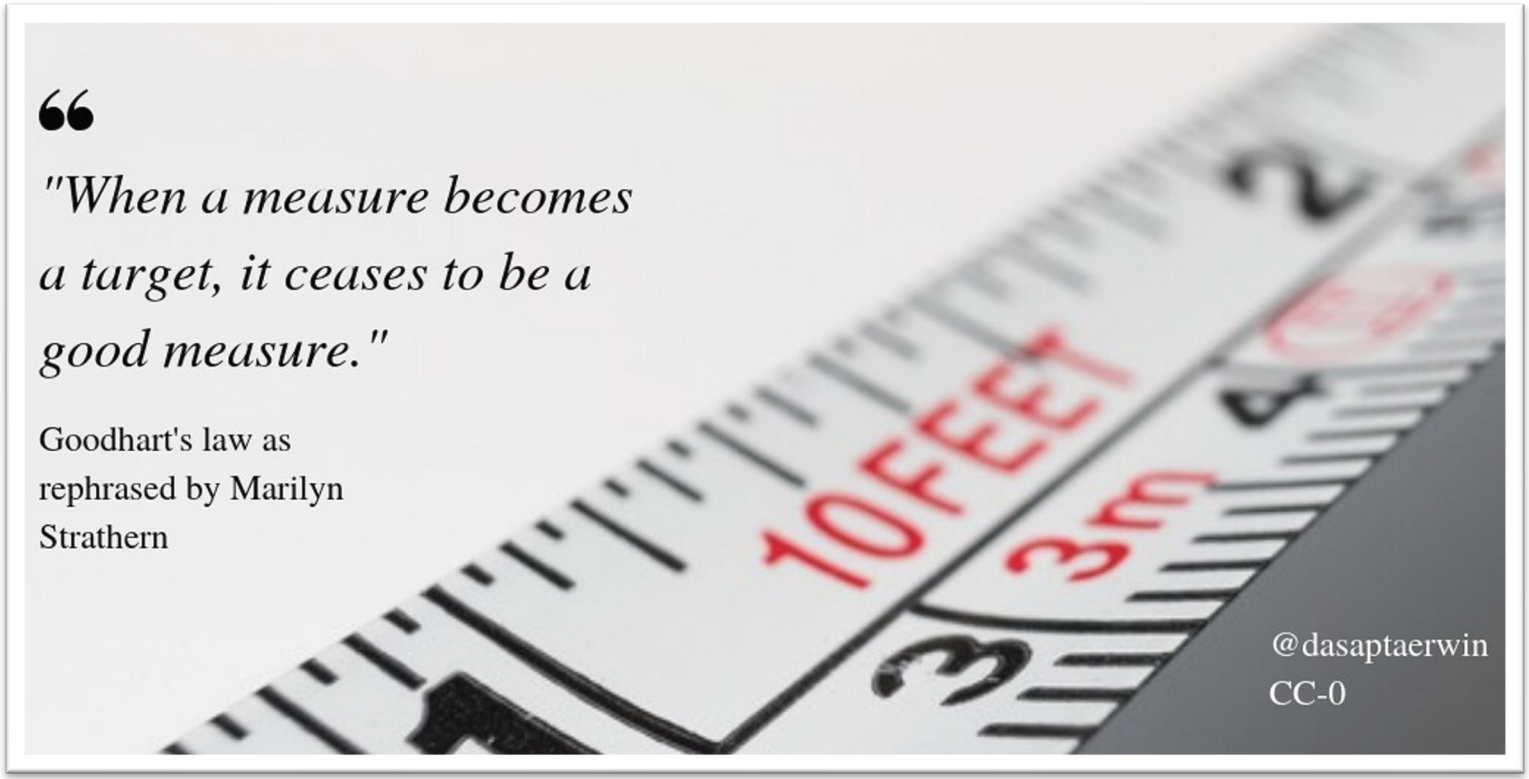


Fig. 7. Aggregated click heatmaps for broader audiences (rotary, policy and librarians combined).

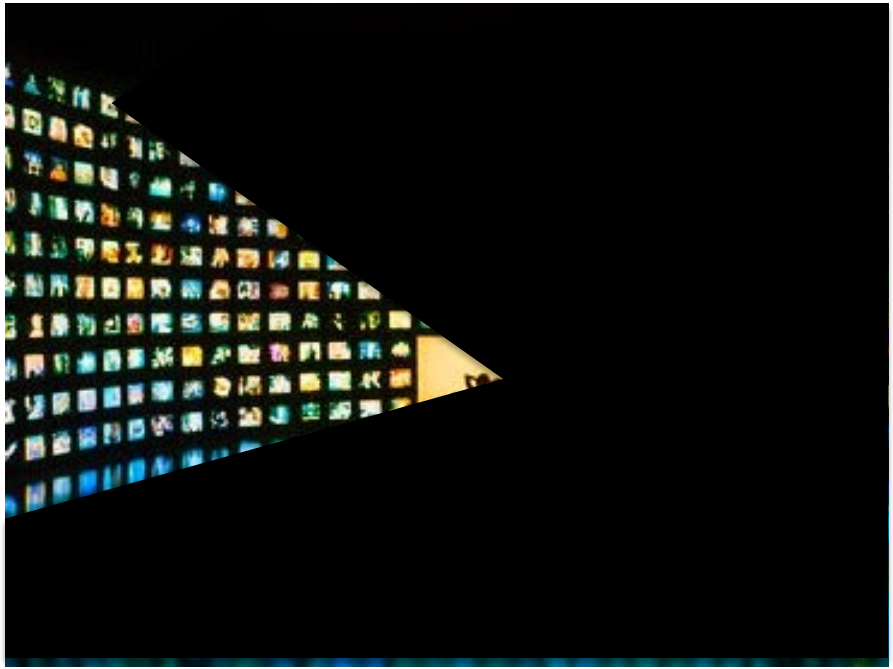
Het goede antwoord?

Perverse prikkels?



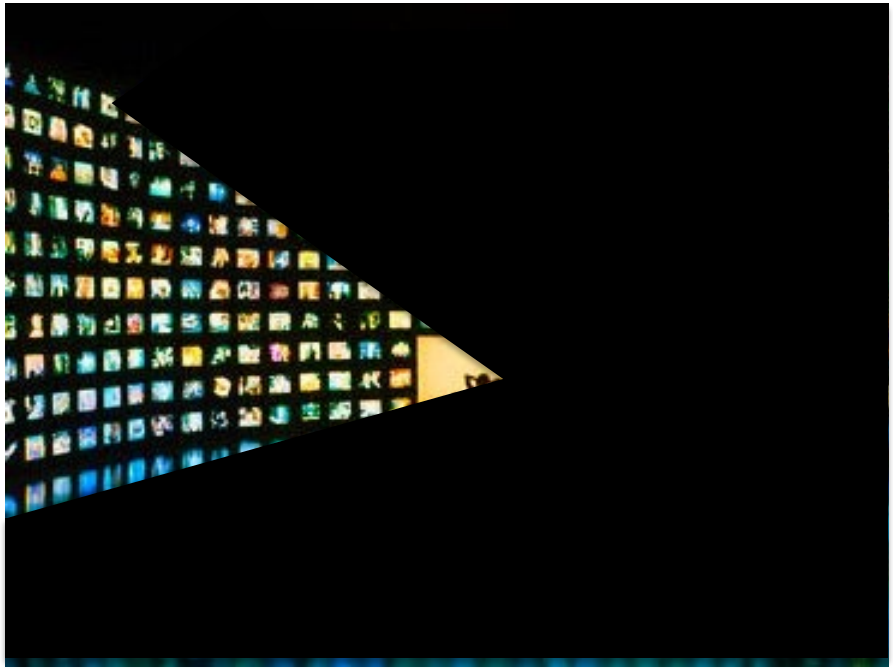
Een lang bestaand ideaal

Een lang bestaand ideaal



Een lang bestaand ideaal

relevant



Een lang bestaand ideaal

Een lang bestaand ideaal

irrelevant

Een lang bestaand ideaal

voorspellingsfout

irrelevant

Een lang bestaand ideaal

voorspellingsfout

irrelevant

oninteressant

Een lang bestaand ideaal

voorspellingsfout

niet de tijd waard

irrelevant

oninteressant

Een lang bestaand ideaal

voorspellingsfout

niet de tijd waard

irrelevant

oninteressant

niet-bestaand

Een lang bestaand ideaal

voorspellingsfout

niet de tijd waard

irrelevant

oninteressant

niet-bestaand

onbereikbaar

‘Eerlijkheid’

- ‘Eerlijkheid’ / fairness in AI is momenteel een hot topic
- Er zijn echter heel veel mogelijke definities van ‘eerlijkheid’, en wiskundig zijn ze met elkaar in conflict
- Ook hier moet je dus scherp zijn op wat je eigenlijk bedoelt!

Tutorial: 21 fairness definitions and their politics

https://www.youtube.com/watch?v=jlXluYdnyyk

Hoe omgaan met fouten en risico's?

- Tel je hoe vaak je het goed hebt?
- Of geef je aan hoe je zo robuust mogelijk met fouten omgaat?
- Wat weeg je af bij beperkte middelen?

Extreem voorbeeld

- Als we alle criminelen van Nederland willen vangen, kunnen we de hele bevolking arresteren. Dan zijn we 100% succesvol.

Belangen en beloning

- Waar worden verschillende stakeholders voor beloond, en wat zien ze als schade?
- Afreken- en sensatiecultuur in politiek en media

Belangen en beloning

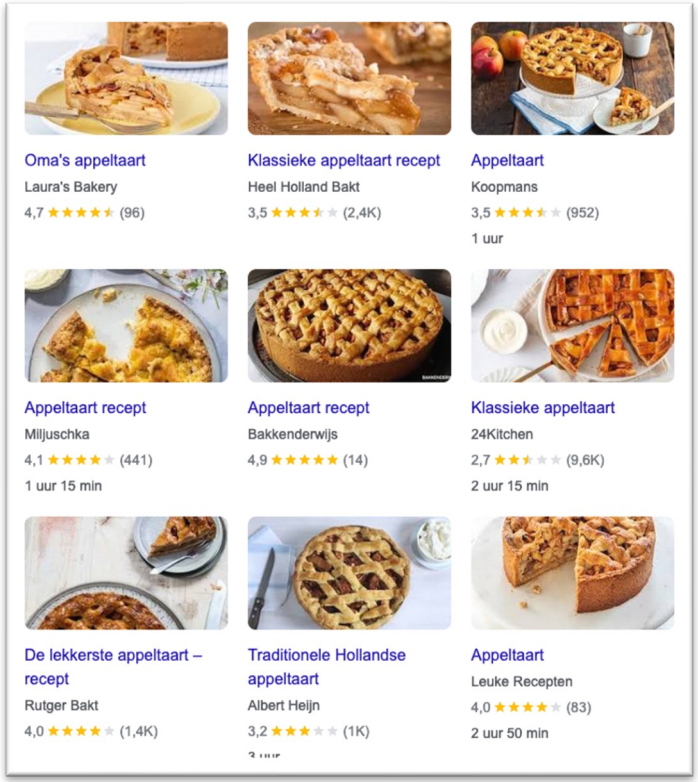
- Genuanceerde afwegingen nodig
  - Goed voorbeeld: afwegingen en moreel beraad in TU Delft rond Kennisveiligheid
    https://www.youtube.com/watch?v=KJTF0blzvkA
  - (morele) beginselen boven (praktische) gevolgen
  - aandacht voor schadebeperking en moreel leerproces

Een metafoor die mij vaak hielp

Algoritmes zijn als recepten

- Objectief repliceerbaar (als we willen)

https://www.youtube.com/watch?v=tnBcVwcoMYY



Algoritmes zijn als recepten

- Eindresultaat is niet ‘universeel optimaal’ of ‘universeel geldend’




Algoritmes zijn als recepten

- Machine learning als een Airfryer?
  Generatieve AI als een nieuwe technologie voor een kunstmatige smaakstof?
- Producent moet regels volgen (met toezicht)
- Burger kan niet meepraten over Airfryer-details, en zich toch positioneren
- En ook eigen verantwoordelijkheid nemen qua een ‘gezond’ dieet

Onze verantwoordelijkheid

- Gan we onze menselijke imperfectie te gretig uit de weg?

essay

Illustratie gemaakt door DALL.E, een programma van de makers van ChatGPT dat beelden genereert op basis van tekst. DALL.E werd gevraagd om van een zin uit dit essay een illustratie te maken in de stijl van Escher. De zin luidt: 'Silicon Valley snelt wel sneller een oppervlakkig efficiëntie-ideaal boven een grondigere gedachte over wat mensen nodig hebben'.
ILLUSTRATIE DALL.E

ChatGPT berooft ons van waardevol denkwerk

Sinds ChatGPT, een slim schrijvende chatbot, eind vorig jaar zijn intrede deed, gaat het vooral over plagiaat. De verkeerde discussie, vindt Cynthia Liem, docent kunstmatige intelligentie en pianist. Stel de vraag waarom moeite doen zo uit de gratie is?

tekst **Cynthia Liem**

Aan het begin van de jaren 10 woonde ik een aantal maanden in Californië, in Silicon Valley, waar ik stage liep op het hoofdkwartier van Google. Elke vrijdagmiddag was er een gezamenlijke werkbijeenkomst: een borrel met snacks, drankjes en nieuwtjes over het bedrijf. En soms een presentatie over een project waar binnen de muren van Google aan werd gewerkt.

Daar hoorde ik voor het eerst over virtuele assistent-technologie. Veel mensen maken in hun dagelijks leven intensief gebruik van Google-producten. Ze mailen met Gmail, houden hun afspraken bij in Google agenda en navigeren naar hun werk met Google Maps. Het idee was om hun gedrag te analyseren en te koppelen aan een functie die automatisch herinneringen op hun telefoon geeft.

Jubelend was het projectteam over het idee dat de telefoon zo een slimme virtuele assistent zou worden, die al wist wat eraan nodig had voordat diegene dat zelf doorhad.

Deze technologie kwam in veel Google-producten terecht, in de vorm van Google's Assistent-functie op Androidtelefoons. Zo kan je nu automatisch een waarschuwing krijgen dat een vriend bijna jarig is, of dat er een zware avondspits aankomt en je dus beter wat vroeger van kantoor kunt vertrekken.

Ik herinner me deze presentatie nog goed, omdat die gemengde gevoelens bij me opriep. Natuurlijk zijn dit soort meldingen handig. Vaak genoeg ben ik druk bezet, en word ik in de chaos van alles wat er moet gebeuren minder scherp op wat mijn volgende taken en afspraken precies zijn. Maar ik merk ook dat ik gemakzuchtiger en onopletender word als technologie mij te veel werk uit handen neemt.

Ook toen al. Als onbekende in Californië reed ik op automatische navigatie, tot ik me realiseerde dat ik veel heel erg achter een pijltje aan reed. Ik had geen fluwe idee waar ik was. Om mezelf uit te dagen, probeerde ik

ZATERDAG 4 MAART 2023 TROUW TIJDGEEST — 41

https://www.trouw.nl/tijdgeest/chatgpt-berooft-ons-van-waardevol-denkwerk-wanneer-zijn-we-gestopt-met-ergens-moeite-voor-doen~bda521b4/

Meepraten over betrouwbare en verantwoorde AI

Wat moet ik weten?

Cynthia C. S. Liem
Multimedia Computing Group, TU Delft
@cynthiacsliem@akademienl.social.nl
c.c.s.liem@tudelft.nl

Verdere links

- Nog een aantal presentaties/commentaren over risicomodellering en het toeslagenschandaal:
  - https://tudelft.zoom.us/rec/share/gipC7rMIuoijUv82eNEpGKYreTHQyTLM0q3Kv7AX3sIBr-jhuo2mvyeAf-Vk7M6k.HGp0V_X-4-puKPFa?startTime=1648047506000
  - https://www.vpro.nl/argos/media/luister/argos-radio/onderwerpen/2021/In-het-vizier-van-het-algoritme-.html
  - https://www.vpro.nl/argos/lees/onderwerpen/algoritme/2023/fraude-algoritme-op-de-snijtafel-rotterdam-verdenkt-vooral-jonge-moeders.html
  - https://www.youtube.com/watch?v=t_Mic6Q5A_M