**OVERIG MATERIAAL (BIG) DATA
(Crash course Four)**

Wij bieden crashcourses aan op [www.tict.io](http://www.tict.io). Deze zijn in het Engels. Daarnaast hebben we in het verleden veel materiaal gemaakt in het Nederlands, dit is allemaal verzameld in dit word-document.

In dit document staat alles rond crash course vier, Big Data. Het is verzameld (oud) materiaal, dus we raden je vooral aan om lekker te grasduinen!

De Engelse geupdate online course staat hier: [www.tict.io/course-four](http://www.tict.io/course-four).

**Introductie**
Op technofilosofie.com doen we aan **toegepaste technofilosofie.**Het is ons doel om zoveel mogelijk mensen na te laten denken over de impact van technologie. Wij denken namelijk dat iedereen dat zou moeten doen. Iedereen zou een beetje een technofilosoof moeten zijn. Of je nu technologie ontwerpt, programmeert, implementeert of gewoon gebruikt. Of je nu een professional bent, of een vader, een coach, een vriend, vakantieganger of gewoon een mens.

Moderne digitale technologie rukt op en wij moeten leren om ons daartoe te verhouden. Het gaat daarbij niet om een abstract gegeven, maar gewoon om onze relatie alledaagse, hedendaagse technologie. Smartphones, e-mail, platformen zoals Uber, Tinder, sociale media, sensoren, slimme meters, en ga zo maar door.

**DAT IS ONZE MISSIE: ZOVEEL MOGELIJK MENSEN ZO GOED MOGELIJK TE LATEN NADENKEN OVER DE IMPACT VAN TECHNOLOGIE**.

We bieden daarom ook materiaal aan, zoals in dit word-document. Hierin staan verwijzingen naar artikelen, video's e.d. De tekst valt onder creative commons, maar de afbeeldingen, verwijzingen, clips en video's niet altijd. Het is dus belangrijk dat je je daar zelf van op de hoogte stelt, als je materiaal wilt gebruiken. Bij twijfel, check de [website](https://creativecommons.nl/) van creative commons.

**Onderwerpen in dit document:**

1. **Een long read met 7 onderwerpen;**
2. **Bijlage A, met dingen om van te genieten & achtergrondinformatie**

**LONGREAD**

De volgende onderwerpen worden behandeld:

1. Wat is (big) data eigenlijk? En waarom schrijven zoveel mensen het met hoofdletters?
2. Wat zijn de voordelen van (big) data?
3. Wat zijn de nadelen van big data? Waarom is data-driven een slecht idee? Tien problemen!
4. Hoe zit het met privacy?
5. Wat is Quantified Self en hoe zit het met de voor,- en nadelen met dat onderwerp?
6. Wat is de dataficatie van onze maatschappij en spionagekapitalisme?
7. Wat zijn de tips voor de toekomst?

Hier staat de [PowerPoint](https://maken.wikiwijs.nl/136737/Technofilosofie___Aandacht#!page-5633958) van deze onderwerpen.

Uiteindelijk, als je alles gelezen hebt, laten we je – hopelijk – verward achter. Of wijzer. Maar misschien ben jij zo iemand, die geen zin heeft in al die tekst. Bekijk dan de onderstaande video. Die is minder uitgebreid dan de tekst, en alweer een beetje verouderd, maar je bent wel in  dik 5 minuten klaar!

Link: <https://youtu.be/WSnULs1TWQs>

**(1) Wat is (big) data eigenlijk? En waarom schrijven zoveel mensen het met hoofdletters?**

Om te beginnen, de grote vraag: waarom schrijven we toch altijd (big) data met hoofdletters? In deze long read doe ik dat bewust niet, zeker omdat niemand me kan vertellen waarom dat zou moeten. Verder lijkt het me niet zo belangrijk. Mocht je het niet met me eens zijn, dan kun je er [hier](http://english.stackexchange.com/questions/161823/should-big-data-be-capitalised) (discussieforum op English Language – site) lekker over discussiëren. Daarnaast is het je wellicht al opgevallen dat ik (big) consequent tussen haakjes schrijf. In de vorige zin ook. Superconsequent dus. De reden daarvoor is dat discussies over de voor,- en nadelen van data en de ethiek van data zowel horen te gaan over data die wordt gezien als big als data die niet onder die definitie valt.

Ik leg dat later nog wel uit. Hoop ik.

Er is overigens geen vaststaande definitie van big data. Er zijn natuurlijk wel pogingen gedaan. Een manier die vaak gebruikt wordt om te bepalen of iets big data is, is om te kijken of de data voldoet aan de [5 V’s](https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know)‘ (artikel via LinkedIn):

–  Het gaat allereerst – verrassing – om veel data (**Volume**);
–  Het gaat om data die snel wordt geproduceerd en snel verandert (**Velocity**);
–  Het gaat om veel verschillende type data, vaak ongestructureerd (**Variety**);
–  Het gaat om de betrouwbaarheid (**Veracity**);
–  Het gaat er om dat je er waarde uit kunt halen (**Value**).

Je zou ook kunnen zeggen (of misschien wel beter kunnen zeggen) dat [big data een ontwikkeling](https://decorrespondent.nl/296/wat-is-big-data/14414312-a609db7d) (artikel, De Correspondent) is. Het wordt steeds eenvoudiger (en goedkoper!) om data te verzamelen, bewaren en bewerken en de statistiek (en algoritmes) worden steeds beter. Maar óók als je data niet voldoet aan één van deze v’s dan is het nog steeds belangrijk om goed te kijken naar de voor,- en nadelen. Ook data zonder het predicaat big heeft zo zijn voor,- en nadelen. Laten we – zoals beloofd – beginnen met de aan de horizon glorende voordelen.

**(2) Wat zijn de voordelen van (big) data?**De volgende voordelen worden hieronder besproken:

1. Het denken in digitale tweelingen;
2. Het voorspellen (vooral van de toekomst);
3. Het mogelijk maken van objectieve besluitvorming;
4. De brandstof voor Machine Learning.

A. Het denken in Digitale Tweelingen
Er komt steeds meer data. Vrijwel alles wat je doet laat een dataspoor na en dat wordt alleen maar meer. Denk maar aan al die e-mails, documenten, posts, likes, appjes en ga zo maar door. Inmiddels hebben ook veel apparaten een internetverbinding. We noemen die apparaten dan vaak smart. Geen idee waarom, alleen een internetverbinding maakt iets nog niet smart. Sterker nog, ik ken hele domme apparaten met een internetverbinding, zoals de [Hapifork](https://www.hapi.com/product/hapifork) (website van het product). Deze houdt bij hoe snel je eet, en als je te snel eet gaat de vork trillen. Smart! Je auto heeft een internetconnectie, je tandenborstel, je weegschaal, je energiemeter en steeds vaker je speelgoed, wat weer leidt tot intrigerende verschijnselen als een [smart Barbie.](https://www.theguardian.com/technology/2015/mar/13/smart-barbie-that-can-listen-to-your-kids-privacy-fears-mattel) (Artikel in The Guardian). Langzaam krijgen alle apparaten een internetverbinding,  zijn er overal camera’s en ga zo maar door.

Naast al die apparaten meten we ook steeds meer aan ons zelf. We noemen dat de Quantified Self – beweging. Het gaat dan nu nog meestal om aantal stappen, hartslag en slapen, maar het is de voorspelling dat we steeds meer gaan meten.

Al die data samen zorgen ervoor dat we anders naar onze ‘werkelijkheid’  kunnen kijken. We zien een wereld met onze ogen, maar we kunnen ook naar een digitale afdruk van die wereld kijken. **De digitale tweeling.** Dat betekent dat we er een perspectief bij krijgen en dat biedt vele mogelijkheden om onze wereld beter te begrijpen. Mijn favoriete voorbeeld, is van de data van [Pornhub](https://www.pornhub.com/insights/2017-year-in-review) (2017 year in review op Pornhub – let op voor je werk!). Daar kun je zien dat Russen daadwerkelijk snel klaar zijn.

Ook zie je op Pornhub dat bij de 5 meest gezochte termen van 2018 vermeld staan: Mom, Stepmom en Milf. Waarmee, bijna 100 jaar na zijn dood, het Oedipuscomplex van Freud in de data bewezen lijkt. Veel mannen willen daadwerkelijk sex met hun moeder.

Meer data is dus vaak beter. Je ziet nieuwe dingen, zoals [Kenneth Cukier](https://www.ted.com/talks/kenneth_cukier_big_data_is_better_data?language=nl) in zijn Ted Talk betoogt. Hij geeft daar – tussen zijn irritante manier van ademen door – een fascinerend voorbeeld over appeltaart. Lang is gedacht dat appeltaart de favoriete taart van Amerika was. Immers, het werd het meest verkocht. Tot dat de supermarkten ook kleine éénpersoonstaarten gingen verkopen. Toen bleek appeltaart helemaal niet zo populair. Appeltaart bleek vooral die taart, die iedereen wel lekker vond en die je dus kocht als er alleen maar grote taarten waren. Overigens kent Kenneth aan zijn kledingstijl te zien, ook de term Big Suit. Met hoofdletters! Zonder haakjes.

Dit zou dus ook kunnen betekenen dat de grote problemen van onze wereld (klimaat, gezondheid, vervuiling, eten, etc…) veel beter kunnen worden aangepakt als je over grote hoeveelheden data beschikt om deze problemen te begrijpen. Gezien de omvang van deze problemen is (big) data misschien wel de énige manier om deze data aan te pakken. Of misschien wel niet, maar dat zien we later.

Nog een voorbeeld, uit Amerika, waar ze via een app op een smartphone kuilen in de weg detecteren. De app voelt de schok, registreert dat, en alle gegevens worden slim vertaald in een opdrachtenbon voor de afdeling gaten vullen van de wegendienst. Maar misschien ook niet.

B. Het voorspellen (vooral van de toekomst)
Als je veel data hebt, en je begrijpt dingen, dan kun je ook de toekomst gaan proberen te [voorspellen](http://www.marketingfacts.nl/berichten/predictive-analytics-kijken-door-de-glazen-databol) (artikel van Marketingfacts). Dit kun je doen op een macro-niveau. Je kunt bijvoorbeeld voorspellen waar er veel misdaad zal plaats vinden of waar het gaat branden en op die manier je politie,- of brandweer- corps slimmer inzetten. Bij Fontys kunnen we (bijna) voorspellen wat de impact is van het weer (regen!) op het aantal studenten dat naar school komt. Op die manier kun je voorspellen wat de drukte zal zijn, en je inkoop van broodjes in de kantine (bijvoorbeeld) daarop aanpassen. Restaurants kunnen veel beter voorspellen wanneer het druk is, maar ook wat mensen op welke dagen en bij welk weer bestellen en op die manier inkoopvoordelen halen. Als je weet hoe mensen door de winkel of de winkelstraat bewegen, kun je die slimmer inrichten.

Maar je kunt ook voorspellen op micr0-niveau. Als je veel data hebt over de normale houding van een chauffeur, dan weet je dus ook wanneer een chauffeur in slaap dreigt te vallen en kan de auto ingrijpen. Als je veel data hebt over hoe iemand online winkelt, kun je ervoor zorgen dat hij of (meer waarschijnlijk) zij precies de aanbiedingen krijgt die bij de persoon passen. Er is steeds meer informatie, de tijd die je hebt is beperkt, dus filters die precies laten zien wat jij wilt weten zijn van groot belang. Je kunt gedrag zien bij een  student dat duidt op uitval, en tijdig ingrijpen.

En zo zijn er ontzettend veel mogelijkheden.

C. Het mogelijk maken van objectieve besluitvormingDoor middel van (big) data kun je – althans in theorie – ook objectievere besluiten nemen. In plaats van dat je iemand laat beoordelen door een subjectief iemand die slechts beschikt over een beetje gegevens en veel gevoel, laat je de besluiten nemen op basis van veel data en geen gevoel. Op die manier kun je bijvoorbeeld kijken of iemand bij je bedrijf past. Of bij je verzekering. Of dat een student bij een numerus-fixus-opleiding wel of niet toegelaten moet worden. Je kunt data ook mee laten wegen bij strafbepaling. En ga zo maar door. Er zijn veel mensen die juist beweren dat dit leidt tot oneerlijkheid. Als de systeem bevooroordeeld is, dan is het algoritme ook bevoordeeld. Dat klopt, maar uit onderzoek blijkt ook, dat de algoritmes in ieder geval beter functioneren dan mensen.

Algoritmes hoeven niet perfect te zijn, alleen maar beter als mensen.

D. De brandstof voor kunstmatige intelligentie
(Big) data is de brandstof waarop machine – leren ontwikkeld kan worden. Het idee van machine – leren (een onderdeel van kunstmatige intelligentie) is dat je bepaalde basisregels programmeert en dat de computer daarna, op basis van enorme hoeveelheden data, zelf gaat leren. Elke keer als je zoekt op een afbeelding van een kat op Google, en je klikt op een kat bij Google Afbeeldingen, dan leert Google op de achtergrond hoe een kat er uit ziet. Machine – leren biedt enorme mogelijkheden. Denk aan het herkennen van kankercellen, zelfrijdende auto’s, chatbots en ga zo maar door. Hoe meer data je hebt, over hoe meer zaken, hoe meer je machine-leren kunt inzetten voor allerlei problemen. Tegenwoordig heb machines die helemaal zelfstandig (en ondoorzichtig) nieuwe patronen vinden of zelfs zaken verzinnen. Daarin veel meer in de [sectie](http://technofilosofie.com/kunstmatige-intelligentie/) over kunstmatige intelligentie.

Nog een voorbeeld. Stel je wil een computer Pac-Man leren spelen. Vroeger moest je de computer dan helemaal programmeren. Dit zijn de spookjes, die zijn gevaarlijk. Je moet de blokjes opeten. Als je er links in gaat, kom je er rechts uit. Dit zijn de beste looplijnen, etc…. Nu vertellen ze de computer alleen dat het de bedoeling van het spel is om zoveel mogelijk punten te halen. De computer begint, verliest, begint opnieuw, verliest, leert, verliest, leert en een half uur later heb je de beste Pac-Man speler aller tijden. Helaas werkt dat niet bij alle spellen. En waarom, dat is nog niet duidelijk.

Samenvattend
Oké, dus (big) data gaat ons helpen om meer inzicht te krijgen, de grote problemen aan te pakken, gedrag te voorspellen en ons te voorzien van de juiste informatie.  Het gaat de besluitvorming eerlijker maken en ons helpen allerlei processen te verbeteren middels machine-leren. Perfect! Ja, maar alleen als je je volledig bewust bent van de nadelen en tekortkomingen van data.

**(3) Wat zijn de nadelen van (big) data? Waarom is data-driven een slecht idee? Tien problemen én nog twee extra!**

De volgende tien problemen met (big) data (plus één extra) worden worden hieronder besproken:

1. Data is niet neutraal;
2. Je beïnvloedt wat je meet;
3. (Big) data. Big problems;
4. Snap je de correlatie wel?
5. Oorzaak en gevolg;
6. Algoritmes zijn vaak niet transparant en niet eerlijk;
7. De feedback ontbreekt;
8. Data kijkt naar het verleden…. duh…
9. Ik ben veel complexer dan duizend datapunten;
10. Roofdieren;

En, de bonus, (1) wat niet weet wat niet deert en (2) met statistieken valt alles aan elkaar te liegen.

Probleem 1. Data is niet neutraal
Er wordt vaak gedaan alsof (big) data er ‘gewoon is’. Maar dat is natuurlijk niet zo. Data is geen natuurverschijnsel. Data noemen we wel eens ge-geven. Maar het is ons niet ge-geven. Je moet het oogsten. Je moet er iets voor doen, je moet meten. Je moet data verzamelen. En dus maak je keuzes. Wat ga je meten? Wat kun je meten? Hoe ga je meten? Wanneer ga je meten? En die keuzes zijn bepalend. En verre van objectief.

Ik zou het liefst van mijn studenten weten hoe ze zich ontwikkeld hebben in vier jaar, maar dat kan ik niet meten. Daarom meten we rendement of studenttevredenheid of wat voor banen studenten krijgen, maar dat zijn **benaderingen**. En daarmee zijn het dus keuzes. En dus is data niet neutraal. Dat is niet erg, maar wel superbelangrijk om te weten wanneer je er daarna dingen mee gaat doen.

In zijn [Godwin – lezing](https://decorrespondent.nl/6692/de-datahonger-van-staten-en-bedrijven-zet-veel-meer-op-het-spel-dan-uw-privacy-alleen/184413960192-62861fec) (De Correspondent) stelt Maxim Februari op een heldere manier vragen bij de neutraliteit van data en daarmee het in zijn ogen misplaatste data – optimisme. Data heeft dus misschien de schijn van neutraliteit, maar data is verre van neutraal. Vooroordelen zijn ook onderdeel van keuzes om data te verzamelen. En omdat data geen natuurverschijnsel is, moet je er steeds kritisch naar kijken. Klopt die verzameling wel? Bijvoorbeeld: als we kijken naar de gebruikers van het spelletje Farmville, dan zien we dat er heel veel uit Beverly Hills komen. Dat is raar. Of is misschien 90210 gewoon de bekendste postcode in Amerika?

Data is eigenlijk net zo neutraal als de vragen die je stelt in een vragenlijst. Niet dus. Zoals Kate Crawford (Harvard) stelt, data is iets wat we creëren, maar ook bij elkaar fantaseren. Een ander voorbeeld. Rondom Hurricane Sandy werden 20 miljoen tweets gestuurd. Die kun je analyseren, maar zonder context trek je de [verkeerde conclusies](https://www.wired.com/insights/2013/05/more-data-more-problems-is-big-data-always-right/) (artikel op Wired.com). Immers, de hardst getroffen gebieden (Breezy Point, Coney Island) twitterden weinig simpelweg omdat ze te druk waren met de storm, minder geïnteresseerd zijn in twitter of geen batterij of verbinding (meer) hadden. In Manhattan, waar het wel meeviel, daarentegen werd volop getwitterd. De data wees er op dat de ramp zich vooral voltrok in Manhattan. De werkelijkheid was heel anders.

Nog een ander beroemd voorbeeld is Google Flu Trends (GFT). Op basis van zoekopdrachten op griep-gerelateerde symptomen kon Google al in 2008 griepepidemies volgen en voorspellen. Het werd het uithangbord van (big) data, maar later, toen het meer media-aandacht kreeg en de data toenam, zat Google er 140% naast en werd het het uithangbord van problemen met (big) data. Lees de analyse [hier](https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/) (artikel op Wired.com).

En nog één dan. Eerder gaf ik het voorbeeld van iphones die kuilen in de weg registreren die vervolgens gevuld kunnen worden. Lijkt niks mis mee. Maar wat bleek. In arme buurten gaan mensen vaak met de bus én hebben ze geen smartphone of wel wat beters te doen. Dat geldt ook voor buurten waar veel ouderen wonen. Het gevolg is dat juist in die buurten minder gaten gevuld worden.

Probleem 2. Je beïnvloedt wat je meetEen tijdje geleden, voordat Facebook in kwaad daglicht kwam te staan, vertelden ze trots dat ze door je likes te analyseren, precies wisten wie je was. Dat was creepy. Maar het was ook raar, want Facebook was ook de partij die bepaalde wat je te zien kreeg op Facebook. Dus, ze beïnvloeden ook wat je ‘liket’. Dit patroon zie je continu. Op het moment dat je iets meet, beïnvloedt je ook hetgeen dat je meet. Direct of indirect.

Weet je nog dat ik – bij de voordelen – zei, dat één van de voordelen van (big) data is dat je kunt voorspellen? Zo’n applicatie is bijvoorbeeld PredPol of Comstat. Die applicaties laten zien waar de meeste misdaad wordt gepleegd, en voorspellen dat met bewegende vierkanten op een kaart. Op die manier kan je de politie goed inzetten, en vermindert de misdaad. Klinkt goed, toch. Maar als we beter kijken, zien we dat vooral kleine misdaad goed te voorspellen is, want dat gebeurt veel, en daar is veel data van beschikbaar. Data is immers niet neutraal en niet altijd beschikbaar. Je kunt prima voorspellen waar in het openbaar gedronken wordt, geluidsoverlast is, wordt rondgehangen, in een auto ingebroken, geblowd, etc… Het is bijna niet te voorspellen waar ingebroken wordt, laat staan waar wordt verkracht of vermoord of gefraudeerd met de belastingdienst. PredPol en Comstat zorgen er dus voor dat de politie ingezet wordt op kleinere misdaad, en die vindt plaats waar armoede is. Daar worden dus meer overtredingen gesignaleerd en een negatieve spiraal wordt versterkt.

Je zou dus ook kunnen zeggen dat het werken met deze data, misschien onbedoeld, armoede criminaliseert.

Probleem 3. (Big) Data. Big Problems
[Apofenie](https://nl.wikipedia.org/wiki/Apofenie) is een verschijnsel waarbij mensen in willekeurige gegevens patronen herkennen en daar betekenis aan toekennen. Bijvoorbeeld een konijn in de wolken. Je weet dat dat toeval is, en niet God, die je iets wil vertellen middels een konijn. Een ander geweldig voorbeeld is dat in Amerika mensen continu overal de afbeelding van Jezus of Maria zien. In een olieplas, in gemorste latex en mijn favoriet: in geroosterd brood. Zeker 200 keer per jaar. Helaas kun je al sinds 2011 ook [toasters kopen, die Jezus in je brood drukken](http://www.dailymail.co.uk/news/article-2075932/Holy-toast-The-Jesus-toasters-flying-shelves-Christmas.html) (artikel Daily Mail), wat natuurlijk de magie wat minder maakt. Apofenie wordt ook gebruikt in de [Holzman Inktblot Test](https://nl.wikipedia.org/wiki/Holtzman_Inkblot_Test) (Wikipedia site) waarbij mensen dingen moeten herkennen in inktvlekken.

Apofenie zie je ook Big Data. Sterker nog als je maar genoeg data hebt, dan vind je altijd en overal wel [overtuigende correlaties](http://scienceblogs.com/insolence/2014/05/12/a-lesson-about-correlation-and-causation/) (artikel op scienceblog) dat zegt echter niets over causaliteit. Even voor de niet-statistici onder ons (zoals ik). Correlatie is als twee variabelen op een ordelijke manier een samenhang vertonen. Causaliteit is een oorzaak – gevolg relatie.

Op de site [Spurious Correlations](http://www.tylervigen.com/) van Tyler Vigen, staan allerlei fantastische voorbeelden. Mijn favoriet (en ik ben daarin niet alleen) is de correlatie tussen de kaas-consumptie per hoofd van de bevolking en het aantal mensen dat sterft doordat ze verstrikt raken in de bedlakens.

Een ander toppertje van Tyler Vigen is de overtuigende correlatie tussen het aantal mensen dat verdrinkt in een zwembad en het aantal films waarin Nicholas Cage speelt. Kortom, als Nicholas Cage een jaartje vrij neemt, hoef je niet meer op je kinderen bij het zwembad te letten. Of toch wel? Zeker wel, want correlatie zegt dus niets over causaliteit. Om oorzaak en gevolg te kunnen bepalen heb je meer nodig dan (big) data. Je hebt ook gezond verstand nodig en verstand van het fenomeen waar je naar kijkt. Hoe meer data er is, hoe belangrijker dat wordt.

Je zou dus ook kunnen zeggen dat het werken met deze data, misschien onbedoeld, armoede criminaliseert.

Probleem 4. Snap je de correlatie wel?Een correlatie is natuurlijk ook heel vaak geen toeval. Vaak is er wel degelijk iets aan de hand. Maar wat precies? Je kunt best wel een eind komen om dat te onderzoeken vanuit de data. Hieronder zie je een voorbeeld, dat vooral supergaaf is als je houdt van statistieken, katten en lange mensen.

Maar meestal is het vooral handig als je domeinkennis hebt. Stel dat je bijvoorbeeld in de data vindt dat mensen statistisch gezien vaak overlijden in de ambulance op weg naar het ziekenhuis. Dan zou je daaruit – zonder domeinkennis – kunnen concluderen dat het beter is – na een ongeluk – om niet in te stappen in de ambulance. Het helpt dan om te weten dat ambulances vaak mensen meenemen die al overleden zijn, maar dat mag eigenlijk niet. Daarom is de verklaring op papier dat die mensen overleden zijn in de ambulance. Ook helpt het om te weten dat een veel grotere groep juist overleeft omdat ze meegaan in de ambulance.

Probleem 5. Oorzaak en gevolgOorzaak en gevolg zijn gemakkelijk door elkaar te halen. Want was is nu de oorzaak en wat is het gevolg. In het boek [Appen is het Nieuwe Roken](http://technofilosofie.com/appen-is-het-nieuwe-roken/) wordt onderstaand voorbeeld gebruikt: Mensen die bij de hoofdingang van het ziekenhuis staan te roken, doen dat waarschijnlijk omdat ze door het roken in het ziekenhuis zijn beland. Of weten ze misschien dat roken slecht voor je is, hebben ze snel een pyama aangetrokken en gaan ze voor de zekerheid dus alvast bij de ingang van het ziekenhuis staan?

Succesvolle bedrijven hebben tevreden medewerkers. Of leiden tevreden medewerkers tot succes? Succesvolle studenten maken zijn vaak on-campus. Of leidt vaak on-campus zijn tot succes? En daarnaast zijn er natuurlijk vaak heel veel factoren die elkaar beïnvloeden en zijn rechtlijnige oorzaak-gevolg relaties hartstikke zeldzaam. Vaak is er een derde factor. Simpel voorbeeld: studenten die vaak naar de sportschool gaan, doen het beter op school. Maar vaak naar de sportschool gaan, leidt niet tot betere schoolprestaties. De derde factor is namelijk: discipline.

Probleem 6. Algoritmes zijn vaak niet transparant of eerlijkWe hebben hierboven een aantal problemen met data opgesomd. Toch gebruiken we data, met al deze tekortkomingen, vaak om beslissingen te automatiseren of ondersteunen. Dat gaat niet altijd goed.

In haar boek [Weapons of Math Destruction](https://www.bol.com/nl/p/weapons-of-math-destruction/9200000074706126/) (ja,ja, woordspeling!) laat Cathy O’Neil stap voor stap zien hoe algoritmes volgebakken zitten met vooroordelen en op die manier bijdragen aan ongelijkheid en discriminatie in plaats van het tegenovergestelde. Je kunt bijvoorbeeld met data nagaan of een verdachte kans maakt op recidive. Dit doe je door je bijvoorbeeld af te vragen of hij/zij al eerder in aanraking is gekomen met de politie. Máár als je dan ook weet dat in sommige staten in Amerika de zwarte bevolking 5% is, en het aantal random aanhoudingen van de verkeerspolitie van zwarte bestuurders bijna 40%, dan weet je ook dat het antwoord voor een afro-american dan sneller ‘ja’ zal zijn. Zet je zo iemand dan ook langer vast, komt hij/zij langer in aanraking met andere criminelen, wordt de kans op een baan kleiner, en de kans op recidive groter. En dus zeg je, het algoritme werkt. Maar dat is dan dus niet zo.

Algoritmes zijn, volgens Cathy, meningen verpakt in code. Aan de andere kant blijkt uit [dit artikel](https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms) op Harvard Business Review, dat algoritmes misschien niet objectief zijn, maar altijd nog objectiever dan mensen.

Het is ook heel moeilijk om de vooroordelen uit de historische data of de algoritmes te houden. Want doet soms geen recht aan de feiten. Alle Duitse mannen hebben een bierbuik, is een vooroordeel, maar Duitse mannen zijn de dikste van Europa is een feit.

Algoritmes die op basis van data allerlei conclusies trekken of nieuwe inzichten geven zijn meestal niet **transparant**. De reden hiervoor is uiteraard dat anders mensen het systeem gaan ‘zitten fucken’. Bijvoorbeeld (shocker) de trending topics in Twitter zijn écht niet de topics die het meest genoemd worden. Maar hoe je dan wel een trending topic wordt, is onduidelijk. Ergens wel terecht, want als je zou weten hoe het werkt, gaan mensen allerlei onderwerpen in hun eigen belang trending maken. Echter, het wordt veel problematischer als het allerlei gevolgen voor je heeft. En je weet het vaak niet eens.

Een ander voorbeeld van een niet transparant algoritme is wat er gebeurde met YouTube. Een tijdje geleden besloot YouTube om het algoritme zo te veranderen dat vloggers met clips die langer duren (15 minuten +) en elke dag verschijnen hoger scoren op de aanbevolen clips. En als je verschijnt in de aanbevolen clips leidt dat tot meer views. De reden hiervoor was uiteraard dat YouTube dan meer kijkers krijgt, die langer kijken (en dus meer data, advertentie-inkomsten). Maar wie kan er elke dag 15 minuten of meer posten, juist mensen die bagger maken. Die meelullen met videospelletjes of zomaar wat filmen. Kortom, het algoritme leidt tot een afname van de kwaliteit. Of, oneerbiedig gezegd, het algoritme voedt onze kinderen bagger.

Overigens als data analyse wel transparant is, ontstaan er ook allerlei problemen. Stel je hebt, net als in Amerika, een ranking van universiteiten. Deze ranking kan niet het belangrijkste meten (wat heb je geleerd), maar wel allerlei andere indicatoren, zoals scores, uitval, donaties van alumni, sponsors, kans op een baan, etc… Het gevolg is dan dat de scholen zich juist volledig op deze indicatoren gaan richten, omdat een goede school leidt tot een goede baan en of je een goede school bent, bepaald de ranking. Omdat de score de studiegelden niet meerekent (weet je nog, data is niet neutraal), krijg je ook een oneerlijke vergelijking en een enorme toename van studiekosten. Het ergste is dat dit soort data-modellen, ook als ze transparant zijn, zichzelf versterken. Immers goede scholen krijgen meer aanvragen, kunnen beter selecteren, krijgen betere studenten, trekken betere professoren aan, etc…

Probleem 7. De feedback ontbreekt
Een goed datamodel kent feedback. Cathy O’Neil geeft het volgende voorbeeld. Stel je moet een nieuwe basketballer selecteren uit de collegespelers. De LA Lakers, een profteam, selecteren een pointguard niet, want die heeft niet genoeg assists. De pointguard wordt geselecteerd door de Utah Jazz en daar heeft hij heel veel assists. het gevolg zal zijn dat de LA Lakers opnieuw naar hun modellen gaan kijken, en uitzoeken wat ze nog meer moeten bekijken om beter te kunnen voorspellen. Feedback maakt het model beter.

Hetzelfde geldt voor Jurgen Streppel die als hoofdtrainer van Willem II besloot om afscheid te nemen van Frenkie de Jong. Niet genoeg talent. Foutje natuurlijk, maar dankzij feedback kon Streppel ervan leren. Zo maakt Frenkie niet alleen Barcelona, maar ook Jurgen beter.

Stel nu dat iemand solliciteert op een baan bij de McDonalds. Daarvoor moet hij een aantal persoonlijkheidsvragen invullen en op basis daarvan wordt hij niet geselecteerd. Daarna gaat hij bij de Burger King werken (die stellen andere vragen) en daar wordt hij binnen no-time franchisehouder. Het systeem van de McDonalds zal dat nooit weten, en blijft dus even dom. Hetzelfde geldt voor studenten. Misschien is de snelste manier naar succes wel uitvallen bij Fontys.

De reden is eenvoudig. In het eerste twee voorbeelden gaat het over mensen die miljoenen waard zijn. Het tweede voorbeeld is bedoeld om een scheiding aan te brengen in de kudde. In dat laatste voorbeeld is feedback onbelangrijk en maakt het niet uit of er wat koren bij het kaf zit. Kinderen die met het badwater worden weggegooid zijn een ingecalculeerd risico. Oninteressant. Tenzij je één van die kinderen bent, natuurlijk.

*Probleem 8. Data kijkt naar het verleden…. duh…*De term predictive analysis komt vaak langs in de datawereld. Het idee is simpel. Je kijkt naar de data (die altijd uit het verleden komt, uiteraard) en je gebruikt allerlei slimme software die patronen herkent die de toekomst kan voorspellen, Misschien, maar alleen als de toekomst grotendeels op het verleden lijkt. Als het systeem min of meer stabiel is. Als er juist grote wijzigingen in het systeem nodig zijn (denk aan klimaatproblemen, migratie, mobiliteit, etc….) dan is er iets heel anders nodig dan data, namelijk verbeeldingskracht. Dan moet je out of the box denken en data is per definitie in the box.

Probleem 9. Ik ben veel complexer dan duizend datapunten
Zoals we al eerder zeiden, met (big) data toepassingen kun je veel beter mensen doorgronden. Je snapt ze beter, krijgt indicaties wat ze drijft en je ziet patronen. Op die manier kun je mensen beter bedienen als klant, student, patiënt of burger. Zo lang het bij suggesties tot verbetering blijft is er niets aan de hand. Maar het wordt twijfelachtiger als er conclusies aan verbonden worden voor een individu.

Ik heb al aangegeven dat algoritmes meningen zijn die verpakt zijn in code en dat dataverzamelingen ook bevoordeeld zijn. Dat is een probleem. Maar stel dat je het beste met mensen voor hebt, dan is er nog misschien wel een groter probleem en dat is dat de werkelijkheid  veel complexer is dan de data en dat is dus vervelend als je de voorspelling vanuit al die (big) data weer terug brengt naar het individu. Hieronder leg ik het uit met een voorbeeld van een zeehond en Ben Bernanke.

Je kent een afbeelding van een zeehond, toch? Je hebt dat gededuceerd aan de hand van een aantal kenmerken (kaal, snorharen, droevige blik). En misschien heb je wel gelijk, maar weet je het zeker?

Heel zeker?

Weet je zeker dat het geen afbeelding is van een pelsrob? Of een zeeleeuw? Je weet dat – tenzij je een marinebioloog bent – waarschijnlijk niet zeker. En als je het wel zeker weet, weet je al helemaal niet wat het karakter van deze zeehond is. Is het een pinguïnjagende killer of gezellige loebas die met ballen op zijn neus kan jongleren.

Voor een gedeelte doet (big) data dat ook. Aan de hand van een groot aantal kenmerken kom je tot een conclusie over iets of iemand. Maar hoe geraffineerd en edelmoedig de (big) data techniek ook is, de belangrijke vraag is natuurlijk of die conclusie wel recht doet aan een veel gecompliceerdere werkelijkheid. Getrouwde mannen leven ongetwijfeld gezonder dan alleenstaande mannen. Maar als jij nou net gaat scheiden van die kettingrokende vrouw die je stress bezorgt, en elke dag frikandellen bakt, moet je dan óók meer ziektenkostenpremie gaan betalen?

Een ander voorbeeld, Ben Bernanke, die van de Federale Reserve wilde zijn hypotheek in 2014 opnieuw financieren. Dat moet mogelijk zijn zou je zeggen. Bernanke barst van het geld en hij had op dat moment een enorme boekdeal. Maar, hij had ook net een nieuw baan. [Het gevolg, het algoritme wees zijn aanvraag af.](https://www.nytimes.com/2014/10/03/upshot/why-ben-bernanke-cant-refinance-his-mortgage.html?_r=0) Net een nieuwe baan? Risico! De algoritmes keken naar zijn **kaalheid,** **snorharen en droevige ogen** en besloten dat het een zeehond was. Niet heel ver gezocht, maar wel fout.

Probleem 10. Roofdieren
In bovenstaande paragrafen heb ik een aantal problemen aangestipt met betrekking tot (big) data. Meestal ging dat echter om mensen, organisaties of systemen die niets kwaads in zin hadden. (Big) data en algoritmes leiden onbedoeld tot problematische situaties. Helaas zijn er ook heel veel mensen die minder goedbedoelend zijn, en die krijgen een geweldige nieuwe toolset. Met (big) data kun je bijvoorbeeld persoonlijke advertenties aan mensen laten zien. Dat klinkt sympathiek, dan zie ik die schoenenreclames van Zalandoo niet meer, maar het betekent ook dat je mensen kunt uitzoeken die kwetsbaar en naïef zijn en die heel gericht kan targeten. Hiervan zijn enorm veel voorbeelden en het probleem is dat het allemaal niet of nauwelijks gereguleerd is.

Probleem – extra 1. Wat niet weet, wat niet deert
In zekere zin is het fijn om onwetend te zijn. Immers als je niet weet welke studenten waarschijnlijk gaan uitvallen dan hoef je je er ook geen zorgen om te maken. Weet je het wel, dan krijg je hele nieuwe vragen. Ben je moreel verplicht om in te grijpen? Moet je deze data delen met de student? Of juist niet?  Als je ontdekt dat bepaalde klanten producten kopen met minder winstmarge? Ga je die klanten dan actief ontmoedigen? Of als je ontdekt dat sommige bezoekers van discotheken meer amok maken dan andere. Ga je die dan weigeren? Of als blijkt dat bepaalde mensen meer ongelukken maken. Geef je die geen polis? Mag je discrimineren op basis van een dataprofiel? Is dat goed zaken doen of juist niet? Het zijn die vraagstukken zullen meer en meer een rol gaan spelen in de toekomst.

Lees ook mijn column over de [buienradar](http://technofilosofie.com/wp-content/uploads/2018/01/Columns_Bron_RensvdVorst_2018_29.pdf) (link naar PDF).

Probleem – extra 2. Statistieken en visualisatiesOok op grote lijnen, kun je van alles met data – analyses. De oude uitspraak: je hebt leugens, grote leugens en statistieken, geldt natuurlijk ook voor hoe je data presenteert. Experimenten hebben aangetoond dat je met dezelfde data-visualisatie software elkaar compleet tegensprekende visualisaties kan maken van dezelde dataset (in opdracht van het Rathenau – instituut, paragraaf [5.1](file:///C%3A%5CUsers%5C879817%5CDownloads%5CRapport_Datagedreven_samenleving_Rathenau%20Instituut.pdf)). Kortom, de data is niet neutraal, de analyse ook niet, de algoritmes niet, noch eventuele visualisaties.

Hier nog een mooi artikel op [LinkedIn](https://www.linkedin.com/pulse/van-data-naar-informatie-de-valkuilen-martin-van-bers/?trackingId=3PsB%2BplLTGdWZK05GzgazQ%3D%3D) van Martin van Bers, waarin hij een aantal manipulaties met grafieken heeft verzameld.

**(4) Hoe zit het met privacy?**
Natuurlijk moet je je ook  afvragen welke data je wel kunt verzamelen en welke niet. Er is veel discussie over ethische vraagstukken rondom meten en onderzoek. Waar kun je wel onderzoek naar doen? Wat niet? Als je Facebook – data ontdoet van iemands personalia, kun je nog steeds met 95% zekerheid de [etniciteit](https://bits.blogs.nytimes.com/2013/06/01/why-big-data-is-not-truth/?_r=0) vaststellen en daar advertenties aan koppelen. Mag dat? We komen dan heel dicht bij het onderwerp gegevensbescherming en privacy. Een onderwerp dat we uitgebreid bespreken [hier](http://technofilosofie.com/privacy/).

**(5) Wat is de Quantified Self Movement en wat zijn de voor,- en nadelen?**Lees alles daarover meer in crash course 2

**(6) Wat is de dataficatie van onze maatschappij en spionagekapitalisme?**
Lees alles daarover meer in crashcourse 2

**(7) Tips voor de toekomst**Ondanks alle nadelen is er toch  – de laatste jaren – in hoge mate sprake van zogenaamd  data-ism of [datafundamentalisme](https://hbr.org/2013/04/the-hidden-biases-in-big-data) (artikel van Harvard). Kate Crawford (Microsoft) muntte die term en het staat voor een blind vertrouwen in de kracht van data. Het idee dat correlatie altijd wijst op causaliteit (o ja, kaas en bedlakens!) en dat grote datasets en voorspellingen op basis van data objectief zijn (niet dus!).  En het komt vaak voor, meer en meer bedrijven adverteren met het feit dat ze datadriven zijn. Datadriven marketing, datadriven e-commerce, datadriven decisionmaking en ga zo maar door. Data-driven heeft nu 23,5 miljoen hits op Google en dat groeit snel. Het is blijkbaar iets om trots op te zijn.

Google maar een beetje rond. Overal zijn er bedrijven die zich er op voor staan dat ze data-driven zijn. Er zijn overal [consultancy bureau’s](https://godatadriven.com/) die hun data verkopen meestal met slogans die soepel inspelen op angst. Stelt u zich de zware typische Hollywoodstem voor. In een wereld die snel verandert, waarin uw concurrenten uw inhalen. Is er maar één oplossing. Aanpassen. Data-driven zijn. Overal vind je beleidstukken waarin serieuze organisaties zeggen dat ze data centraal stellen, dat ze data-driven zullen zijn. Datafundamentalisme, dus.

Maar gezien alle bovenstaande opgesomde nadelen lijkt mij dat een organisatie misschien wat minder trots zou moeten zijn op het feit dat ze zich laat sturen door data. Datadriven decisionmaking is – eufemistisch gezegd – een besmette term. Persoonlijk, heb ik liever te maken met een organisatie die zich niet te laten leiden-en-rijden door data, zeker niet gezien alle uitdagingen die hierboven geschetst zijn. Geef mij maar een organisatie die zijn klanten centraal stelt of mensen. Geef mij maar een organisatie die human-driven is en ondersteund wordt door data. Dat klinkt minder sexy wellicht, maar wel beter.

Ik heb uiteraard geen allesomvattende oplossingen, maar ik heb op basis van hetgeen ik zelf geleerd heb, en heb neergeschreven in bovenstaande long read toch een aantal vuistregels proberen te formuleren, die helpen bij om goed om te gaan met alle mogelijkheden van (big) data.

1. Elke dataset ***moet*** sceptisch bekeken worden. Besef dat er fouten in de data zitten. Besef dat de data is verzameld door mensen met een mening. En vraag je vooral af welke data ***NIET*** verzameld is en/of geen onderdeel uitmaakt van de dataset. Wat je ook doet, beschouw data niet als iets neutraals;
2. Data is een **tool**, geen beslisser. Je vraagt je hamer ook niet hoe je je huis moet bouwen. Als je ergens de term data driven decision making hoort, loop dan weg;
3. Een data scientist is alleen van waarde als hij of zij verstand heeft van het vak waar ze data over verzamelen. Vraag je af of je genoeg verstand hebt om de data op waarde te kunnen schatten;
4. Wantrouw zogenaamde proxies. Als je iets niet kunt meten (bijv. wat heb je geleerd op school, ben je een goede werknemer) dan wordt vaak allerlei data eromheen verzameld (proxies). Het gevaar is dat de proxies (uitval, rendement, aantal docenten met een titel, etc…) belangrijker worden dan de vraag-die-je-niet-kon-meten;
5. Wantrouw data die conclusies verbindt aan **WIE** je bent. Het moet in heel gevallen gaan om **WAT** je doet, niet om wie je bent. Je krijgt een gevangenisstraf om wat je gedaan hebt, niet om wie je bent. Je krijgt een lening omdat je een bepaald verleden en inkomen hebt, niet om waar je woont, etc….
6. [Stel de juiste vragen](http://www.addmustard.com/blog/when-big-data-too-much-data/). Kevin Kelly wijdt er in [The Inevitable een heel hoofdstuk aan](https://www.nobb.nl/nobb/nieuws/5350-citaat-385). Goede vragen worden veel belangrijker dan goede antwoorden. Dat geldt zeker in de wereld van de (big) data. Laat data je helpen om vragen te stellen.

Kortom, kies voor organisaties die (big) data vooral als een tool beschouwen waar je verstandig mee om moeten gaan. Misschien dat in de toekomst blijkt dat ik het verkeerd zie, dat machine-leren en data-analyse zo goed wordt, dat het veel beter is dan wat wij mensen kunnen (want ja, we geven het toe, wij zijn ook niet perfect). Als dat zo is, dan kom ik er op terug, maar voor nu kies ik voor organisaties die (big) data met bovenstaande 6 vuistregels in gedachten beschouwen. En bij voorkeur het gewoon met kleine letters schrijven.

Lijkt me wel zo gepast.

**BIJLAGE A. ACHTERGROND EN GENIETEN MAAR**

### BOEKEN



Cathy O’Neil schrijft blogs op [mathbabe.org](http://mathbabe.org/) en is auteur van het fascinerende boek Weapons of Math Destruction. In haar boek geeft ze talloze voorbeelden die laten zien hoe het gebruik van data ongelijkheid en discriminatie in de hand werkt. Het meest ontluisterende van het boek is de meeste voorbeelden van organisaties of mensen zijn die (big) data gebruiken om goed te doen, maar de resultaten zijn vaak het tegenovergestelde. Lees over Predictive Policing of over de ranking van universiteiten en je begrijpt dat een data – driven society het tegenovergestelde is van een utopie.



In zijn boek HOMO DEUS, de opvolger van de wereldwijde bestseller Homo Sapiens, verhaalt Yuval Noah Harari over een toekomst vol met data-ism. Daarin wordt ieder individu ontleed in al zijn losse onderdelen, van koopgedrag tot ziektegeschiedenis, met grote voorspellende kracht. Het systeem is nu al betrekkelijk zichtbaar in advertentiemachines op internet. Zoals God ooit werd ontmaskerd als een product van menselijke verbeelding, zo zal ook de menselijke verbeelding worden ontmaskerd als het product van biochemische processen en berekeningen, voorspelt Harari. Nu al schrijven computers verrassend goede muziek en verslaan ze professionele Go-spelers. Het ‘systeem’ zal ons beter gaan kennen en voorspellen dan wij ooit zelf kunnen. En dus zal het nu nog samenbindende geloof in de uniciteit van het individu ernstig gevaar gaan lopen, denkt Harari. Gezellig boek, dus, maar een absolute aanrader.

En natuurlijk over Privacy: [Je hebt wel iets te verbergen](https://www.bol.com/nl/p/je-hebt-wel-iets-te-verbergen/9200000063121305/?suggestionType=suggestedsearch) van [Maurits Martijn](https://decorrespondent.nl/mauritsmartijn) en [Dimitri Tokmetzis](https://decorrespondent.nl/dimitritokmetzis). In dit boek wordt haarfijn uit de doeken gedaan waarom privacy zo belangrijk is, hoe slecht het daarmee gesteld is Ã©n waarom de uitspraak “privacy interesseert met niet, ik heb niets te verbergen”, écht heel dom is. De vraag is overigens of de mensen die dat soort dingen zeggen, dit soort boeken lezen. Jij moet het in ieder geval wel doen, aanrader!