dialogic
innovatie • interactie

Big Data in onderwijs en wetenschap

Inventarisatie en essays

In opdracht van:
Ministerie van Onderwijs, Cultuur en Wetenschap (OCW)

Publicatienummer:
2014.086-1507

Datum:
Utrecht, 24 februari 2015

Auteurs:
Frank Bongers
Cor-Jan Jager
Robbin te Velde

Managementsamenvatting

Achtergrond inventarisatie van en essays over Big Data in onderwijs en wetenschap

De verwachtingen over Big Data zijn groot. Steeds meer organisaties proberen een antwoord te vinden op de vraag welke betekenis Big Data heeft voor hun producten, processen en diensten. Deze vraag is allerminst ongegrond. De hoeveelheid data die wereldwijd wordt opgeslagen, is de afgelopen tien jaar elk jaar verdubbeld. Niet alleen het volume groeit exponentieel maar ook de omloopsnelheid en de variatie in de data. Men mag verwachten dat deze ontwikkeling nieuwe kansen biedt voor allerlei maatschappelijke en economische sectoren, zo ook het onderwijs en de wetenschap. Tegelijkertijd roepen Big Data ook vragen op. De beschikking over een grote hoeveelheid data en het kunnen koppelen van deze data aan andere bronnen vormen geen garantie voor een verbetering van het onderwijs en de wetenschap.

Big Data is zowel voor het onderwijs als de wetenschap in tweeërlei opzichten relevant: als *methode* en als *onderwerp*. De methode verwijst naar de toepassing van Big Data-technieken in het onderwijs- en wetenschapsdomein zelf. Dat kan zowel in het primaire proces (in steeds meer wetenschapsdisciplines wordt Big Data op grote schaal in het onderzoek toegepast) als in het secundaire proces (sturingsinformatie over leerlingen en [wetenschappelijk] personeel). Het onderwerp verwijst met name naar de bredere ontwikkelingen rond Big Data, zoals de kansen en bedreigen die Big Data biedt voor de samenleving.

Vanuit deze optiek heeft het ministerie van OCW een opdracht gegeven voor een verkenning van het gebruik en de impact van Big Data in het onderwijs en de wetenschap. Deze verkenning bestaat enerzijds uit een inventarisatie en anderzijds uit een essayistische beschouwing in de vorm van een utopie en dystopie. In de inventarisatie ligt de nadruk op de methode (dus op het gebruik van Big Data in het onderwijs en de wetenschap). In de beschouwing valt de nadruk op het onderwerp (dus beschouwing over Big Data).

De verkenning moet het ministerie van OCW helpen beter zicht te krijgen op Big Data in het onderwijs en de wetenschap. Deze verkenning is geleid door de volgende vier onderzoeksvragen:

- In welke mate wordt Big Data in het onderwijs en in de wetenschap nu al gebruikt? Is er sprake van een hype?
- Wat (of wie) zijn de potentieel disruptieve krachten die van buiten middels Big Data de instituties en institutionele verhoudingen in onderwijs en wetenschap kunnen beïnvloeden?
- Wat is de (structurele) impact op onderwijs en wetenschap en is deze impact op lange termijn (2025) positief (utopie) of negatief (dystopie)?
- Welke rol kan/moet het ministerie van OCW op het gebied van Big Data pakken?

Big Data: afbakening & verwachtingen

Big Data refereert – ondanks de naamgeving – niet alleen aan een grote hoeveelheid data. Het gaat ook om de slimme combinatie en toepassing van (on)gestructureerde data die als restproduct van digitale gegevensverzameling ontstaat. Computertechnologie wordt steeds kleiner (miniaturisatie), goedkoper, krachtiger, (draadloos) verbonden en steeds meer geïntegreerd in andere producten en diensten. We dragen steeds meer apparaten bij ons die middels sensoren, antennes, netwerken en applicaties continu gegevens verzamelen, opslaan en transporteren. Het gaat dan al lang niet meer om elektronica, maar ook om

kleding, meubels, gebouwen, auto's en andere producten en diensten die met elektronica worden uitgerust, ondersteund en gekoppeld aan netwerken. De grote hoeveelheden veelal ongestructureerde data die deze producten voortbrengen lijken haast onbegrensde mogelijkheden te bieden om nieuwe kennis te ontwikkelen, patronen te detecteren en voorspellingen te doen. Een meer theoretische definitie van de wetenschappers Mayer-Schönberger en Cukier stelt dat Big Data betrekking heeft op het vermogen van de samenleving om informatie op nieuwe manieren in te zetten voor het verkrijgen van nuttige inzichten of waardevolle goederen en diensten. Het is in onze optiek zinloos om Big Data te verbinden aan een aantal bytes. Het gaat immers om een dynamische ontwikkeling waarin steeds meer data verzameld en bewerkt wordt door een groeiende hoeveelheid onderling verbonden apparaten die te pas en te onpas data opslaan en transporteren. Wat vooral nieuw is aan Big Data in onze tijd zijn de nieuwe mogelijkheden die deze ontwikkeling biedt, ook voor onderwijs en wetenschap.

Uit de literatuur en de gesprekken die wij in het kader van deze opdracht voerden in het onderwijs en de wetenschap leiden wij vooralsnog een niet-uitputtende lijst *hoge verwachtingen* af over Big Data, bijvoorbeeld:

- Betere voorspellingen over natuurlijke en sociale fenomenen.
- Meer en betere innovaties.
- Hogere arbeidsproductiviteit.
- Beter onderbouwd beleid.
- Aanvullende methoden en bronnen voor het verrichten van onderzoek.

Tegelijkertijd worden verwachtingen over de impact van Big Data ook getemperd:

- Impact van Big Data op de bescherming van de privacy.
- Gebrek aan generieke en specifieke competenties om met Big Data om te kunnen gaan.
- Big Data betekent zoveel informatie dat 'door de bomen het bos niet meer wordt gezien'.
- Te groot vertrouwen in de mogelijkheden van Big Data waarbij inhoudelijke duiding het aflegt tegen 'contextarme' kwantificering.
- Risico van 'bad science', wetenschappelijk onderzoek dat onjuist of oneigenlijk gebruik van gegevens maakt om bepaalde causale verbanden aan te tonen.

Essays: dystopie en utopie voor onderwijs en wetenschap

In het vervolg van dit rapport presenteren wij vier *essays*, namelijk een *dystopie* en *utopie* voor het *onderwijs* en een *dystopie* en *utopie* voor de *wetenschap*. In deze essays kijken wij terug vanuit het jaar 2025.

Als *hoofddimensie* voor het opstellen van de essays gebruiken we *de mens is leidend* (utopie) versus *de technologie is leidend* (dystopie). Deze dimensie maakt dus een onderscheid tussen een situatie waarin het publiek 'in control' is hoe Big Data worden ingezet. Aan de andere kant staat een meer technologiegedreven toepassing van Big Data. Wij zijn niet 'in control' met betrekking tot gebruik en toepassing van Big Data. Als afgeleide dimensie nemen wij het onderscheid tussen:

- Utopie: Publieke toegang tot Big Data gekoppeld aan democratische controle en brede verspreiding van Big Data-vaardigheden (‘publiek is de baas’), versus
- Dystopie: Private toegang tot Big Data gekoppeld aan beperkte controle en smalle verspreiding van Big Data-vaardigheden (‘elite is de baas’).

Bij het schrijven van de utopieën en dystopieën is de uitdaging aangegaan om deze dimensies te verwerken in essays die meer zijn dan louter een spiegelbeeld van elkaar. Het reflectieve en essayistische karakter zit in het feit dat we in de beschrijving van elke dystopie (in 2025) ergens een ‘verkeerde’ beslissing hebben genomen of dat we een negatieve ontwikkeling omtrent Big Data niet hebben kunnen ombuigen naar een positieve ontwikkeling. De negatieve krachten overheersen de positieve krachten op weg naar 2025. In de utopie beschrijven we hoe we deze negatieve invloed ten positieve hebben kunnen keren. Deze insteek impliceert dat de beschrijving van de utopie uitgebreider zal zijn dan de beschrijving van de dystopie. De dystopieën en utopieën zijn het resultaat van een creatief denkproces en pretenderen niet een waarheid over de toekomst te presenteren. Het gaat eerst en vooral om het stimuleren van de gedachtenvorming en aan te geven op welke terreinen zich ontwikkelingen voordoen waar het ministerie op zou kunnen reageren.

Keuzemomenten

Op basis van deze essays hebben we een aantal keuzemomenten geformuleerd. Dat zijn keuzes die zich de komende jaren *kunnen* aandienen. Aan deze keuzes kan het ministerie van OCW (met partners) aandacht schenken bij vervolgstappen in de domeinen van Big Data in het onderwijs en in de wetenschap. Deze keuzemomenten (op hoofdlijnen) zijn:

1. Identificeren van rol(len) die het ministerie van OCW moet spelen in ontwikkelingen rondom Big Data in het onderwijs en in de wetenschap.

2. Stimuleren van het gebruik van Big Data binnen de kaders van de bescherming van privacy enerzijds en de stimulering van het benutten van kansen anderzijds.

3. In welke mate moeten het ministerie en het onderwijsveld investeren in generieke en specifieke competentie-ontwikkeling om met Big Data om te kunnen gaan.

4. Stimuleren dat partijen in het onderwijs en in de wetenschap actief data gaan delen.

5. Het onderwijs en de wetenschap actief stimuleren om aansluiting te zoeken bij internationale ontwikkelingen omtrent Big Data.

Inhoudsopgave

1 Inleiding

1.1 Achtergrond

De verwachtingen over Big Data zijn groot. Steeds meer organisaties proberen een antwoord te vinden op de vraag welke betekenis Big Data heeft voor hun producten, processen en diensten. Deze vraag is allerminst ongegrond. De hoeveelheid data die wereldwijd wordt opgeslagen, is de afgelopen tien jaar elk jaar verdubbeld. Niet alleen het volume groeit exponentieel, maar ook de omloopsnelheid en de variatie in de data (ook aspecten van Big Data).

Men mag verwachten dat deze ontwikkeling nieuwe kansen biedt voor allerlei maatschappelijke en economische sectoren, zo ook het onderwijs en de wetenschap. Hoewel er veel optimistische verhalen de ronde doen over welke meerwaarde Big Data kan bieden, is het de vraag of deze ook verzilverd gaan worden.[1] De beschikking over een grote hoeveelheid data en het kunnen koppelen van deze data aan andere bronnen vormen geen garantie voor een verbetering van het onderwijs en de wetenschap. Bovendien zijn er waarborgen nodig om Big Data niet te laten vervallen in 'Big Brother' of een "heerschappij van de Datameesters".[2] De mogelijkheden en impact van Big Data zijn dus nog niet uitgekristalliseerd. Hebben we met een hype te maken of is er werkelijk sprake van een ontwikkeling die uiteindelijk een grote impact zal hebben op onderwijs en wetenschap? Of andersom: welke bijdrage kunnen onderwijs en wetenschap leveren aan een effectief gebruik van Big Data?[3] Aan welke randvoorwaarden moet worden voldaan om de vruchten van Big Data te plukken en de risico's van Big Data te minimaliseren? Wat kan de overheid doen met Big Data? Etc.

De trends in het gebruik van Big Data in het onderwijs en in de wetenschap kunnen niet los worden gezien van de ontwikkelingen van Big Data in het algemeen. Big Data is zowel voor het onderwijs als de wetenschap in tweeërlei opzichten relevant: als *methode* en als *onderwerp*. De methode verwijst naar de toepassing van Big Data-technieken in het onderwijs- en wetenschapsdomein zelf. Dat kan zowel in het primaire proces (in steeds meer wetenschapsdisciplines wordt Big Data op grote schaal in het onderzoek toegepast) als in het secundaire proces (sturingsinformatie over leerlingen en [wetenschappelijk] personeel). Het onderwerp verwijst met name naar de bredere ontwikkelingen rond Big Data, zoals de kansen en bedreigen die Big Data biedt voor de samenleving.

Onderwerp en methode kunnen niet los van elkaar worden gezien. Een – verlichte – toepassing van Big Data in het onderwijs- of wetenschapsdomein wordt uiteraard gevoed door een grondige reflectie op de mogelijke voor- en nadelen van Big Data in het algemeen. Andersom heeft een reflectie op Big Data in het onderwijs en de wetenschap weinig zin

[1] De Nationale Denktank komt in een recente studie uit op een potentiële toegevoegde waarde van Big Data voor de Nederlandse economie van 45 miljard euro. Dit is een nogal rooskleurige schatting die waarschijnlijk geen rekening houdt met verdringingseffecten, maar geeft desalniettemin een indruk van de grote potentie. Nationale Denktank (2014). *Big data. Samenvatting analysefase*.

[2] Een situatie waarin wiskundigen en computerwetenschappers in een positie verkeren waarin ze kunnen heersen over onze levensinformatie (Baker, S. (2012), De Datameesters. Hoe onze gegevens in ons voor- en nadeel worden gebruikt, Maven Publishing, Amsterdam, p. 20.).

[3] NWO heeft in samenspraak met honderd wetenschappers Big Data geïdentificeerd als een van de zes uitdagingen waarlangs de Nederlandse wetenschap vanuit haar sterktes kan bijdragen aan het tegemoet treden van maatschappelijke vraagstukken en Nederlandse topsectoren (Ministerie van OCW (2014), *Wetenschapsvisie 2025, Keuzes voor de Toekomst*, Den Haag).

zonder dat het subject in kwestie (de student, de wetenschapper, de docent) in de praktijk zelf met Big Data heeft gewerkt.

Vanuit deze optiek heeft het ministerie van OCW ons gevraagd een verkenning van het gebruik en de impact van Big Data in het onderwijs en de wetenschap uit te voeren. Deze verkenning bestaat uit een inventarisatie en een essayistische beschouwing in de vorm van een utopie en dystopie. In de inventarisatie ligt de nadruk op de methode (dus op het gebruik van Big Data in het onderwijs en de wetenschap). In de beschouwing valt de nadruk op het onderwerp (dus beschouwing over Big Data). De opbrengst van de inventarisatie is een overzicht van de huidige stand van zaken met betrekking tot het gebruik van (en kennis over) Big Data in het onderwijs- en wetenschapsdomein. De essays over onderwijs en wetenschap geven een objectieve en onderbouwde duiding aan de verschillende richtingen waarnaar de huidige situatie zich kan ontwikkelen.

1.2 Vraagstelling en aanpak

De verkenning moet het ministerie van OCW helpen beter zicht te krijgen op Big Data in het onderwijs en de wetenschap. Om tegemoet te komen aan de informatiebehoefte zoals hierboven aangegeven, hebben wij ons laten leiden door de volgende vier onderzoeksvragen:

- In welke mate wordt Big Data in het onderwijs en in de wetenschap nu al gebruikt? Is er sprake van een hype?
- Wat (of wie) zijn de potentieel disruptieve krachten die van buiten middels Big Data de instituties en institutionele verhoudingen in onderwijs en wetenschap kunnen beïnvloeden?
- Wat is de (structurele) impact op onderwijs en wetenschap en is deze impact op lange termijn (2025) positief (utopie) of negatief (dystopie)?
- Welke rol kan/moet het ministerie van OCW op het gebied van Big Data pakken?

Aangezien er in eerste instantie behoefte is aan een brede verkenning en niet aan een wetenschappelijk doorwrochte studie, hebben wij een flexibele aanpak gehanteerd die tijdens de looptijd van de verkenning ruimte bood om verschillende inzichten mee te nemen. Ten eerste hebben we continu publicaties verzameld over Big Data (boeken, artikelen, websites). Ten tweede hebben we personen uit het onderwijs en de wetenschap geïnterviewd. De interviews betroffen vooral een reflectie op relevante ontwikkelingen en een projectie daarvan op de toekomst.

1.3 Leeswijzer

In het volgende hoofdstuk schetsen we de algemene ontwikkelingen rondom Big Data (grotendeels los van de domeinen onderwijs en wetenschap). We besteden aandacht aan de positionering en impact van Big Data. Onder meer de vraag of we met een hype te maken hebben wordt beantwoord. We sluiten af met een bespreking van de dimensies die voor de essays worden gehanteerd. De twee opvolgende hoofdstukken inventariseren de stimulerende en remmende krachten voor het gebruik van Big Data in het onderwijs (hoofdstuk 3) en in de wetenschap (hoofdstuk 4). In deze hoofdstukken presenteren we ook een dystopie en utopie voor beide domeinen. Het rapport sluit af met een overzicht van de keuzemomenten die zich aandienen in de verdere ontwikkeling en toepassing van Big Data in het onderwijs en de wetenschap. In de bijlage staat een overzicht van gesprekspartners.

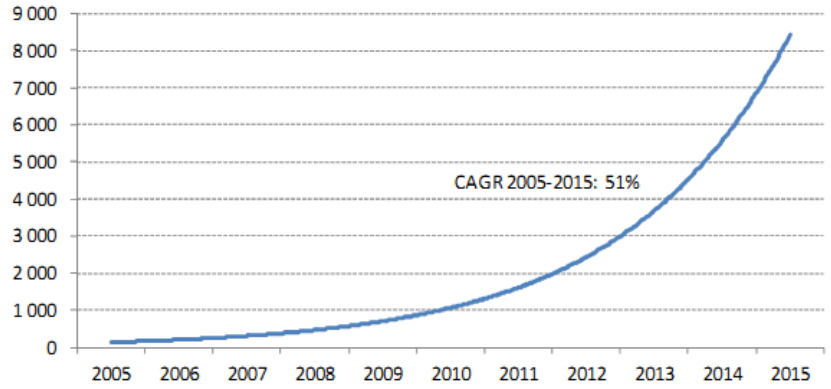
2 Big Data

2.1 Inleiding

Dit hoofdstuk geeft een algemeen overzicht van Big Data (dus niet specifiek gerelateerd aan onderwijs of wetenschap). Onderwerpen die aan de orde komen zijn een positionering en afbakening van Big Data met aandacht voor ontwikkelingen in afgelopen jaren (paragraaf 2.2). Daarna bespreken we de impact van Big Data en de verwachtingen over Big Data met aparte aandacht voor de vraag in hoeverre Big Data een hype of ontwrichtend is (paragraaf 2.3). Tot slot komen de structurerende dimensies aan de orde waarlangs de utopie en dystopie over Big Data in het onderwijs en de wetenschap behandeld gaan worden in de volgende twee hoofdstukken (paragraaf 2.4).

2.2 Positionering van Big Data

We leven in een wereld waarin het volume, de snelheid en de variatie van data die dagelijks verzameld en bewerkt wordt exponentieel groeit. In 2013 was – volgens IBM - de schatting dat 90% van alle data in de wereld in de twee jaren daarvoor was gecreëerd. Dagelijks worden 2,5 miljard gigabytes aan data gecreëerd, genoeg om 27.000 tablets per minuut te vullen.



Figuur 1: Ontwikkeling van totale hoeveelheid data (in exabytes) die wereldwijd is opgeslagen, 2010-2015 (bron: OECD 2013, gebaseerd op IDC Digital Universe research project).

Deze enorme hoeveelheid data wordt continu in onze omgeving gegenereerd, bewerkt en verspreid. Elk digitaal proces en alle sociale media produceren deze data. Systemen, sensoren en mobiele apparaten verspreiden de data. De term Big Data roept bij menigeen associaties op met *heel veel* data. Een praktische grens die daarbij wel eens gehanteerd wordt, is dat de datasets te groot zijn om met reguliere tools te verwerken en te analyseren. Dit leidt nogal eens tot de veronderstelling dat de dataset - bij wijze van spreken – niet meer in een statistisch of spreadsheetprogramma past. Big Data refereert – ondanks de naamgeving – niet alleen aan een grote hoeveelheid data. Het gaat ook om de slimme combinatie en toepassing van (on)gestructureerde data die als restproduct van digitale gegevensverzameling ontstaat.

De nadruk op het aspect volume impliceert dat Big Data voor dit belangrijk onderdeel een relatief begrip is. Gegevens worden al eeuwenlang verzameld, getransporteerd en bewerkt met de hulpmiddelen die op dat moment beschikbaar waren. Napoleon voerde rond 1800 in de Lage Landen de Burgerlijke Stand in hetgeen tot een – voor die tijd – "explosie" van gegevens leidde die gebruikt werden voor doop-, trouw- en overlijdensakten, dienstplicht en

belastingheffing. Artikel 1, Lid 2 van de Amerikaanse grondwet bepaalt sinds 1790 dat er iedere tien jaar een "United States Census" plaatsvindt. Dit is een volkstelling die bepalend is voor de belastingen die een staat moet afdragen aan de federale overheid en hoeveel vertegenwoordigers elke staat mag afvaardigen in het Huis van Afgevaardigden. Dit Grondwettelijk artikel leidde voor die tijd tot Big Data. Zo zijn er ongetwijfeld meer voorbeelden van (publieke) maatregelen die resulteerden in Big Data. Elke tijd heeft zijn eigen Big Data.

De (relatief) grote hoeveelheden data die sinds het begin van de negentiende eeuw over de samenleving worden verzameld, hebben echter weldegelijk tot een radicale omwenteling geleid die op dit moment, na circa twee eeuwen, in de 'Big-Data-hype' lijkt te culmineren. Ironisch genoeg begint de ontwikkeling in de sociale wetenschappen, waar statistici avant la lettre (zoals Quetelet) op basis van sociale statistieken allerlei regelmatigheden in sociale fenomenen (zoals misdaad en zelfmoord) gaan beschrijven.[4] De omwenteling nu is gelegen in het feit dat waarschijnlijkheidsrekening en statistiek bewijzen dat *regelmatigheden voortgebracht kunnen worden door toevalsprocessen*.[5] Eeuwenlang was daarvoor de leidende gedachte geweest dat alle fenomenen (zowel fysieke als sociale) in beginsel zijn terug te voeren op vaste regelmatigheden en oorzakelijkheden. De hele klassieke Griekse filosofie draait in wezen om het identificeren van dit soort 'onderliggende Goddelijke wetmatigheden'. Vanaf Quetelet wordt de argumentatie omgedraaid: orde – het berekenbare – is het resultaat van chaos, in plaats van dat aan chaos een onbekende orde wordt toegekend.[6] Sindsdien is het belang van statistiek en waarschijnlijkheidsrekening sluipenderwijs toegenomen. In vrijwel elk domein worden steeds meer besluiten genomen op basis van probabilistische argumentaties. Met de opkomst van operations research neemt het gebruik van geavanceerde statistische technieken om besluitvorming te ondersteunen in de jaren 1930-1950 al een enorme vlucht, decennia vóór de opkomst van het huidige vertoog rond Big Data. Het huidige vertoog staat dus eerder aan het eind dan aan het begin van de 'probabilistische revolutie'.

Het is echter onjuist en onterecht om Big Data af te doen als oude wijn in nieuwe zakken. Een dergelijke bewering legt onterecht te veel nadruk op het volumeaspect. De huidige ontwikkeling is eerder te duiden als een *tweede* golf. Zoals het werk van Quetelet was gebaseerd op het beschikbaar komen van (voor die tijd) grote hoeveelheden statistische data, is de huidige opkomst van Big Data gebaseerd op een tweede exponentiële toename van de beschikbaarheid van data. Die toename is sterk technologisch gedreven. De digitalisering van de afgelopen decennia heeft een revolutie veroorzaakt in opslag, transport en bewerking van data.

Computertechnologie wordt steeds kleiner (miniaturisatie), goedkoper, krachtiger, (draadloos) verbonden en steeds meer geïntegreerd in andere producten en diensten. We dragen steeds meer apparaten bij ons die middels sensoren, antennes, netwerken en applicaties continu gegevens verzamelen, opslaan en transporteren. Het gaat dan al lang niet meer om elektronica, maar ook om kleding, meubels, gebouwen, auto's en andere producten en diensten die met elektronica worden uitgerust, ondersteund en gekoppeld aan netwerken (zgn. 'smart devices').[7] De grote hoeveelheden veelal ongestructureerde data die

---

[4] Pas later wordt de statistiek binnen de natuurkunde (statistische mechanica) geïntroduceerd door Maxwell, die zich daarbij liet inspireren door het statistische werk van Quetelet.

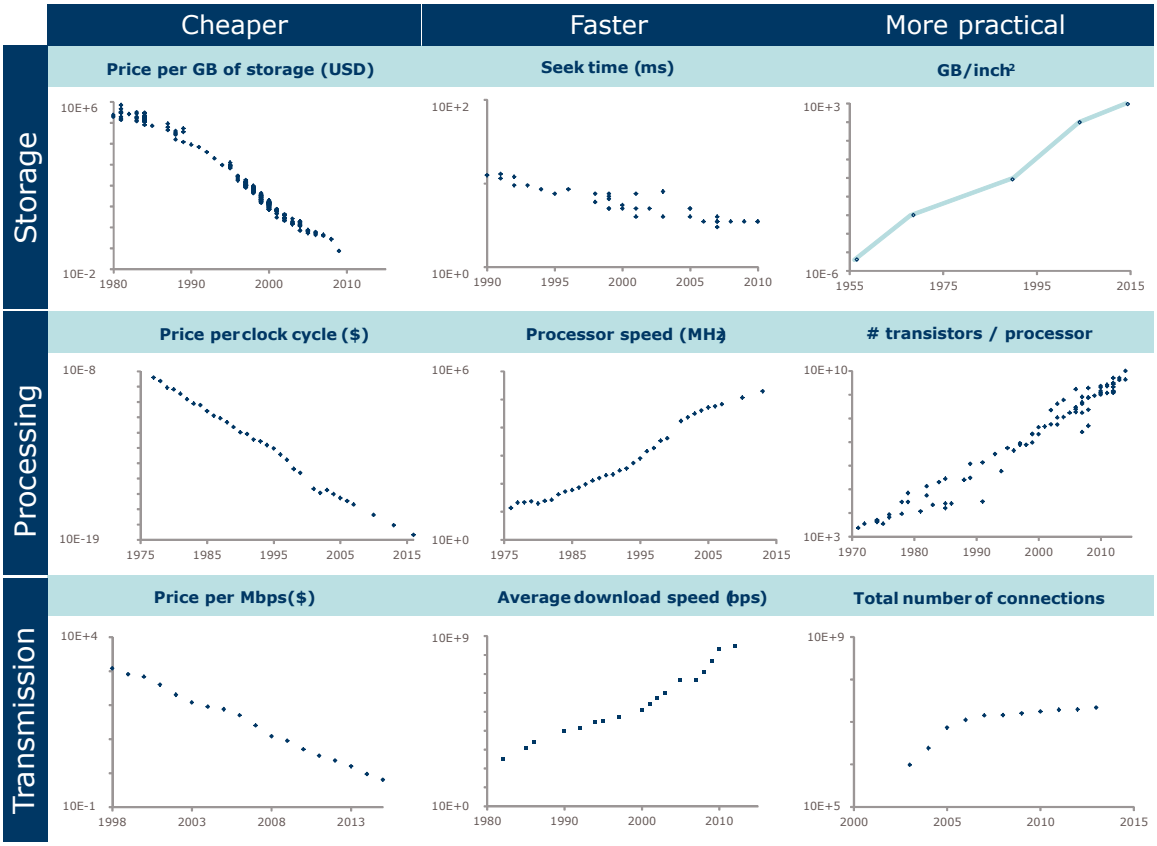
[5] Gerard de Vries (1999). De onbekende revolutie. De Groene Amsterdammer (10 maart).

[6] Ibid.

[7] In dit verband valt de term 'Internet of Things' (IoT). Dit betreft de onderlinge verbondenheid van miljarden uniek identificeerbare digitale apparaten via internet. Deze connectiviteit (ook van 'smart

deze producten voortbrengen lijken haast onbegrensde mogelijkheden te bieden om nieuwe kennis te ontwikkelen, patronen te detecteren en voorspellingen te doen.

De totale hoeveelheid data die de samenleving genereert is, zoals eerder gezegd, exponentieel gegroeid. Tegelijkertijd is het technologische vermogen om data op te slaan, te verwerken en te vervoeren ook exponentieel toegenomen en de relatieve prijs evenzeer afgenomen (zie Figuur 2). Het is nu dus mogelijk om (i) veel meer processen (ii) veel gedetailleerder en (iii) veel vaker (near real-time) door te meten.



Figuur 2: Ontwikkelingen in de computerindustrie in de afgelopen jaren (Dialogic, forthcoming).

Het gaat bij Big Data dus niet alleen om grote hoeveelheden data (een interpretatie die beter past bij de vroegere Big Data). Het gaat nu evenzeer om data die als bijproduct van andere processen van digitale gegevensverzameling optreden, data die ongestructureerd en ongelijksoortig zijn (bijvoorbeeld tekstdata en data uit verschillende systemen) en bovenal creatieve en slimme combinaties van data uit verschillende bronnen die leiden tot nieuwe en onverwachte inzichten. Samenvattend komt het bij Big Data neer op de volgende drie eigenschappen ('de drie V's'):

- Groot volume aan data ('Volume').
- Grote snelheid waarin deze data verzameld en getransporteerd worden ('Velocity').

---

devices') luidt naar verwachting een tijdperk in van ongekende automatisering van allerlei domeinen, inclusief de ontwikkeling van geavanceerde toepassingen als 'smart grid'. O. Monnier (2013), *A smarter grid with the Internet of Things*. Texas Instruments, 2013. Het aantal via verbonden apparaten wordt geschat op enkele tientallen miljarden in 2020 (Gartner, ABI Research).

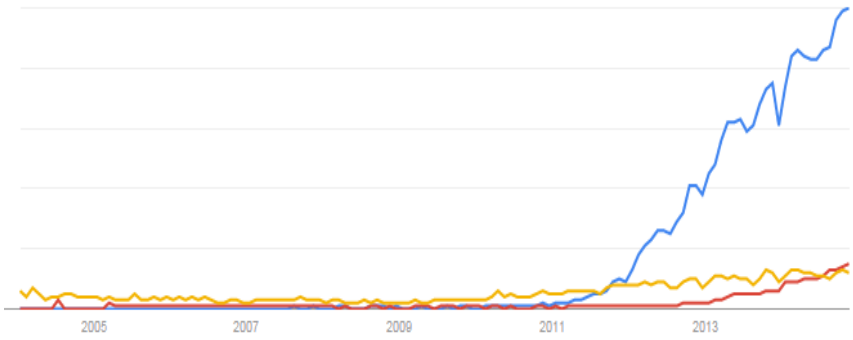
- Grote variatie van de data ('Variety').

Een meer theoretische en minder technische definitie stelt dat Big Data betrekking heeft op het vermogen van de samenleving om informatie op nieuwe manieren in te zetten voor het verkrijgen van nuttige inzichten of waardevolle goederen en diensten.[8] Er is echter geen sprake van een radicale breuk met de 'probabilistische manier van denken', omdat het probabilistisch denken natuurlijk al langer aanwezig is dan Big Data. Wel neemt de snelheid waarmee statistische correlaties worden gevonden meer toe dan de snelheid waarmee wij verklaringen of hypothesen kunnen formuleren. We kunnen nu alleen nog meer 'orde uit chaos' scheppen. Tegelijkertijd betekent dit dat er nog minder ruimte voor toeval en 'noodlot' overblijft.

Het is in onze optiek zinloos om Big Data te verbinden aan een aantal bytes. Het gaat immers om een dynamische ontwikkeling waarin steeds meer data verzameld en bewerkt worden door een groeiende hoeveelheid onderling verbonden apparaten die te pas en te onpas data opslaan en transporteren. Wat vooral nieuw is aan Big Data in onze tijd zijn de nieuwe mogelijkheden die deze ontwikkeling biedt, ook voor onderwijs en wetenschap.

2.3 Impact van Big Data

Vanwege het dynamische karakter van de ontwikkeling en het gebruik van Big Data is het relatief onzeker welke impact hiervan verwacht mag worden. Big Data zou een hype kunnen zijn die na verloop van tijd vanzelf wegebt. In dat geval zou de impact van Big Data beperkt zijn. Gegeven de ontwikkelingen die wij in de vorige paragraaf beschreven verwachten wij echter niet dat Big Data een tijdelijk karakter heeft. Overigens observeren we wel dat de term Big Data zelf in toenemende mate een negatieve connotatie meekrijgt en wellicht minder houdbaar zal blijken dan het fenomeen dat het beschrijft. Er is ongetwijfeld een piek in de aandacht voor Big Data (bijvoorbeeld in zoekmachines) op het internet, maar ook na deze piek zal Big Data vaste grond verwerven.

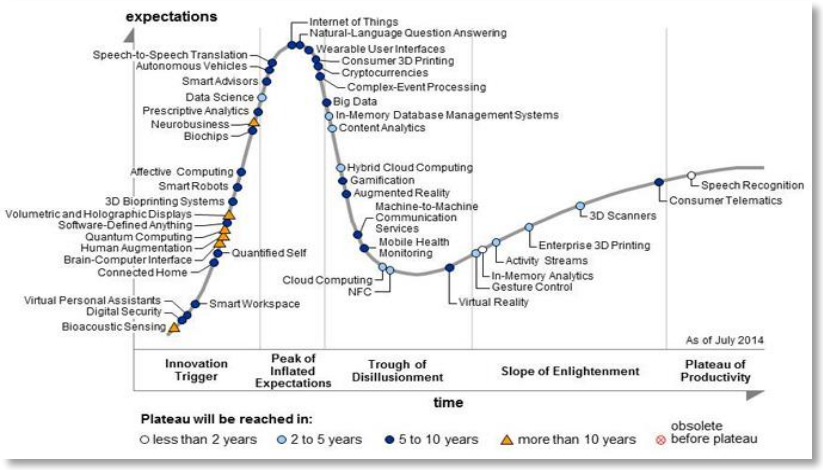


Figuur 3: Populariteit van de termen 'Big Data' (blauwe lijn), 'Open Data' (oranje lijn) en 'Data Science' (rode lijn) op Google, 2005-2013 (bron: Google Trend Charts).

Het is mogelijk dat Big Data na de verhevigde aandacht zich in deze jaren ontwikkelt tot 'gewone' data en verweven raakt met allerlei maatschappelijke en economische processen, zonder dat deze inbreng goed zichtbaar is of dat daar apart aandacht aan wordt besteed. De

[8] Mayer-Schönberger, V. en K. Cukier (2013), *De Big Data Revolutie. Hoe de data-explosie al onze vragen gaat beantwoorden*, Maven Publishing BV, Amsterdam, p. 10. Naast de genoemde drie V's wordt in het kader van Big Data gewezen op twee andere V's, namelijk Visualization en Value.

intensieve aandacht op dit moment gaat meestal gepaard met hoge verwachtingen. De Gartner Group visualiseert deze verwachtingen in een 'hype cycle' (zie de volgende figuur).



Figuur 4: Big data op de Gartner Hype Cycle.

Uit dit figuur blijkt dat de verwachtingen over Big Data de piek inmiddels gepasseerd zijn. De tempering van deze verwachtingen komt meestal voort uit een groeiend besef dat een technologie nog niet volwassen genoeg is voor brede verspreiding, maatschappelijke en economische voorwaarden onvoldoende vervuld zijn om een innovatie goed te 'absorberen' of dat de innovatie gepasseerd wordt door een andere en betere innovatie.

Uit de literatuur en de gesprekken leiden wij vooralsnog *hoge verwachtingen* af over Big Data (we doen hier geen uitspraak of deze verwachtingen ook uit zullen komen), bijvoorbeeld:

- Dankzij Big Data zullen wij *betere voorspellingen* doen over allerlei uiteenlopende fenomenen, of het nu om een natuurverschijnsel als het klimaat gaat of om sociale verschijnselen als menselijk gedrag.[9]
- *Innovaties* zullen dankzij Big Data gerealiseerd worden, omdat de slimme en creatieve uitwisseling, koppeling en bewerking van grote gegevensbestanden de opmaat is naar nieuwe producten, diensten en processen.
- Big Data zal bijdragen aan een *hogere productiviteit*, net zoals andere digitale toepassingen – bijvoorbeeld kantoorautomatisering – dat in het verleden ook hebben bewerkstelligd.
- Big Data zal bijdragen aan meer *'evidence-based' beleid*, omdat Big Data beleidsprocessen voedt met meer geobjectiveerde en 'real-time' informatie en

[9] De OESO stelt dat waardecreërende mechanismen van data-analyse zich bevinden op het vlak van verbeterde inzichten en kenniscreatie en op het vlak van geautomatiseerde besluitvorming. Data-analyse helpt om inzichten af te leiden en betere instrumenten te ontwikkelen om dataobjecten (natuurverschijnselen, sociale systemen, individuen) beter te begrijpen, te beïnvloeden en te controleren. Daarnaast helpt data-analyse om machines en systemen uit te rusten om van data te leren en deze apparaten autonome besluiten te laten nemen gebaseerd op data-analyse. OESO (2014), *Data-driven Innovation for growth and Well-being*, Interim synthesis report, Parijs.

nieuwe toepassingen het mogelijk maken om deze informatie direct te ontsluiten.[10] Recentelijk is de Europese Commissie nog een onderzoek gestart naar de impact van Big Data op *'evidence-informed beleid'*.[11]

- Big Data geeft *aanvullende methoden en bronnen* voor het verrichten van onderzoek. Zo kan menselijk gedrag soms beter gemeten worden aan de hand van (veel) gegevens uit informatiesystemen dan op basis van laboratoriumexperimenten en enquêtes met de risico's van zelfselectie en sociaalwenselijke antwoorden. Met Big Data krijgen wij de kans om de samenleving in al haar complexiteit te zien, via miljoenen netwerken van interpersoonlijke relaties.[12]

De betekenis van deze enorme hoeveelheid data voor samenleving, economie, innovatie en onderzoek wordt dus hoog ingeschat, ook door beleidsmakers en onderzoekers.[13] Zo voorziet de Europese Commissie een grote impact van een *data driven economy*,[14] maar tegelijkertijd worden verwachtingen over de impact van Big Data ook getemperd:

- Er bestaan vooral grote zorgen over de impact van Big Data op de bescherming van de privacy van degenen waarover data worden verzameld, gekoppeld en geanalyseerd. Zeker wanneer derden deze gegevens over velen opslaan, onderling delen en gebruiken om ons leven 'aangenamer' te maken of voorspellingen te doen over ons gedrag. In veel gevallen kan dit leiden tot betere en op maat gemaakte producten en diensten, maar de keerzijde is dat er ook persoonlijke informatie vrijgegeven en gebruikt wordt die men liever persoonlijk houdt.
- De samenleving moet de competenties ontwikkelen om met Big Data om te kunnen gaan. Het gaat zowel om meer generieke vaardigheden als om specialistische kennis, bijvoorbeeld om hoogopgeleide 'data scientists'. Er dreigt nu bijvoorbeeld een tekort aan 'data scientists'.[15] Dit zijn experts in het koppelen, bewerken en analyseren van big datasets. Het goed omgaan met Big Data vraagt om competenties als het stellen van de goede vragen, het identificeren van de juiste data en de kunde om data te koppelen, te analyseren en te duiden.[16] Eveneens is er behoefte aan Big-Data-

---

[10] Sommige adviseurs spreken daarom al van het "einde van de beleidsambtenaar" omdat Big Data een nog meer datagedreven beleid mogelijk maakt die geen boodschap heeft aan trage beleidscyclussen. De aanwezigheid en snelheid van data maakt directe terugkoppeling mogelijk waardoor de overheid sneller kan bijsturen. Peter Joosten (2014), Big data bij de overheid: einde van de beleidsambtenaar?, Frankwatching.com (2 juli 2014).

[11] Europese Commissie (2014), Data Technologies for evidence-informed policy-making (including Big Data) – Smart 2014/0004. Tender announcement.

[12] Pentland, A. (2014), *Sociale Big Data. Opkomst van de datagedreven samenleving*, Maven Publishing BV, Amsterdam.

[13] European Commission (2014), *Towards a thriving data-driven economy*, Brussel. Australian Government (2013), *The Australian Public Service Big Data Strategy Improved understanding through enhanced data-analytics capability*, Canberra. HM Government (2013), *Seizing the data opportunity. A strategy for UK data capability*, Londen.

[14] Europese Commissie (2014), *Towards a thriving data-driven economy*, Brussel.

[15] In het VK zal het aantal Big Data specialisten dat werkzaam is bij grote bedrijven tussen 2012 en 2017 groeien met 240%. Big Data Analytics (2012), *An assessment of demand for labour and skills 2012 – 2017*. E-skills UK report on behalf of SAS UK. De data scientist is eerder een onderzoeker of adviseur met ICT-competenties dan andersom.

[16] Verschillende universiteiten en hogescholen starten Data-Science-opleidingen en afstudeerrichtingen (TUD, TU/e, UU, Fontys). Deze opleidingen moeten tegemoetkomen aan de dreigende tekorten op de

gebruikers, ofwel allerlei professionals die in hun dagelijks werk Big Data gebruiken en daarvoor meer generieke datavaardigheden benutten.

- Big Data draagt bij aan de 'informatie overload'. Verschillende informatiesystemen verzamelen zodanig veel informatie dat 'door de bomen het bos niet meer wordt gezien'. Overigens staat hier tegenover dat Big Data door slimme toepassing en presentatie 'informatie overload' juist zou kunnen reduceren, doordat het in staat stelt om 'orde te scheppen' in de chaos van deze grote hoeveelheden data, op manieren die met traditionele methodes niet mogelijk zijn.[17]

- Big Data alleen vertelt het verhaal onvoldoende ("Data does not speak for itself"). Een te groot vertrouwen in Big Data gaat voorbij aan de behoefte aan inhoudelijke duiding. De *betekenis* van data hangt af van de vraagstelling en het kader waarin de data functioneert. [18] Big-Data-analyse vraagt om inhoudelijke domeinkennis, bijvoorbeeld over onderwijsprocessen.

- Big Data verhoogt het risico van 'bad science'. Dat is wetenschappelijk onderzoek dat onjuist of oneigenlijk gebruik van gegevens maakt om bepaalde causale verbanden aan te tonen. Meer specifiek kan Big Data resulteren in meer respect voor correlaties in plaats van het voorzetten van een voortgaande zoektocht naar ongrijpbare causaliteit. Daarbij vergroot Big Data de kans op een waslijst aan andere statistische problemen, door de enorme datasets en de vaak sterkere aanwezigheid van ruis. Zolang veel wetenschappers niet geschoold zijn in de vraagstukken die Big Data met zich meebrengt voor onderzoek, is de kans groot dat ook dit tot onjuiste analyses en conclusies leidt.[19]

Impact van Big Data op (wetenschappelijke) methodologie[20]

- De mogelijkheid om enorme hoeveelheden gegevens over een onderwerp te analyseren in plaats van noodgedwongen te werken met kleinere verzamelingen.

- De bereidheid de rommeligheid (lees: variëteit en probabilistische karakter) van gegevens uit de werkelijkheid te accepteren in plaats van de voorkeur te geven aan exactheid (lees: een veronderstelde diepgaande orde die niet bestaat; of althans nooit gekend kan worden).

- Toenemend respect voor correlaties in plaats van een voortgaande zoektocht naar een ongrijpbare causaliteit. Dit is echter een heikel punt in de wetenschapsfilosofie. Onderzoek doen zonder enige ex ante theoretische veronderstellingen over de relatie tussen variabelen kan al snel tot allerhande onzinnige uitkomsten leiden die zijn gebaseerd op schijnverbanden.

- Nieuw aan Big Data is dat de statistische analyses niet langer op een (representatief) deel van de dataset worden uitgevoerd maar op de gehele

arbeidsmarkt. Ook de Nationale Denktank pleit dit jaar voor een zogenaamde Big Data Academy. Gartner schat dat in 2014 slechts 1/3 van de 4,4 miljoen vacatures wereldwijd vervuld kan worden.

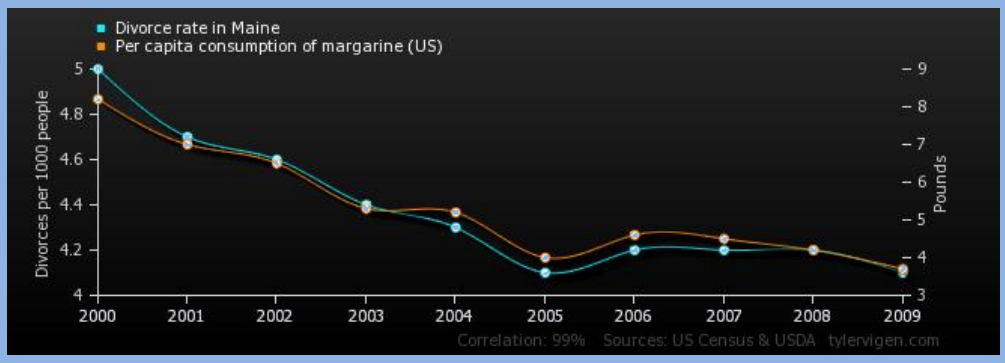
[17] IMERGE Consulting (2014). Big Data and The Crucial Need for Information Governance.

[18] Gerard de Vries (1999). De onbekende revolutie. De Groene Amsterdammer (10 maart).

[19] Fan, J., Han, F. & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review,* 1, 293-314.

[20] Zie voetnoot 8.

dataset. Met andere woorden, er hoeft niet meer met steekproeven te worden gewerkt – met alle problemen van dien.



De figuur toont een bijna perfecte correlatie tussen de verkoop van margarine en het aantal scheidingen in de Amerikaanse staat Maine. Hoewel de onderwerpen vrijwel niks met elkaar te maken hebben, lopen we met het 'eindeloos' koppelen van talloze gegevens (de crux van Big Data) het risico correlaties (en zelfs causale verbanden) te detecteren die er niet zijn.

De impact van Big Data kan zo groot zijn dat gesproken kan worden van maatschappelijke en economische ontwrichting. De vraag is echter of deze ontwrichting positief of negatief is en in welke mate deze ontwrichting zich zal voordoen. Bij ontwrichting moeten we denken aan grote verschuivingen op de arbeidsmarkt (verdwijnen van werk, opkomst van nieuwe beroepsgroepen), het verdwijnen van maatschappelijke verbanden en structuren (implosie van het verenigingsleven, opkomst van nieuwe sociale netwerken) en institutionele veranderingen (herinrichting van het onderwijs).

Sommigen spreken in het verband van Big Data over:

- De **Vierde Revolutie:** (na landbouw, industrie, ICT) nu naar intelligente machines en geavanceerde analysetechnieken om productieprocessen nog beter te maken. De Amerikanen noemen het "the industrial internet". De Duitsers spreken van "Industrie 4.0" en "cyber physical systems". De vierde revolutie zou zelfs uit vier revoluties tegelijk kunnen bestaan[21]:
    - Automatiseringsrevolutie: Alle onderdelen van een productieproces communiceren met elkaar en zorgen dat deze processen efficiënt en effectief worden ingericht. Het werk wordt slim en flexibel verdeeld tussen de apparaten en de enkele productiemedewerkers die nog nodig zijn. Ook buiten de fabriek dringt automatisering in alle haarvaten van onze samenleving door.
    - Makersrevolutie: Consumenten kunnen zichzelf dankzij kleinschalige en goedkope gereedschappen (bijvoorbeeld 3D-printers) thuis of in de lokale gemeenschap ontwikkelen tot ontwerpers, producenten en aanbieders van hun eigen producten ('prosumeren'). Internet en sociale media helpen hun producten en diensten wereldwijd aan te bieden.

[21] Zie de blog van de futuroloog Marcel Kreijveld: Leve de vierde Industriële Revolutie!, (http://wisdomofthecrowd.nl/2014/03/15/industriele-revolutie-4-0/).

- Duurzaamheidsrevolutie: Dankzij 'smart cities' en 'energy grids' worden, in combinatie met hernieuwbare energiebronnen, steden energiezuiniger gemaakt en treden consumenten steeds vaker als producent van energie op (zonnepanelen, etc.). Data-analyse draagt bij aan inzicht en optimalisatie van de productie en het gebruik van (groene) energie.
- Circulaire Revolutie: Deze revolutie gaat verspilling van (consumptie)goederen tegen door meer nadruk op hergebruik. Dit wordt gestimuleerd door de opkomst van de deeleconomie, marktplaatsen en door bij het ontwerp van producten al rekening te houden met hergebruik.

- De automatiseringsgolven sinds de jaren tachtig van de vorige eeuw hebben eerst vooral laagopgeleide beroepen geautomatiseerd in de maakindustrie ('lopendebandwerk'). Vervolgens is veel werk in de (standaard) dienstverlening geautomatiseerd, vooral dankzij de opkomst van internet (bank- en verzekeringswezen, detailhandel, …). Big Data zal het beroep van **hoger opgeleide professionals** beïnvloeden, omdat bijvoorbeeld grote gegevensstromen gedetailleerd 'real time' inzage geven in menselijk gedrag, klimaat, etc.. Hoewel deze data om inhoudelijke interpretatie vragen, zal het leiden tot een andere positie van kenniswerkers. Ook leidt dit tot een vraag naar een andere soort kennis en vaardigheden.[22]
- Big Data zal leiden tot **institutionele vernieuwing**. Simpel gesteld: Big Data impliceert dat bedrijven en instellingen datasets voor het optimaliseren van hun bestaande processen gaan combineren die op het eerste gezicht weinig met elkaar te maken hebben. De volgende stap is dan een volledige omkering: bedrijven nemen data als startpunt en ontwikkelen op basis van die data geheel nieuwe producten en diensten. Zo kan het gebeuren dat 'branchevreemde' organisaties op basis van de data waarover zij beschikken producten en diensten gaan ontwikkelen die zij als buitenstaander aanbieden in voor hun 'vreemde' sectoren. Dit biedt mogelijkheden voor institutionele vernieuwing (KPN in het onderwijs, etc.). Parallel daaraan kan de opkomst van 3D-printen in combinatie met Big Data een deel van de maakindustrie terughalen naar individuele huishoudens en lokale gemeenschappen, wat ook gezien kan worden als een voorbeeld van institutionele vernieuwing.[23] Tevens zijn deze ontwikkelingen niet alleen relevant voor de private sector: zoals in eerdere paragrafen uitgelegd is ook voor overheden het potentieel van Big Data dusdanig groot dat dit kan leiden tot een omkeer in hoe er over datagebruik nagedacht wordt.[24]
- Big Data stelt grote vragen over **privacy**. Enerzijds is het nodig bescherming te bieden en te voorkomen dat - bij wijze van spreken - alle (persoonlijke) gegevens ongecontroleerd in handen komen van een geprivilegieerde groep van enkele bedrijven en personen (dataspecialisten) die alles over ons weten (en daar misbruik van kunnen maken). Anderzijds moeten we onderkennen dat het gebruik van persoonsgegevens, bijvoorbeeld in medisch onderzoek, betere kansen biedt om meer effectieve diagnoses en therapieën te ontwikkelen. De mate waarin Big Data ontwrichtend werkt op bescherming van de persoonlijke levenssfeer zal voor iedereen verschillend zijn, omdat niet iedereen dezelfde normen ten aanzien van privacybescherming hanteert. Bovendien hangen het correct gebruik van data en het

[22] http://www.policy-network.net/pno_detail.aspx?ID=4642&title=The-societal-impact-of-technology

[23] Anderson, C. (2013), *Makers. De Nieuwe Industriële Revolutie*, Nieuw Amsterdam Uitgevers, Amsterdam.

[24] McKinsey & Company (2011). Big data: the next frontier for innovation, competition, and productivity.

tegengaan van misbruik van data samen met datacompetenties van betrokkenen, zowel specialisten als leken (over wie data wordt verzameld). Hiermee samenhangend is een cultuur van digitaal risicomanagement noodzakelijk in het data-ecosysteem.[25]

2.4 Dimensies voor de essays onderwijs en wetenschap

In deze paragraaf presenteren wij de dimensies die de basis vormen voor de essays over onderwijs en wetenschap waarin zowel een utopie als een dystopie wordt opgenomen. De dimensies zijn een structurerend kader voor de essays over onderwijs en wetenschap in de volgende twee hoofdstukken. Deze dimensies slaan op de positie van Big Data in de samenleving. Dit is dus het reflectieve vertoog over Big Data.

Wij gebruiken als hoofddimensie *de mens is leidend* (utopie) versus *de technologie is leidend* (dystopie). Deze dimensie maakt dus een onderscheid tussen een situatie waarin het publiek 'in control' is hoe Big Data worden ingezet. Aan de andere kant staat een meer technologiegedreven toepassing van Big Data. Wij zijn niet 'in control' met betrekking tot het gebruik en de toepassing van Big Data.[26] Als afgeleide dimensie nemen wij het onderscheid tussen:

- Utopie: Publieke toegang tot Big Data gekoppeld aan democratische controle en brede verspreiding van Big Data-vaardigheden ('publiek is de baas'), versus
- Dystopie: Private toegang tot Big Data gekoppeld aan beperkte controle en smalle verspreiding van Big Data-vaardigheden ('elite is de baas').

Bij het schrijven van de utopieën en dystopieën is de uitdaging aangegaan om deze dimensies te verwerken in essays die meer zijn dan louter een spiegelbeeld van elkaar. Het reflectieve en essayistische karakter zit in het feit dat we in de beschrijving van elke dystopie (in 2025) ergens een 'verkeerde' beslissing hebben genomen of dat we een negatieve ontwikkeling omtrent Big Data niet hebben kunnen ombuigen naar een positieve ontwikkeling. De negatieve krachten overheersen de positieve krachten op weg naar 2025. In de utopie beschrijven we hoe we deze negatieve invloed ten positieve hebben kunnen keren. Deze insteek impliceert dat de beschrijving van de utopie uitgebreider zal zijn de beschrijving van de dystopie. De dystopieën en utopieën zijn het resultaat van een creatief denkproces en pretenderen niet een waarheid over de toekomst te presenteren. Het gaat eerst en vooral om het stimuleren van de gedachtenvorming en aan te geven op welke terreinen zich ontwikkelingen voordoen waar het ministerie op zou kunnen reageren.

In de volgende twee hoofdstukken worden deze dimensies toegepast op het onderwijs en op de wetenschap.

[25] OECD (2014), *Data-driven Innovation for Growth and Well-being*. Interim Synthesis Report, Parijs.

[26] In de dimensie 'de technologie is leidend' is de mens feitelijk nog steeds leidend, maar in dit scenario wordt de technologie op onjuiste manier ingezet (manipulatief, ongecontroleerd, ondemocratisch, …).

3 Inventarisatie en essay onderwijs

3.1 Inleiding

Dit hoofdstuk is een beschouwing op het gebruik en de impact van Big Data op het onderwijs. Het hoofdstuk begint met een kwalitatieve beschrijving van respectievelijk stimulerende dan wel afremmende krachten die van invloed zijn op het huidige gebruik van (Big) Data in het onderwijs. Daarna maken we (los van deze krachten) een inventarisatie van huidige ontwikkelingen op het gebied van Big Data die we hebben kunnen optekenen vanuit gesprekken met partijen in en rondom het onderwijs.

Vervolgens presenteren we de dystopie van Big Data in het onderwijs in 2025. Dit is een korte scenarioschets waaruit blijkt dat Big Data niet heeft gebracht wat we ervan verwachtten. Ergens tussen 2015 en 2025 is een 'verkeerde' beslissing genomen of hebben negatieve krachten de positieve krachten van Big Data gedomineerd. Deze dystopie wordt gevolgd met een kwalitatieve beschrijving van deze positieve en negatieve krachten en hoe deze op elkaar inwerken.

In het laatste deel van dit hoofdstuk presenteren we de utopie van Big Data in het onderwijs in 2025. Dit is een korte scenarioschets waaruit blijkt dat Big Data wel heeft gebracht wat we ervan verwachtten. Ergens tussen 2015 en 2025 is een 'goede' beslissing genomen of hebben positieve krachten de negatieve krachten van Big Data gedomineerd. Deze utopie wordt gevolgd door een kwalitatieve beschrijving van deze positieve en negatieve krachten en hoe deze op elkaar inwerken. Centraal staat: wat hebben we gedaan om niet in de dystopie te vervallen.

3.2 Inventarisatie van datagebruik in onderwijs

Alvorens nader in te gaan op de inventarisatie van het huidige en toekomstige gebruik van (Big) Data in het onderwijs, is het zaak om de context van (toenemend) datagebruik te begrijpen. De ontwikkelingen in het onderwijs kunnen immers niet los worden gezien van technologische en maatschappelijke ontwikkelingen waar het onderwijs mee te maken heeft. Om die reden beschrijven we in deze paragraaf eerst een aantal stimulerende en afremmende krachten op datagebruik in het onderwijs aan de hand van actueel beleid en technologische en maatschappelijke ontwikkelingen.

3.2.1 Stimulerende krachten voor gebruik (Big) Data in onderwijs

Hieronder volgt een overzicht van enkele stimulerende krachten voor een toenemend en/of intensiever gebruik van data in de afgelopen jaren.

Opbrengstgericht werken (OGW)

Een van de belangrijkste drivers voor intensivering van het gebruik van data in het onderwijs kan gevonden worden in de ambitie van de overheid tot meer opbrengstgericht werken en professionalisering. De aanzet daartoe kwam reeds tot uiting in de *Kwaliteitsagenda PO*[27], waarin werd gepleit voor een betere benutting van aanwezige informatie over leerprestaties binnen en buiten de school. In zekere zin is in deze agenda de term opbrengstgericht werken geïntroduceerd die in het vigerend beleid als belangrijk uitgangspunt is doorontwikkeld.

[27] www.rijksoverheid.nl/documenten-en-publicaties/notas/2007/11/28/kwaliteitsagenda-po.html.

In het po is hier middels het *Actieplan Basis voor Presteren*[28] vorm aan gegeven. In dit plan worden ambities besproken die in de kern neerkomen op het beter benutten van bestaande kennis. Dit moet in het DNA van de moderne schoolorganisatie zitten. In de actielijnen komt dit bijvoorbeeld terug in de ambities om (1) toegang te krijgen tot prestaties van de school (2) een verplichte centrale eindtoets in te voeren (3) een verplichting tot een leerling- en onderwijsvolgsysteem en (4) een meetinstrument voor de toegevoegde waarde van een school te ontwikkelen.

Het vo kent een soortgelijk Actieplan, genaamd: *Beter Presteren: opbrengstgericht en ambitieus*[29]*.* De doelen van de overheid met dit plan lopen uiteen, maar opbrengstgericht werken (met een ambitie naar 90% van de scholen in 2019) neemt een prominente plek in. Hierin wordt onder andere als actie genoemd een leerlingvolgsysteem en landelijke diagnostische toets verplicht te stellen. Bovenop dit actieplan is er nog *Leraar2020: Een krachtig beroep*[30]*,* waarin professionalisering van de docent centraal staat en onder andere werd aangekondigd dat toezicht op het leraarschap onderdeel gaat uitmaken van de onderwijsinspectie. In het Hoger Onderwijs zijn diverse initiatieven opgezet voor het inzetten van data ten behoeve van het verkrijgen van inzichten die de onderwijskwaliteit ten goede komen. De termen *Institutional Research* en *Learning analytics* zijn daarbij dominant. SURF heeft een experiment opgezet in 2012 [31] , dat gevolg heeft gekregen middels het innovatieprogramma Learning analytics 2013-2014.

Samengevat is het evident dat opbrengstgericht werken een essentieel ankerpunt van het huidige beleid van het ministerie is en dat dit zich manifesteert in een toenemende monitoringsbehoefte voor zowel scholen zelf als overheid. Beschikbare data speelt daar ontegenzeggelijk een grote rol in.

Onderwijs op maat

Een tweede kracht die inspeelt op de databehoefte in het onderwijs heeft betrekking op de toegenomen personalisering in het onderwijs. Ook in voornoemde actieplannen komt dit al tot uiting, zoals in het Actieplan PO*: "Scholen moeten in het onderwijs nog beter aansluiten bij verschillen in capaciteiten tussen leerlingen*". Vervolgens wordt een serie pilots aangekondigd waarin met behulp van onderwijsvolgsystemen, centrale eindtoetsen en achtergrondkenmerken de ontwikkeling van het leerproces inzichtelijk wordt gemaakt.

Ook in het onlangs afgesloten *Sectorakkoord Klaar voor de toekomst*![32] van de VO-Raad wordt deze trend voor passend onderwijs (op maat) gesignaleerd. "*De krimp die zich momenteel voltrekt plaatst scholen voor grote uitdagingen. Scholen worden geconfronteerd met dalende leerlingaantallen. Daarnaast worden achtergronden van leerlingen diverser. Deze trend zal in de toekomst– mede door de invoering van passend onderwijs – aan belang winnen. Ontwikkelingen in de buitenwereld komen met de leerlingen de school binnen. Dit vraagt om steeds meer ondersteuning van de leerling op maat."*

---

[28] www.rijksoverheid.nl/documenten-en-publicaties/kamerstukken/2011/05/23/actieplan-po-basis-voor-presteren.html.

[29] www.rijksoverheid.nl/documenten-en-publicaties/kamerstukken/2011/05/23/actieplan-vo-beter-presteren.html.

[30] www.rijksoverheid.nl/documenten-en-publicaties/kamerstukken/2011/05/23/actieplan-leraar-2020.html.

[31] Innovatieregeling Learning analytics 2012

[32] www.vo-raad.nl/userfiles/bestanden/Sectorakkoord/Sectorakkoord-VO-OCW.pdf.

Parallel hieraan ontstaat er (met name in het po) een markt voor adaptieve leersystemen die per definitie individuele feedback leveren. Hier wordt in de volgende paragraaf uitvoeriger bij stilgestaan. Duidelijk is in elk geval dat persoonlijke leercycli, terugkoppeling op maat en de behoefte tot meer inzicht in individuele leerprocessen aan de hand van analyse van leeropbrengsten met inachtneming van achtergrondkenmerken, ook een stimulerend effect hebben op datagebruik in het onderwijs.

Transparantie

Een derde kracht heeft te maken met een maatschappelijke ontwikkeling, waarin meer transparantie wordt verlangd van de overheid. Het meest in het oog springende middel hiervoor is de Wet openbaarheid van bestuur (Wob), waarin het recht op informatie van de overheid is geregeld. Nu zal een gemiddelde ouder/verzorger van een leerling of student in het onderwijs geen Wob-verzoek bij de overheid neerleggen, maar bestaat evengoed de wens om zich bijvoorbeeld te informeren ten aanzien van schoolkeuze.

Gelet op de (media)-aandacht voor de jaarlijkse 'Elsevier-ranking' van beste scholen en het investeren door de overheid in 'Vensters PO' en 'Vensters VO' lijken dergelijke informatieproducten belangrijker te worden. Schoolinfo, de stichting achter Vensters PO en VO, geeft op haar website aan een product te willen leveren dat onbevooroordeeld, feitelijk en genuanceerd is. De afnemers van deze producten zijn zowel scholen (voor opbrengstgericht werken), ouders (voor transparantie) en andere belanghebbenden zoals onderzoekers en het ministerie van OCW.

In het eerder gememoreerde Sectorakkoord VO wordt ook veelvuldig gesproken over transparantie, zoals bijvoorbeeld: *"Het groeiende belang dat wordt gehecht aan transparantie en doelmatigheid leidt er in het voortgezet onderwijs onder meer toe dat de eisen over de verantwoording van de resultaten en de besteding van middelen verder zullen toenemen"* en *"Het vertrouwen in de sector is niet vanzelfsprekend aanwezig, maar moet continu verdiend worden".*

De gewenste transparantie van onderwijsopbrengsten richting school, ouders en overheid leidt tot een intensivering van het belang dat aan onderwijsdata wordt gehecht.

Verantwoording

Een vierde stimulerende factor kan feitelijk worden gezien als een combinatie van Transparantie en Opbrengstgericht werken. Het ministerie van OCW heeft vanaf 2014 de studiekeuzecheck verplicht gesteld aan HO-instellingen. Nieuwe studenten in het Hoger Onderwijs dienen zich vanaf 2014 vóór 1 mei aan te melden en hebben dan recht op deze check. Elke hogeschool en universiteit kan deze check anders invullen, maar het instrument dient ertoe om studenten een kritische reflectie mee te geven of de studie waarvoor zij zich willen inschrijven wel de juiste is. Ook is een studiebijsluiter verplicht gesteld aan hogescholen en universiteiten. In deze studiebijsluiter moet onder andere worden weergegeven hoe tevreden studenten zijn over deze opleiding en wat het arbeidsmarktperspectief is.

Beide maatregelen moeten bijdragen aan het terugdringen van studie-uitval en studieswitch in het Hoger Onderwijs, hetgeen aan belang heeft gewonnen na het afschaffen van de basisbeurs. Immers: een verkeerde studiekeuze heeft grotere persoonlijke (financiële) consequenties.

Recentelijk (9 december 2014) heeft de Tweede Kamer ingestemd met het wetsvoorstel *'Wet educatie en beroepsonderwijs',* waarin een publieke taak is doorgeschoven naar de Stichting Samenwerking Beroepsonderwijs en Bedrijfsleven (SBB) voor het verknopen van vraag en

aanbod naar Mbo-gediplomeerden. Actuele data met betrekking tot vraag en aanbod speelt daarbij een grote rol. Universiteiten en hogescholen grijpen de studiekeuzecheck en studiebijsluiter ook aan om hun (nieuwe) studenten een 'data-based' profiel van de opleiding aan te bieden. Ook hier is dus sprake van een stimulerend effect.

Analysepotentieel

De laatste factor (analysepotentieel) is meer een 'enabler', dan een afzonderlijke kracht in het spectrum, maar omdat het een factor betreft die zonder meer een stimulerend effect heeft op voornoemde krachten, noemen we hem toch op deze plek.

De totale hoeveelheid data die de samenleving genereert is exponentieel gegroeid. Tegelijkertijd is het technologische vermogen om data op te slaan, te verwerken en te vervoeren ook exponentieel toegenomen en de relatieve prijs evenzeer afgenomen (zie Figuur 2).

Alle ingrediënten voor meer en intensievere ontginning van data zijn dus voorhanden. Het wordt voor overheden, maar ook bedrijven relatief goedkoop om grote servers in te richten en data gestructureerd op te slaan en te ontsluiten naar informatieproducten, hetgeen ook gretig plaatsvindt in het onderwijs (bijvoorbeeld via elo's, aanbieders van adaptieve leersystemen, onderzoeksbureaus, etc.).

Noodzakelijk is wel dat de afnemers van dit analysepotentieel hier ook de toegevoegde waarde van inzien. Het Sectorakkoord VO laat hier geen misverstand over bestaan: "*De sector gaat de mogelijkheden die ICT biedt beter benutten en werken aan het eigentijdser maken van de voorzieningen in het onderwijs. ICT is daarbij geen doel, maar een middel om gepersonaliseerd leren te ondersteunen en meer maatwerk in het voortgezet onderwijs te realiseren".* Deze ambitie vraagt ook om meer draagvlak en expertise onder docenten (Sectorakkoord VO, 2014). Een gedeeltelijke invulling hiervan is gevonden in het *Doorbraakproject Onderwijs en ICT*[33]*,* waarin scholen ICT kunnen inzetten voor meer gepersonaliseerd leren. Dit project is een gezamenlijk initiatief van de PO-Raad, VO-Raad en Ministeries van OCW en EZ.

Met andere woorden: het analysepotentieel is qua techniek, omvang, kwaliteit, economisch perspectief en draagvlak groter geworden. Tegen de achtergrond van de overige vier hierboven beschreven ontwikkelingen, zijn alle ingrediënten aanwezig om (Big) Data intensiever te integreren in het primaire en secundaire onderwijsproces.

3.2.2 Afremmende krachten voor gebruik (Big) Data in onderwijs

Desalniettemin staan niet alle deuren wijd open voor ongelimiteerde inzet van (Big) Data in het onderwijs. We noemen hieronder een aantal afremmende krachten voor het gebruik van data in het onderwijs. We putten in deze beschouwing vooral uit interviews in het veld en deskresearch.

Gestold wantrouwen

Verreweg de meest genoemde remmende kracht heeft te maken met gebrek aan (maatschappelijk) vertrouwen in een data-intensieve maatschappij. We kunnen dit niet zozeer als oorzaak dan wel afgeleide kracht beschouwen van alle hieronder genoemde factoren. Dit gebrek aan vertrouwen kan zijn gevoed vanuit de angst dat persoonsgegevens op straat belanden of worden doorverkocht, de angst in het onderwijs te verworden tot een

[33] www.doorbraakonderwijsenict.nl/.

'getal' zonder eigen identiteit of het gevaar dat we als maatschappij toegroeien naar aan afrekencultuur.

Veelal werd verwezen naar de plannen van de ING[34] als toonbeeld van de invloed van deze kracht op het maatschappelijke debat rondom Big Data. De publieke moraal is uiterst gevoelig voor dergelijke berichtgeving en opgebouwd vertrouwen kan als sneeuw voor de zon verdwijnen, ook als dit in andere domeinen plaatsvindt.

Angst voor privacy- en informatiebeveiligingsissues

Dat vertrouwen kan worden geschaad door het onrechtmatig verzamelen dan wel distribueren van persoonsgegevens. Recentelijk werd nog een onderzoek van PricewaterhouseCoopers in opdracht van OCW[35] in verschillende media breed uitgemeten.

Uit dit onderzoek kwam onder andere naar voren dat de in dit onderzoek onderzochte scholen beperkt op de hoogte zijn van de Wet bescherming persoonsgegevens (Wbp) en deze wet op basis van 'gezond verstand' toepassen. In een aantal gevallen verzamelden scholen gegevens die niet strikt noodzakelijk waren en mogelijk zelfs in strijd met de Wbp – overigens zonder kwade intenties. In het rapport wordt de oorzaak hiervoor treffend geformuleerd: "*Het ontbreekt bij hen (red. scholen) aan tijd en capaciteit om zich in de materie te verdiepen en additionele capaciteiten zoals regievoering op leveranciers in te richten*." Ook informatiebeveiligingsmaatregelen door scholen worden toegepast op basis van 'gezond verstand'. De maatregelen worden ex ante getroffen op basis van ervaringen en incidenten. Bij hogescholen en universiteiten speelt dit punt minder, omdat de meeste van hen (ook vanwege de schaal) een interne functionaris gegevensbescherming hebben die als vooruitgeschoven post van het College Bescherming Persoonsgegevens opereert.

Uit ditzelfde onderzoek werden gerede twijfels geventileerd ten aanzien van adequate naleving van de Wbp van leveranciers van elektronische leeromgevingen en digitale leermiddelen. Het is niet duidelijk hoe gegevens van scholen verwerkt worden, in hoeverre de gegevens worden gedeeld en welke maatregelen voor informatiebeveiliging zijn getroffen. Het College Bescherming Persoonsgegevens heeft onlangs nog (september 2014) 'Snappet Tabletonderwijs' op de vingers getikt[36]. Snappet is aanbieder van tabletonderwijs en kwalificeert en vergelijkt leerprestaties van kinderen van andere scholen, zonder dat de scholen hier expliciete toestemming voor hebben gegeven. Ook liet de informatiebeveiliging te wensen over.

Dergelijke berichten schaden het vertrouwen in eventuele onderwijsverbeteringen met behulp van data. Vanuit interviews hebben we hier gemengde signalen over ontvangen. Enerzijds pleit men voor een duidelijk wettelijk kader en deelt men deze zorg, anderzijds stipt men hier en daar ook het 'verlammende' karakter van te rigoureuze en stringente wet- en regelgeving aan, zonder weging van wetenschappelijke en/of maatschappelijke belangen. Er lijkt vooralsnog geen middenweg te bestaan.

---

[34] ING lanceerde in maart 2014 het idee om een proef te beginnen met het delen van informatie over betalingsgedrag van klanten met bedrijven ten behoeve van gepersonaliseerde advertenties. Deze proef zou overigens plaatsvinden met vrijwillige proefpersonen die hier expliciet hun toestemming voor hadden verleend.

[35] Nulmeting Privacy en informatiebeveiliging in het Primair en Voortgezet Onderwijs (2014) in opdracht van Ministerie OCW.

[36] www.cbpweb.nl/downloads_rapporten/rap_2013_snappet.pdf.

Angst voor afrekencultuur en reductionisme

Een andere kracht die iets verder afstaat van vertrouwen heeft betrekking op een reductionistische kijk op het onderwijs, waarin alleen ruimte is voor alles dat meetbaar is en een daaruit volgende afrekencultuur.

De angst bestaat dat de te beperkte focus op onderwijsrendement en toetsen op harde criteria verder wordt doorgevoerd en er zich als afgeleide een rigide kijk op het onderwijs meester maakt van de samenleving. Conformeren aan de norm wordt de norm. Meermaals is de vraag gesteld in hoeverre we competenties dekkend kunnen vangen in bestaande (toets)instrumenten en – belangrijker in hoeverre de traditionele instrumenten nog geschikt zijn voor een verschuiving naar andersoortige competenties, door sommigen aangeduid als 21st century skills.

In het verlengde daarvan is een jaarlijks terugkerend fenomeen bijvoorbeeld de kritiek van ouders, maar ook scholen in het po op het belang dat aan de Cito eindtoets voor het basisonderwijs wordt gehecht en de stress die deze test meeneemt voor leerlingen.[37] Sommige scholen in het vo stelden harde eisen voor toelating tot een havo of vwo op basis van deze cito-scores en dat draagt bij, volgens een deel van de ouders en scholen in het po, aan een afrekencultuur.

Of deze kritiek nu terecht is of niet, het zijn sentimenten die blijkbaar bestaan in de samenleving en over het algemeen een remmende invloed hebben op het draagvlak voor gebruik van (Big) Data in het onderwijs.

Gebrek aan vaardigheden

Tot slot noemen we hier een kracht die we, schijnbaar tegenstrijdig, ook hebben genoemd bij 'analysepotentieel' in de vorige paragraaf als stimulerende kracht. Dit betreft de vaardigheden van docenten in omgang met data, of breder gezien ICT als geheel.

Er is nog veel verbeterpotentieel in ICT-vaardigheden van docenten. Stichting Kennisnet onderzoekt al enige jaren via de '*Vier in Balans monitor*' de stand van zaken van ICT in het onderwijs. Uit de laatste monitor[38] blijkt dat vooral onderwijsmanagers van mening zijn dat de *didactische* ICT-vaardigheden van docenten onvoldoende zijn. Docenten zelf zijn aanmerkelijk positiever over hun eigen vaardigheden, maar dit lijkt zich met name te manifesteren ten aanzien van 'basisvaardigheden' en niet met betrekking tot de inzet van ICT in het primaire onderwijsproces.

Eerder zagen we al in het onderzoek van PricewaterhouseCoopers[39], dat er ook rondom het gebruik van data nog de nodige verbeterpunten zijn voor scholen. Het is overigens wel de vraag of dit tot de kerncompetenties van het onderwijs moet toebehoren.

---

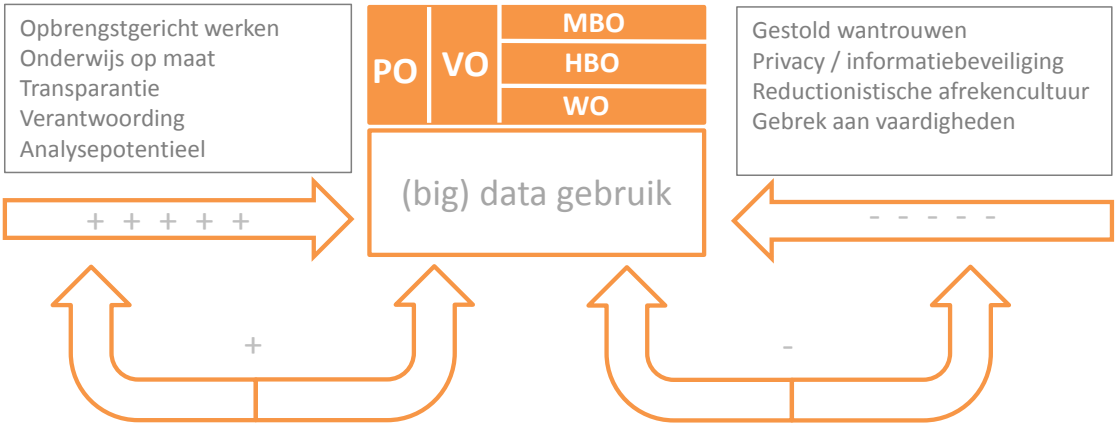
[37] Staatssecretaris Dekker kon zich overigens niet vinden in deze kritiek en pareerde dat er in het onderwijs een angst voor openheid zou heersen. De ambitie is juist om transparant te maken waar scholen zich kunnen verbeteren.

[38] Vier in balans (2013) – de laatste stand van zaken van ict en onderwijs.

[39] Nulmeting Privacy en informatiebeveiliging in het Primair en Voortgezet Onderwijs (2014) in opdracht van Ministerie OCW.

3.2.3 Interactie tussen stimulerende en afremmende krachten datagebruik

Binnen dit krachtenveld opereren vele partijen in het onderwijs. Uit de 'ronde langs de velden' blijkt dat deze krachten door toename van het gebruik van (Big) Data ten behoeve van het onderwijs zichzelf lijken te versterken. Opbrengstgericht werken lokt nog meer opbrengstgericht werken uit. Meer transparantie leidt tot een nog grotere informatiebehoefte, het toegenomen analysepotentieel leidt ertoe dat er een sterke prikkel is om dit potentieel uiteindelijk te benutten.



Figuur 5: Interactie tussen krachtenvelden ten aanzien van gebruik (big) data in het onderwijs.

Aan de andere kant van het spectrum worden de afremmende krachten juist aangewakkerd door de toename van het gebruik van (Big) Data. De angst voor een afrekencultuur wordt bijvoorbeeld gevoed door de verantwoordingsdrang en kwantificering van het onderwijsproces. De angst voor privacy- en informatiebeveiligingsproblematiek wordt groter bij elk 'relletje' dat in de media wordt aangegrepen zoals bijvoorbeeld de nulmeting privacy en informatiebeveiliging van PwC, het onderzoek naar Snappet, maar ook buiten het onderwijsdomein zoals het eerder gememoreerde voorbeeld van ING of het feit dat bijvoorbeeld zeven miljoen Dropbox accounts waren gehackt (oktober 2014). Dit draagt allemaal bij aan een publieke moraal die angstig wordt voor de data-intensieve maatschappij en zich hiervan afkeert. Deze polarisatie van standpunten lijkt zichzelf te versterken en een veel gehoord geluid is dan ook dat de overheid juist op die plek in het krachtenveld (tussen de belangen in) ruimte voor een genuanceerd publiek debat moet bieden.

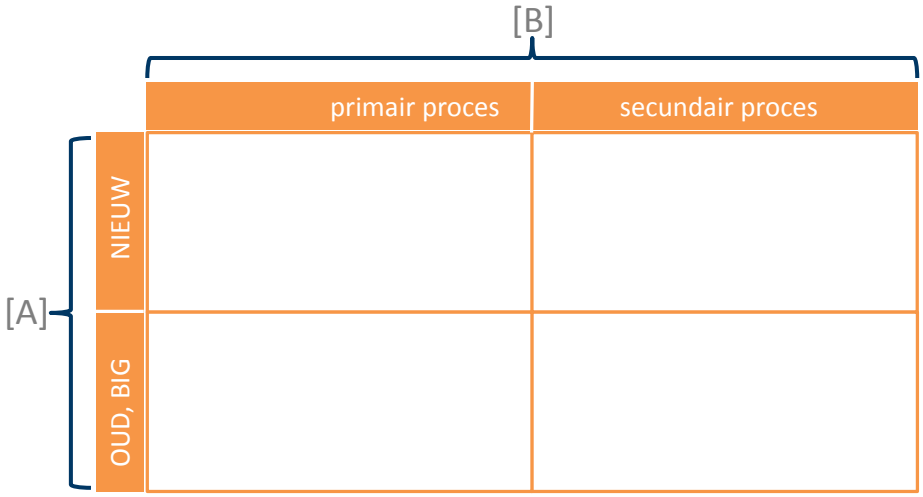
3.2.4 Huidig gebruik oude en nieuwe data in primair en secundair onderwijsproces

Uit inventarisatie van verschillende beleidsstukken vanuit het ministerie van OCW, de nodige onderzoeksrapportages en gevoerde gesprekken in het onderwijsveld zelf, is een beeld ontstaan van het huidige gebruik van (Big) Data in het onderwijs en de projectie daarvan naar de toekomst.

Vanzelfsprekend gaat onderstaande uiteenzetting niet in op afzonderlijke datasets en zal de inventarisatie geen uitputtend karakter hebben. Wel hebben we het vertrouwen middels beschrijving van het krachtenveld in de vorige paragraaf, aangevuld met gesprekken van centrale spelers in het onderwijsdomein, een adequaat beeld op hoofdlijnen kunnen geven van de positie van (Big) Data. We hanteren bij de beschrijving hiervan een inrichtingsraamwerk dat bestaat uit een 2*2 matrix (zie Figuur 6), waarin we onderscheid maken tussen:

> **[A]** De articulatie van verschillende vormen van Big Data in het onderwijsdomein, te weten: '*dezelfde data, maar dan big*' en '*nieuwe Big Data*'

[B] Een onderscheid tussen Big Data in het primaire proces en het secundaire onderwijsproces



Figuur 6: Inrichtingsraamwerk voor het positioneren van partijen op centrale posities.

Dit raamwerk is met name voor het onderwijs een zinvol instrument, omdat het onderwijs voor een groot deel bestaat uit gevestigde partijen (naast de onderwijsinstellingen zelf) die goed te projecteren zijn in dit raamwerk. Ook het onderscheid tussen data *over* het onderwijs (secundair) en data *in* het onderwijs (primair) is nuttig. Het aantal afzonderlijke onderwijsinstellingen is groot en de periode dat een individu acteert binnen dit systeem evenzeer. Er is vanuit een maatschappelijke taak van toezicht, aansturing en kwaliteitsbewaking dus nu eenmaal veel historische data voorhanden (secundair proces). Tegelijkertijd verdwijnt het 'monopolie' van de overheid op onderwijsdata langzaam. Met name door de introductie van 'nieuwe' data die voortkomen uit eigen registraties, loggingsgegevens van digitale lesmethodes en/of externe bronnen. Deze ontwikkeling, die mede wordt gevoed door nieuwe bedrijvigheid binnen het onderwijs (b.v. elektronische leeromgevingen, digitale lesmethodes, adviesdiensten), past goed in de bovenkant van deze matrix.

Dezelfde data, maar big in secundair proces

De meest voor de hand liggende 'cel' om mee te beginnen is die van de historische data met betrekking tot focus op het secundaire proces. Hier vinden we een groot deel van de gevestigde partijen terug, zoals de Dienst Uitvoering Onderwijs, Onderwijsinspectie en het Centraal Bureau voor de Statistiek, met als afgeleide afnemers van deze data het ministerie OCW en de verschillende sectorraden (VSNU, Vereniging Hogescholen, VO-Raad, PO-raad, AOC-raad en MBO-raad) en partijen als Kennisnet en Studiekeuze123.

Niet al deze partijen zijn betrokken bij dit onderzoek, maar over de hele linie zien we hier de krachten van transparantie, verantwoording en opbrengstgericht werken sterk opkomen. DUO houdt op grotere schaal data bij en ziet ook een toename in de vraag naar zijn data (PwC, 2014)[40].

De meest opvallende ontwikkeling die we in deze cel kunnen beschrijven is dan ook niet die van sec een waargenomen toename van data, maar vooral die van een toegenomen *gebruik*

[40] Nulmeting Privacy en informatiebeveiliging in het Primair en Voortgezet Onderwijs (2014) in opdracht van het ministerie van OCW.

ervan, gevoed door de behoefte aan transparantie, verantwoording en opbrengstgericht werken. Dat manifesteert zich vooral door ontwikkeling van verschillende informatieproducten op basis van reeds beschikbare data. Zo ontwikkelt studiekeuze123 diverse producten voor de (aankomende) student, wordt er vanuit SION gewerkt aan de voorwaarden voor eenduidige informatieproducten, zoals die door Schoolinfo (in opdracht van de VO-Raad en PO-Raad) worden uitgerold en wordt steeds meer data uit eigen administraties vastgelegd in dashboards binnen elektronische leeromgevingen.

Mede om de wildgroei van data-aanvragen te voorkomen heeft DUO een portaal ingericht (data.duo.nl) waar op geaggregeerd niveau onderwijsdata van alle sectoren te raadplegen zijn. Ook heeft het ministerie van OCW geïnvesteerd in het opzetten van SION (Samenwerkingsplatform Informatie Onderwijs), waar alle sectorraden samenwerken ten behoeve van stroomlijning van informatievoorziening in de onderwijsketen om eveneens DUO te ontlasten, maar ook om eenduidige definities af te spreken.

Bij de wat grotere uitvoeringsinstellingen (Onderwijsinspectie, CBS, DUO) lijkt voldoende aandacht te bestaan voor informatiebeveiliging en privacy, mede gelet op het feit dat deze organen intern geaudit worden.[41] Zo werkt DUO sinds 2014 met onderwijssectorspecifieke 'omnummers'. Voorheen was er 1 uniek omnummer (ongeacht sector) voor een leerling of student in de administratie. Dat maakte het relatief eenvoudig om bestanden aan elkaar te koppelen, ook voor organisaties die daartoe niet gemachtigd waren. Dat is met deze maatregel (gevoed door een onderzoek van Berenschot) onmogelijk gemaakt. Het CBS heeft middels het Sociaal Statistisch Bestand (SSB), waarin verschillende administraties aan elkaar verknoopt zijn, nadrukkelijk zorg te dragen voor privacyaspecten en werkt dan ook met strikte voorwaarden voor toegang, verwerking en distributie van data uit het SSB.

Echter, zoals beschreven in de vorige paragraaf, bestaat bij andere partijen, zoals leveranciers van elektronische leeromgevingen klaarblijkelijk de nodige onduidelijkheid over toepassing van de Wbp.

Buiten de conventionele onderwijspartijen die hierboven zijn beschreven, zien we bij deze laatste groep, de leveranciers van elektronische leeromgevingen zoals Parnassys, Topicus en Magister, een sterke toename in profilering op type diensten (modules) die onderwijsinstellingen additioneel kunnen afnemen op basis van de opgeslagen onderwijsdata. Het gaan dan om voorzieningen richting ouders, benchmarkmodules en ondersteuning van intern kwaliteitsbeleid. Een markt die sterk in ontwikkeling is rondom deze voorziening is die van Learning Analytics.

Nieuwe Big Data in het secundair proces

Bij het beschrijven van de inventarisatie van nieuwe Big Data *over* het onderwijs, moeten we in beginsel al een fundamenteel onderscheid maken tussen (i) feitelijk nieuwe data en (ii) het koppelen van verschillende databronnen tot nieuwe informatie. In het verlengde van de reeds bestaande onderwijsdata, spelen ook in deze 'cel' de triggers van transparantie, verantwoording en opbrengstgericht werken een grote rol. Meerdere partijen zetten zich in om tot nieuwe informatieproducten te komen, dan wel met behulp van nieuwe data, dan wel door het samenbrengen van data uit andere systemen.

Nieuwe data worden bijvoorbeeld door aanbieders van elektronische leeromgevingen verzameld in samenspraak met scholen ter bevordering van het kwaliteitsbeleid via

[41] Auditrapport 2012, Ministerie van OCW, 2012, Den Haag.

enquêtes, maar ook via verzameling van data uit educatieve apps door marktpartijen of Kennisnet. Er is bijvoorbeeld een enorme hoeveelheid informatie te halen uit de loggingsdata van adaptieve leersystemen. Dit wordt vooralsnog (vanwege privacygevoeligheden) beperkt ingezet voor didactische inzichten, maar die potentie wordt wel als zodanig ervaren in het veld. Onderwijsinstellingen zelf zien zich daarnaast 'genoodzaakt' om steeds meer gegevens bij te houden over leerlingen, ouders en docenten (bijvoorbeeld voor het passend onderwijs of opbrengstgericht werken).

Een grotere groep organisaties spant zich in om vanuit bestaande bronnen tot nieuwe inzichten te komen. Hier zijn legio voorbeelden voorhanden. Een direct gevolg van de studiebijsluiter in het Hoger Onderwijs is de ontsluiting van bestaande data uit verschillende bronnen (Nationale Studenten Enquête, WO-Monitor en DUO) naar een informatieproduct. Meerdere partijen (waaronder SION) verkennen de mogelijkheid om onderwijsdata aan arbeidsmarktontwikkeling te verknopen en ook de Onderwijsinspectie krijgt er een taak bij om toezicht op leraarschap te meten. Binnen Institutional Research afdelingen van Hogescholen en Universiteiten zijn gremia ontwikkeld (bijv. DAIR[42]) om gegevens en inzichten te delen die de betreffende instelling verder kunnen helpen en bijdragen aan het kwaliteitsbeleid. In de omgeving van de uitvoeringsinstellingen, zijn er diverse marktpartijen (onderzoeksbureaus, aanbieders van elektronische leeromgevingen, etc.) die graag bijdragen aan de transparantie en de opbrengstgerichte mindset van het onderwijs door bestaande datasets in te zetten.

Dezelfde data, maar big in primair proces

In het primaire proces zijn vooral de onderwijsinstellingen zelf in beeld en vindt een ontegenzeggelijke opmars plaats van het gebruik van data. Steeds meer gegevens over leerprestaties worden vastgelegd in systemen en partijen en worden met steun van externe partijen ook steeds meer ingezet ten behoeve van het primaire proces.

Sommige onderwijsinstellingen gaan hierin verder dan anderen en geven bijvoorbeeld aan data uit diverse bronnen (bijv. ook 'social media') te willen inzetten in toelatingsgesprekken, studiebegeleiding en bijsturing van de student. De logica daarachter is dat de fysieke wereld zich heeft verplaatst naar de digitale en tegelijkertijd de tijd per student voor begeleiding door een toegenomen werkdruk afneemt. Door het begeleidingsproces met relevante informatie te voeden kan efficiënter worden gehandeld. Wel wordt meermaals benadrukt dat data gezien dienen te worden als een gelimiteerde interpretatie van de werkelijkheid en enkel als 'conversation starter' ingezet zouden moeten worden.

Voor Big Data in het primaire onderwijsproces permitteren we ons hier een ruime opvatting. Dit heeft immers ook te maken met specifieke opleidingen die zijn gestoeld op de ontwikkelingen rondom (Big) Data of data science. In het primair onderwijs en voortgezet onderwijs zelf is er nauwelijks sprake van aandacht voor deze thematiek met uitzondering van het informaticaonderwijs in het vo, maar dat vak is toe aan een grondige hervorming en heeft de nodige problemen om zich als brede discipline staande te houden in het curriculum. Hoe anders is dat in het hbo, maar vooral wo. In het wo worden bestaande tracks van masteropleidingen steeds data-intensiever. Een kleine greep uit deze opleidingen is:

- Computer Science op de Radboud Universiteit Nijmegen biedt sinds 2014 een track data science aan.

[42] Dutch Association for Institutional Research.

- Computer Science aan de Universiteit Leiden heeft een aantal specialisaties waaronder bio-informatics, waarin aspecten van wiskunde, informatica met biologie en biochemie worden verbonden
- Binnen de Master Econometrics aan de UvA wordt een specialisatie gericht op Big Data aangeboden.
- De Universiteit van Twente kent een Master Business Information technology, waarbij een specialisatie is ingericht op het gebied van business analytics, waarin vakken als Managing Big Data en Data Science terugkomen.
- De masteropleidingen Business Analytics en Artificial Intelligence aan de VU hebben een behoorlijk zware 'data'-poot met vakken op het gebied van datamining, data analysis, information retrieval en data management.

Een van de weinige partijen die zich naast de educatieve uitgeverijen en de bekostigde onderwijsinstellingen zelf in het primaire proces 'mengt' is Cito, via een divers toetsingsinstrumentarium, zoals de eindtoets basisonderwijs en de centraal schriftelijke eindexamens in het voortgezet onderwijs. In feite kan hier niet eens gesproken worden van dezelfde data, maar dan big, omdat Cito al vele jaren de eindtoets basisonderwijs en eindexamens in het vo opstelt. Cito is een onafhankelijke instelling, en bestaat uit een stichting die vanuit een wettelijke taak (wet SLOA[43]) projectsubsidie krijgt vanuit het ministerie van OCW. Daarnaast heeft Cito een BV waar commerciële activiteiten zijn ondergebracht. Er wordt volop getrokken aan de data die bij Cito besloten liggen, maar Cito hanteert een steng beleid en levert geen data op individueel niveau terug. Zelfs de Onderwijsinspectie haalt in sommige gevallen de Cito-data voor de eindtoets basisonderwijs bij de po-instellingen zelf op.[44]

Nieuwe Big Data in primair proces

Tot slot kijken we naar de cel met nieuwe (Big) Data in het primaire proces. We beschouwen het primaire proces hier als alles dat direct te maken heeft met onderwijs. Enerzijds betreft dit big data als centraal onderwerp van onderwijs. Anderzijds gaat het om data die afkomstig is uit het primaire leerproces, zoals bijvoorbeeld toetsresultaten. We hanteren dus een bredere opvatting en bezien onder meer de nieuwe opleidingen / onderwijs op het gebied van Big Data / data science. Buiten het feit dat voornoemde tracks in de vorige paragraaf aan bestaande opleidingen bijna alle slechts enkele jaren bestaan, zijn er ook de nodige nieuwe onderwijsinitiatieven.

Zo start men in 2015 met de Data science track binnen de opleiding Computer Science aan de TU Delft. Een groter initiatief is het voornemen van een Graduate School Data Science & Entrepreneurship vanuit de Technische Universiteit Eindhoven en Tilburg University. Deze Graduate School beoogt drie (multidisciplinaire) Masteropleidingen aan te bieden in respectievelijk Eindhoven, Tilburg en 's-Hertogenbosch. Ook wordt voorzien een brede bachelor op te zetten in Eindhoven en Tilburg als geijkt voorportaal / koninklijke route voor deze Masteropleidingen. Ook Fontys Hogescholen is onlangs (27 oktober 2014) gestart met een Big-Data-cursus, maar die is aanzienlijk kleiner qua schaal.

Verreweg de meest relevante partijen in deze cel waar het daadwerkelijk Big Data betreft, zijn de aanbieders van adaptieve leersystemen zoals Oefenweb, Cito, Malmberg, Quayn, etc. Zoals gememoreerd in de vorige paragraaf, zijn adaptieve leersystemen erop gericht om op

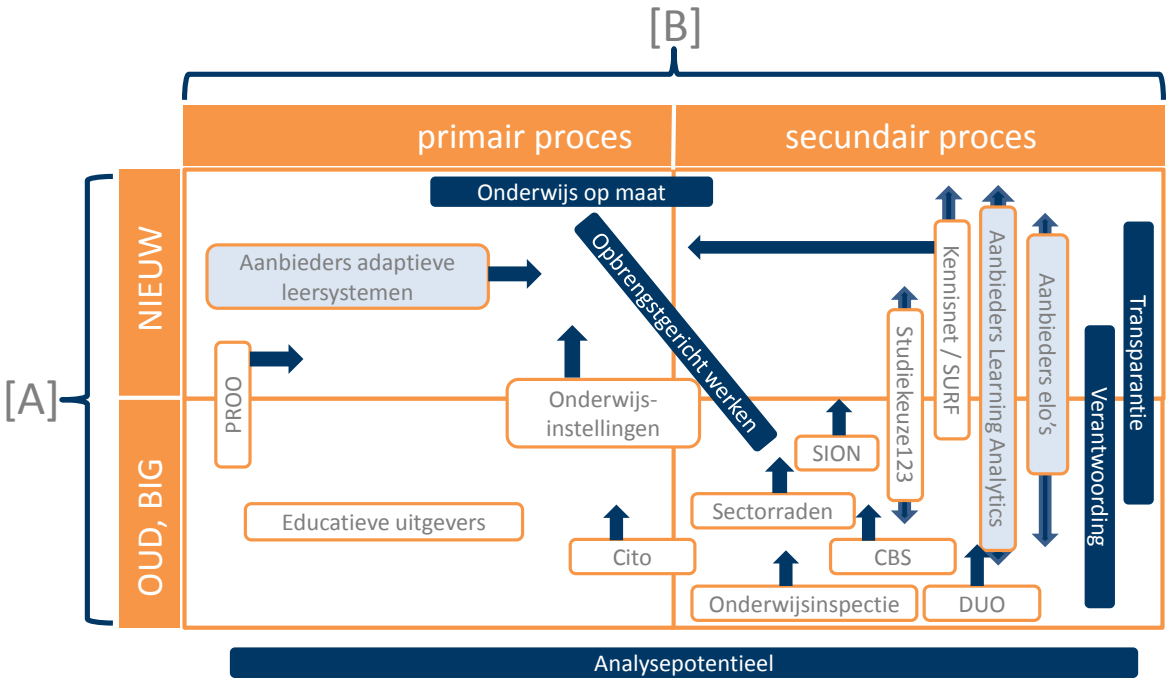
[43] Subsidiëring van landelijke onderwijsondersteunende activiteiten.

[44] De CITO-scores voor de eindtoets basisonderwijs worden overigens ook (geanonimiseerd) in het Basis register Onderwijs (BRON) opgenomen sinds 2010, maar die data gaat uit van de registratiesystemen van de po-instellingen en is niet altijd accuraat.

basis van feedback van het individu (bijv. via testjes) het leerproces op maat bij te sturen. Het niveau van het kind wordt (in de tijd) gemonitord en continu herijkt. Dit zelf-organiserende systeem heeft dus als voordeel (1) dat het kind op zijn/haar niveau opgaven / spelletjes krijgt aangereikt en (2) de ontwikkeling in de getoetste vaardigheden kan worden gemonitord. Dit genereert dus een immense hoeveelheid data en tot op heden wordt deze data beperkt gebruikt voor geaggregeerde didactische doeleinden anders dan reflectie op de prestaties van het individu. Hier ligt echter wel een grote potentie. Steeds meer commerciële partijen begeven zich op deze markt.

Een andere grote ontwikkeling die min of meer parallel loopt aan de adaptieve leersystemen is het formatief toetsen. Bij formatief toetsen wordt in tegenstelling tot summatief toetsen het (continu) bijsturen van het leerproces centraal gezet. Toetsing is geïntegreerd in het leren. Adaptieve leersystemen brengen formatief toetsen een stuk dichterbij.

Figuur 7 hieronder toont het gevulde inrichtingsraamwerk zoals dat kan worden samengesteld vanuit de dominante ontwikkelingen die we hierboven hebben besproken. Het streven is niet om hier een uitputtend overzicht te schetsen, maar vooral om enkele spelers te plotten in de 2*2 matrix samen met de aanwezigheid van het krachtenveld dat in Figuur 5 was weergegeven.



Figuur 7: Inrichtingsraamwerk voor het positioneren van (onderwijs) partijen en krachten

3.3 Dystopie Onderwijs 2025

In deze paragraaf kijken we vanuit het jaar 2025 terug op de impact van Big Data op het onderwijs. De veronderstelling in dit essay is dat Big Data in het onderwijs niet gebracht heeft van wat wij ervan verwachtten. Het gaat dus om een dystopie. We benadrukken dat de dystopie en de utopie in dit en het volgend hoofdstuk bedoeld zijn om bewustwording, kennisopbouw en dialoog over Big Data te stimuleren. De essays zijn geen toekomstvoorspelling of resultaat van (wetenschappelijk) onderzoek.

3.3.1 Onderwijs in het primair proces vanaf het begin op achterstand

In het onderwijsdomein was er aanvankelijk sprake van een ambivalente houding tegenover Big Data. In het primaire proces werd er lange tijd nauwelijks aandacht aan besteed. Debet daaraan was het feit dat veel docenten (met name in het po) zelfs de meest basale kennis van kwantitatieve methoden ontbeerden. Rekenen was al moeilijk, statistiek onbegrijpelijk. Big Data was iets dat vooral buiten het klaslokaal speelde; het was iets voor grote Amerikaanse bedrijven. Wat er tijdens lestijden allemaal ongezien via schoolcomputers en smartphones van leerlingen binnenkwam, had niets te maken met onderwijs. Slechts in ad hoc modules als mediawijsheid werd er aandacht besteed aan het mogelijke misbruik van data, dat wil zeggen van het oneigenlijke hergebruik van persoonsgegevens.

Ondertussen stonden de technologische ontwikkelingen niet stil. Met name in de ontwikkeling van adaptieve leersystemen werden grote sprongen gemaakt. Bij de inzet van deze systemen, die steeds complexer van aard werden, leidde dat tot grote problemen bij de implementatie, omdat de meeste leerkrachten noch de kennis noch de wil hadden om de mogelijkheden van de systemen volledig te benutten. De workaround van de softwareproducten was om de systemen zelflerend te maken en de aanpassingen aan de leerling zoveel mogelijk automatisch te laten verlopen. Die aanpassingen voltrokken zich dus grotendeels aan het zicht van de leerkrachten. Tegelijkertijd werd er aan de achterkant grote hoeveelheden gedetailleerde informatie over de leerlingen verzameld. Dit was nodig om de zelflerende systemen door te ontwikkelen. Binnen een tijdsbestek van enkele jaren wisten de ontwikkelaars van leeromgevingen en digitale lesmethodes al meer over leerlingen en studenten dan DUO. Veel meer. Het werd commercieel pas echt interessant toen bedrijven hun eigen informatie gingen verrijken met data uit andere bronnen en op maat adviezen aan leerlingen en hun ouders gingen uitbrengen.

Met betrekking tot het secundaire proces verliepen de ontwikkelingen heel anders. In het streven naar opbrengstgericht werken werden de mogelijkheden van Big Data juist enthousiast omarmd. Voor het eerst was het mogelijk om een grote set aan kengetallen bijna continu te meten en bijna real-time het rendement van onderwijsinstellingen te volgen. Dat maakte het mogelijk om steeds sneller bij te sturen op uitval en slechte toetsresultaten. Het management van de instellingen was zich zeer wel bewust van de grotendeels inzichtelijke maar permanente controle van bovenop. Op hun beurt hielden zij de *performance* van vakgroepen en individuele leerkrachten nauwgezet bij.

3.3.2 Simplistische toepassing van Big Data

In de jaren 2015-2020 stond het Big Data regime nog in de kinderschoenen. Maar er werd in deze periode nogal simplistisch met data omgegaan. Omdat een duidelijk beeld op de mogelijkheden – en vooral de onmogelijkheden – van Big Data ontbrak en omdat de kosten van opslag in die tijd nog laag waren, werd er ongebreideld van allerlei soorten data verzameld. *Catch as catch can*. Men wist vooraf immers nooit waar de data allemaal voor konden worden toegepast. Dat was, volgens de Big-Data-goeroes van die tijd, nu juist de grote belofte: dat er patronen konden worden ontdekt en verbanden worden gelegd die vooraf niemand had kunnen bedenken. De goudkoorts voor data leidde tot een hausse aan pilots en projecten rond Big Data. Consultancybedrijven en onderzoeksbureaus verdrongen zich voor de bureaus van onderwijsmanagers en beleidsmakers. Allemaal claimden ze waardevolle inzichten uit de data-analyses te kunnen halen. Deze claims bleken later weliswaar overdreven – maar het (doen) ontstaan van dit soort hooggespannen verwachtingen was nu eenmaal inherent aan de dialectiek van de introductie van nieuwe

technologieën.[45] In technologisch en technisch opzicht werden er wel grote stappen gezet op het terrein van data handling en datamining.[46] Debet aan de toenemende complexiteit van de analysemethoden en de steeds grotere hoeveelheid data die werd verzameld was het feit dat er grote druk ontstond om met verrassende inzichten uit de data te komen. Als de berg mest maar groot genoeg was vond je immers altijd wel een keer een gouden munt. Anders gezegd: als er maar genoeg data werd verzameld en genoeg verschillende analyses werden uitgevoerd, kwam er altijd wel een opvallend verband uit naar voren. De bottleneck zat echter niet in de ICT en methodologie, maar in het gebruik aan domeinkennis waarvan het belang toen nog niet zo werd ingezien. In de wedloop naar betere en nog dieper gaande analyses was een specialisatie en verkokering ontstaan die ten koste ging van het belang van domeinkennis en 'gezond verstand'. De whizzkids werden louter afgerekend op significante en ogenschijnlijk doorwrochte analyses, gepresenteerd in oogstrelende rapportages. De kloof tussen hen en het veld waarover zij uitspraken deden groeide echter.

Dat leek in eerste instantie ook geen probleem – de methoden waren immers generiek en voor de validatie van de data kon een beroep worden gedaan op materiedeskundigen uit het veld. Probleem was dat deze deskundigen lang niet altijd de empirische kennis hadden die de mensen op de werkvloer (hier: de leerkrachten) wel hadden en ook niet bepaald representatief waren voor de mening van de docenten. En net als de leerkrachten waren ook deze experts niet in staat om de complexe Big-Data-technieken te volgen. Ze hadden eigenlijk geen idee wat er met de data gebeurde, en welke waarde ze nu precies moesten hechten aan de indrukwekkende classificaties en visualisaties die hen werden voorgelegd.

Bevindingen van Big-Data-analyses moesten echter altijd in de specifieke context van de toepassing worden geplaatst en dit vereiste grondige (ervarings)kennis van die context. Patronen en correlaties konden niet zomaar worden vertaald naar conclusies. *It's human nature to see only what we expect to see*.[47] En met de enorme hoeveelheden data ter beschikking werden er altijd wel (gewenste) patronen gevonden. *Apophenia* – patronen zien die er eigenlijk niet zijn – was schering en inslag. Er werden in deze tijd tal van conclusies getrokken die niet relevant waren of zelfs ronduit misleidend. In eerste instantie leidde dit nog niet tot grote problemen. Veel Big-Data-projecten waren nog in een pilotfase en de aanbevelingen waren vrijblijvend – er werd meestal nog geen concreet gevolg gegeven aan de (soms harde) conclusies ten aanzien van de *performance* van onderwijsinstellingen of van individuele leerkrachten en leerlingen. Daarnaast waren de meeste analyses nog vrij beperkt van aard; ze richtten zich meestal op rendement in de enge zin van het woord. Voor dit soort beperkte kwantitatieve analyses was de rijkere context nog relatief onbelangrijk voor de duiding. Geleidelijk aan bleek dat de scope van de analyses te beperkt was. De performance van leerkrachten en leerlingen bleek veel moeilijker te modelleren dan aanvankelijk was aangenomen. Er bleken veel meer (voornamelijk kwalitatieve) variabelen van belang te zijn. Dat zorgde er wel voor dat de juiste interpretatie van de data nu een nog heikelere onderneming werd dan ze altijd al was geweest. Tegelijkertijd nam vanuit de politiek de druk toe om ook meer rendement te halen uit de (dure) Big-Data-exercities. Er zou, net als in

---

[45] R.A. te Velde (2004), Schumpeter's Theory of Economic Development Revisited: About True Entrepreneurship, High Bandwidth, False Hopes, and Low Morale, In: T. Brown en J. Ulijn (red.), Technology, Entrepreneurship and Culture, Edward Elgar: Cheltenham UK and Northampton, MA, USA; R.A. te Velde en B. Pegge (2011). De Nederlandse softwaresector: ICT als innovatie-as. In: CBS (eds.) ICT, kennis en economie 2011. Den Haag: CBS (pp. 214-225).

[46] Zoals de ontdekking van het Cassina-Kovacs-Kolman-theorema in 2016, dat ten grondslag heeft gelegen aan een geheel nieuwe generatie van random forest classification models.

[47] S. Klous en N. Wielaard (2014). Wij zijn big data. De toekomst van de informatiesamenleving. Amsterdam/Antwerpen: Business Contact (p. 143).

bijvoorbeeld de Verenigde Staten, veel meer moeten worden gedaan met de uitkomsten van al het onderzoek. Na de motie Kalsbeek (2021) werden de continue (data-driven) beoordelingsrondes voor leerkrachten en leerlingen verplicht gesteld.

3.3.3 Probabilistische kloof groeide

Sluipenderwijs ontstond er in het onderwijsdomein een tweedeling tussen zij-die-analyseren en zij-die-worden geanalyseerd, tussen de leerlingen en leerkrachten (die niet of nauwelijks van Big Data gebruik maken) en de combine van managers, beleidsmakers en bedrijven (die op steeds grotere schaal gebruik gingen maken van Big Data). De tweedeling vertaalde zich ook door in de polarisatie van standpunten ten opzichte van privacy. Leerkrachten stonden huiverig ten opzichte van het verzamelen en hergebruiken van gegevens, maar omdat niemand precies wist wie wat verzamelde, wat ermee gebeurde en wat de (on)mogelijkheden waren, bleven de spanningen lange tijd onderhuids. Pas nadat er steeds meer gevallen optraden van bindende beoordelingen op basis van evident onjuiste analyses barstte de bom. Nu kwam ook de volle omvang van de data-verzameling – en de vaak verstrekkende conclusies die aan de analyses werden verbonden – naar boven. Het was onduidelijk wie welke (cruciale) data verzamelde en beheerde, waar deze data (allemaal wel niet) voor werden gebruikt en hoe (incorrecte of verouderde) data met terugwerkende kracht konden worden uitgevoerd. De publieke verontwaardiging was in eerste instantie groot. Er bleek niet of nauwelijks relevante wetgeving te zijn en de wetgeving die er was werd nauwelijks gehandhaafd door de overheid, die zelf zo langzamerhand door het woud aan statistische mogelijkheden de beslisbomen niet meer zag. De sturingsmogelijkheden van de overheid waren sowieso beperkt omdat het gros van de onderwijsdata inmiddels in handen van private partijen was gekomen. Onderzoeksjournalisten doken op de zaak en brachten ettelijke gevallen van misbruik van privacygevoelige informatie aan het licht. Er werd openlijk burgerlijk verzet gepleegd tegen het verzamelen van onderwijsdata. Puristen binnen de beweging (de 'gebeten digi's') pleitten voor een algeheel verbod in het klaslokaal op alles elektronische apparatuur die in verbinding stond met de buitenwereld.

Dat ging de meeste Big-Data-antagonisten te ver. Geleidelijk aan ontstond er een bredere consensus om ICT in het primaire proces alleen nog maar te gebruiken voor de meest basale functies: het doorvoeren van rapportcijfers, agendabeheer en emails. Alle lesmethoden moesten offline. Data zouden alleen nog maar lokaal moeten worden opgeslagen en alleen de leerkracht had toegang tot de data van haar of zijn leerlingen. Omdat men als de dood was dat gegevens alsnog op straat kwamen te liggen en verder werden verspreid zou die mogelijkheid bij de wortel moeten worden uitgeroeid: alle meetcycli zouden verplicht moeten worden vernietigd nadat ze door de leerkracht waren gebruikt. Eén school haalde het nieuws nadat zij via een antiquair een 30-tal gereviseerde Commodore 64's had aangeschaft. Zo konden digitale vaardigheden worden onderwezen zonder risico op ongeoorloofde dataverzameling.

Het bleek echter allemaal *too little and too late*. Anno 2020 was de samenleving al volledig doordrenkt met data. De protagonisten betoogden met succes dat het benutten van de mogelijkheden van Big Data allang niet meer een keuze maar een noodzakelijke voorwaarde was voor elke moderne samenleving. Het was buitengewoon naïef om te denken dat we nog terug konden keren naar een wereld van vóór de Big-Data-revolutie. Zo wezen verschillende commentatoren fijntjes op het feit dat de onderzoeksjournalisten die de privacy-debacles aan het licht hadden gebracht, zelfs op grote schaal gebruik hadden gemaakt van Big-Data-technieken. Datzelfde gold voor de leerkrachten en leerlingen, die buiten het klaslokaal enthousiast gebruik maakten van op maat gemaakte online diensten die op basis van Big Data waren ontwikkeld en continu werden doorontwikkeld.

De politiek trok echter weldegelijk het boetekleed aan. Allerwege werd toegegeven dat er tot dan toe te veel ongerichte data was verzameld, en dat er te veel analyses waren uitgevoerd die van onvoldoende kwaliteit waren. *Niet meer maar beter* werd een gevleugelde kreet. Ook ging men mee in de kritiek dat de overheid tot nu toe gebrek aan visie op Big Data had. De overheid zou zich duidelijker moeten uitspreken dan voorheen. En dat doet ze ook.

Zo stelde ze ten aanzien van de privacy dat de issues technisch uitstekend op te lossen waren en in de praktijk ook al grotendeels opgelost waren. De oorzaken lagen veel meer in de beeldvorming rond de opgeblazen 'data-schandalen' dan in de techniek zelf. Een betere publieke voorlichting was de logische remedie. Dat gold volgens de overheid ook voor de bedrijven die in de markt voor onderwijsdata opereerden. Die zouden veel actiever moeten zijn met pre-marketing om de koudwatervrees bij hun databronnen c.q. gebruikers weg te nemen. Ten slotte zouden er zowel voor publieke als voor private partijen veel strengere eisen worden gesteld aan transparantie zodat een ieder wist wat er met zijn of haar data gebeurde.

3.3.4 Reductionistische aanpak domineerde

Ten aanzien van Big Data in het algemeen had de overheid een hele duidelijke visie: het gros van de problemen rond het gebruik van Big Data in het onderwijsdomein was niet veroorzaakt doordat er te veel maar juist *nog te weinig* in Big Data was geïnvesteerd. Die investeringen moesten juist worden geïntensiveerd en dan zou het onderwijs (eindelijk) de zoete vruchten plukken van Big Data – de positieve synthese in de dialectiek van beloftevol onderzoek.[48] Het gebruik van Big Data moest naar een hoger niveau worden getild. Data moest selectiever worden verzameld en er moesten de juiste verbanden tussen datasets worden gelegd. Een belangrijk beleidsvoornemen was om de scope van de analyses verder te vergroten – die hadden tot nu toe immers een te eng, kwantitatief karakter gehad. Het nieuwe beleid ging uit van een holistisch perspectief op de mens, in de volle rijkheid van diens bestaan.

Om tot betere voorspellingen te komen ten aanzien van het leersucces van leerlingen en het didactische vermogen van docenten betekende dit onvermijdelijk dat er nu ook andere soorten van ongestructureerde data werden verzameld en geanalyseerd die men tot dan toe, vanwege de complexe aard van de data en de potentiële privacygevoeligheid, links had laten liggen. Het gedrag van mensen kon alleen goed worden beschreven door de gevonden patronen te plaatsen in de context van de persoonlijke sfeer. Eén van de consequenties was dat er steeds meer achtergrondgegevens over leerlingen en leerkrachten werden verzameld. In de vigerende privacywetgeving vond men hiervoor de ruimte. Er was immers nog steeds sprake van doelbinding; alle data werden verzameld met het doel om de leer- c.q. didactische vaardigheden te verbeteren. Een andere consequentie was dat het gedrag van ouders (in het geval van leerlingen) en van partners (in het geval van docenten) nu ook een belangrijk onderdeel van de analyse werd. Aanvankelijk stuitte dit nog op enig juridisch verzet – de privacy van deze derde personen moest immers worden beschermd. Na de uitspraak van het Hof in de zaak Zelle werd het gebruik van data over derden in het laatste geval niet toegestaan – in het eerste geval wel voor zover het waargenomen gedrag van de ouders direct verband had met de leerprestaties van hun kind.

---

[48] "[…] if only additional investments are made the technology will be able to fulfil its promises. Skilful entrepreneurship requires the timely formulation of antitheses in such a way that the additional resources are geared towards the resolution of the critical problem – the 'reverse salient' that hamper further diffusion of the innovation." R.A. te Velde (2014) op.cit., p.119-120.

Deze uitspraak was ook in moreel opzicht van groot belang. Zij onderstreepte het groeiende besef dat ouders medeverantwoordelijk waren voor het streven om een maximaal maatschappelijk rendement uit hun kind te halen. Maximaal in de zin dat de kosten van de opleiding in verhouding moesten staan tot de (geschatte) opbrengsten over het gehele werkzame leven (20 tot 70 jaar). Die opbrengsten konden, op basis van geavanceerde Big-Data-analyses, steeds nauwkeuriger worden bepaald. Uit deze opbrengsten konden ook de maximale opleidingskosten voor een kind worden bepaald. Vanuit het maatschappelijk economisch oogpunt was noch een onderinvestering noch een overinvestering in het kind c.q. de leerling gewenst. Men vroeg zich af het wel verstandig was om mensen met studieleningen op te zadelen die – gezien hun achtergrond - deze waarschijnlijk niet terug zouden verdienen. Mensen die voor een dubbeltje geboren waren, werden immers zelden kwartjes.

De vraag naar onderwijs- en loopbaanadviezen, die sinds 2015 al gestaag was gegroeid, nam nu een hoge vlucht. Elke ouder wilde – binnen het gegeven wettelijke opleidingsbudget - komen tot het meest optimale opleidingsaanbod. De private aanbieders van leer- en loopbaanadviezen zagen hun markt nog verder groeien. Er vond een consolidatie van de markt plaats en er bleef nog een half dozijn grote bedrijven (*The Big Six Sigma's*) over die alle internationaal opereerden en zo over gigantische datasets beschikten. Deze bedrijven maakten gebruik van steeds geavanceerdere predictive statistics om risicoprofielen van hun cliënten te kunnen kwantificeren en ze in nog preciezere categorieën in te kunnen delen.

De differentiatie in vraag naar opleidingen zette zich ook onvermijdelijk door in een differentiatie in het aanbod. De verschillen in aanbod en kwaliteit tussen scholen en onderwijsinstellingen namen in korte tijd sterk toe. Om zich te onderscheiden van de duizenden andere opleiders positioneerden alle opleidingen zich op een specifieke plaats op de continue arbeidsschaal. Aan de bovenkant van de markt waren de opleidingen van internationaal topniveau, maar de ouderbijdrages waren ook navenant hoog. Aan de onderkant werden voor bescheiden budgetten basale vaardigheden aangeleerd. Om vertraging te voorkomen, werkten instellingen binnen de verschillende kwaliteitsniveaus steeds nauwer met elkaar samen. Zo ontstonden er tienduizenden unieke opleidingspaden, maar daarbinnen was er niet of nauwelijks ruimte voor creatieve ontplooiing van de leerling. De opleidingen stonden immers onder grote druk van de ouders om het maximale rendement uit hun kind te halen. Bij vertraging of, in het extreme en ongewenste geval, uitval, bleven de ouders met hoge kosten zitten. Uiteraard kwamen deze kosten dan geheel voor hun eigen rekening. Ze hadden vooraf immers precies kunnen weten wat het optimale opleidingspad voor hun kind was. Dat had één van de Big Six Sigma's ze precies kunnen vertellen.

3.4 Utopie Onderwijs 2025

In deze paragraaf kijken we vanuit het jaar 2025 terug op de impact van Big Data op het onderwijs. De veronderstelling in dit essay is dat Big Data in het onderwijs gebracht heeft wat wij ervan verwachtten.

3.4.1 Big Data in dienst van beter onderwijs

Het belang van personalisatie was al twee decennia eerder erkend in de marketing.[49] Uiteindelijk groeide ook zowel bij de leerkrachten als bij het management het besef dat Big Data zeer veel mogelijkheden bood tot het verbeteren van de kwaliteit van het onderwijs. Die verbeteringen zaten met name in het aanbieden van onderwijs op maat. Die belofte werd

[49] Pine II, J. (1992). Mass Customization: The New Frontier in Business Competition. Boston, Mass.: Harvard Business School. ISBN 0-87584-946-6.

met behulp van Big Data waargemaakt. Steeds meer producten en (online) diensten werden op basis van Big Data op maat aangeboden. Vanwege de lengte en de intensiteit van het contact tussen de aanbieder (de onderwijsinstelling) en de afnemer (de leerling/student) leek personalisatie bij uitstek geschikt voor het onderwijs. En dat bleek ook al snel zo te zijn.

De personalisatie had zowel betrekking op de inhoud van de lesstof als op de wijze waarop de lesstof werd aangeboden. Leidende gedachte was dat leerlingen en studenten in grote mate een individueel traject konden volgen. Deze trajecten werden ondersteund door adaptieve leersystemen die geheel in dienst stonden van het leerproces. Via tal van experimenten onder auspiciën van het ministerie van OCW leidde dit binnen enkele jaren tot een afname van het traditionele klassikale onderwijs. Andersom maakte de aloude studiehuisgedachte een revival, maar nu met een twist. In de oorspronkelijke opzet bleek dat er noodzakelijkerwijs een trade-off was tussen breedte en diepte. De balans leek toen te zijn doorgeschoten naar breedte. Destijds had dit tot klachten geleid van ho-instellingen over het – al dan niet vermeende – te lage kennisniveau van instromende vo-studenten. In de nieuwe opzet was het juist mogelijk om veel selectiever te zijn in het aanbod van de lesstof, en op bepaalde vakken veel meer de diepte in te gaan. Hiervoor was wel een radicale verandering nodig in de formele definitie van de eindkwalificaties. De eis om over een breed palet aan vakken een voldoende te scoren werd losgelaten. In plaats daarvan werd louter gestuurd op het totaal aantal behaalde studiepunten. Leerlingen konden dat aantal ook halen door in een beperkt aantal vakken (extreem) hoge cijfers te halen.

Het loslaten van de eis tot een breed (en dus noodgedwongen relatief homogeen) eindexamenpakket leidde tot een onvermoede specialisatie aan de kant van de ontvangende partijen. Al snel gingen veel ho-instellingen ertoe over om hun toelatingseisen ook veel preciezer te definiëren dan in termen van de traditionele brede studieprofielen. Wie een bepaalde specifieke studie wilde doen, moest aan specifieke eisen voldoen. Dit was minder verplichtend dan het leek: omdat het aantal specifieke opleidingen sterk toenam, was er voor individuele leerlingen nog steeds veel te kiezen. Het aantal generieke opleidingen nam wel sterk af, zodat bedrijven veel beter wisten welk vlees ze in de kuip hadden. Dit leidde tot een betere aansluiting met de arbeidsmarkt. In vacatures werden steeds vaker specifieke studies of zelfs specifieke instellingen genoemd – voor zover er nog vacatures werden geplaatst, want in veel gevallen hadden bedrijven direct contact met hun 'huisleveranciers'.

De specialisatie en differentiatie zetten zich geleidelijk aan ook naar beneden toe door. In 2019 werden de generieke eindkwalificaties voor het po losgelaten en in 2021 werden de eisen verder versoepeld. Dat betekende dat po-leerlingen zich, na een gedegen basisopleiding, vanaf groep 6 konden toeleggen op de vakken waar ze het beste in waren. Leerlingen die beter waren in rekenen dan in taal konden vanaf hun 9^{e} jaar dus al voorsorteren voor een bepaalde studierichting. Het karakter van het onderwijs veranderde langzaam maar zeker. Summatief toetsen verdween nagenoeg geheel en feitenkennis werd veel minder belangrijk. Domeinkennis daarentegen bleef echter belangrijk en bleek, hoe meer Big Data in het (hoger onderwijs) leren zelf werd verweven, ook steeds belangrijker omdat zij een essentieel ingrediënt is voor correcte data-analyses. Bij het verwerven van domeinkennis werd wel steeds meer gebruik gemaakt van actief leren (praktijkopdrachten om *hands-on experience* op te kunnen doen) in plaats van passieve (e)boekenkennis.

3.4.2 Brede verspreiding van Big-Data-vaardigheden in het onderwijs

De focus kwam sterk te liggen op het kunnen stellen van goede onderzoeksvragen en het kunnen leggen van verbanden – om kennis op een originele manier te kunnen combineren om zo steeds nieuwe vraagstukken op te kunnen lossen. In onderwijsconcepten vertaalde zich dat onder andere in het benadrukken van harde analytische vaardigheden maar ook van

'soft skills' zoals creativiteit (om ogenschijnlijk losse feiten met elkaar te kunnen verbinden) en communicatieve vaardigheden (om verbindingen met mensen van buiten het eigen 'kennis-ecosysteem' te kunnen maken, en om daarmee te kunnen samenwerken).[50] Samenwerken werd het nieuwe '*Leitmotiv'* in onderwijsland. Immers, kennis lokte kennis uit. Er werd door professionals en onderwijsexperts aanvankelijk volop geëxperimenteerd met verschillende vormen van samenwerking tussen verschillende groepen (leerling, studenten, wetenschappers, experts uit het bedrijfsleven) en op verschillende schaalniveaus (binnen opleidingen, tussen opleidingen, tussen landen). Door wél duidelijk aan te geven *dat* er samengewerkt moest worden maar niet *hoe* dat moest gebeuren liet het ministerie aanvankelijk duizenden bloemen bloeien. Uiteindelijk kristalliseerden van onderaf een beperkt aantal ideaaltypische manieren van samenwerking uit die voor bepaalde contexten en typen kennis het meest geschikt bleken te zijn.

Vanaf 2017 nam formatief toetsen een hoge vlucht. Om leerlingen al in een dergelijk vroeg stadium de beste keuze te kunnen laten maken, werden er in steeds vroegere stadia op basis van *learning analytics* gedetailleerde profielen gemaakt van de specifieke capaciteiten van een leerling. Deze profielen dekten een breed spectrum aan capaciteiten en vaardigheden (dus niet alleen de cognitieve vaardigheden, maar ook sociale, verbale en motorische vaardigheden). Ze gaven een verwachting van de verdere ontwikkeling van de leerling in de volgende tien (na 2022 vijftien) jaar en werden voortdurend bijgesteld op basis van de meest recente ontwikkelingen. Met andere woorden, de profielen stuurden weliswaar de individuele leertrajecten maar waren verder volgend aan de natuurlijke ontwikkeling van de leerling (adaptieve leersystemen). De profielen waren slechts een aanbeveling – ouders en de leerlingen zelf konden er te allen tijde van afwijken en bleven in 'the driver seat'. Het aanbod van het onderwijs werd dan ook op basis van de individuele amendementen bijgesteld. Drie jaar na de invoering werd de wijzigingsprocedure wel minder vrijblijvend gemaakt. Als uit de monitoring bleek dat de amendementen niet tot de resultaten leidden die de leerling en ouders zelf voor ogen hadden, werd het aanbod alsnog weer bijgesteld (*evidence-based education).*

3.4.3 Leerling bleef eigenaar van zijn of haar gegevens

Het Individuele Studenten Portal (Instupo), dat in feite de poort was tot een groot, gepersonaliseerd informatiesysteem, werd in 2018 landelijk ingevoerd. Het werd in Nederland gehost in een betrouwbare en veilige omgeving. De functionaliteiten werden daarna voortdurend uitgebreid. Instupo werd zowel gebruikt voor de uitvoer (online onderwijsmateriaal op maat) als voor de invoer van data. De individuele keuzes die werden gemaakt in het aangeboden materiaal – maar ook in het dagelijkse leven (patronen van mobiliteit en aanwezigheid, communicatiepatronen) – werden gebruikt om de inhoud en vorm van het onderwijs nog beter op maat te maken voor het profiel van de leerling.

Leerlingen en studenten konden op Instupo niet alleen voortdurend zien welke progressie ze boekten maar ook alle gegevens controleren. Het principe van Instupo was (en is nog steeds) dat studenten de eigenaar waren van hun eigen data en daar ook altijd controle over hielden. Zij bepaalden wie er toegang kreeg tot de data – en wie niet. Ze konden vanaf de allereerste invoering van Instupo dus ook hun ouders (vanaf 12 jaar) en leerkrachten (vanaf 16 jaar) uitsluiten. In eerste instantie werd dit ultieme recht van datazelfbeschikking weggehoond maar al snel bleek de vrijheid om zich te allen tijde te kunnen onttrekken van het delen van informatie juist bij te dragen aan verantwoord datagedrag.

[50] Ibid.

Het leidde alras tot een omslagpunt in het denken over privacy, dat rond 2010 nog in een scherpe patstelling verkeerde tussen de voorstanders van het vrijelijk hergebruik van data (met name Big-Data-multinationals die in die tijd nog zeer veel invloed hadden) en privacybeschermers (die zich principieel verzetten tegen elke vorm van hergebruik). Dat denken werd nu veel genuanceerder; omdat mensen van jongs af aan werd geleerd om zelf verantwoordelijk te zijn voor hun data, leerden ze de meerwaarde van het delen kennen en ook de kosten van het *niet* delen van de data. Mensen – en de organisaties waar ze deel van uitmaakten – gingen een breder spectrum aan belangen meewegen en zochten van geval tot geval een goede balans tussen het al dan niet toegang geven tot hun data. Anno 2025 konden we ons niet meer voorstellen dat het afschermen van data de default was, in plaats van het delen. Toch was dit echt de situatie tot in de tweede helft van de jaren '10. Deels had dat natuurlijk ook te maken met de stand van de techniek; *provenance technology* stond toen nog in de kinderschoenen. Men had in 2015 nog niet het overzicht wie wanneer tot welke data toegang had gekregen. De techniek was leidend. Dit gebrek aan overzicht leidde waarschijnlijk tot het breed gedeeld gevoel van onbehagen dat de samenleving niet (meer) *in control* was wat betreft het (her)gebruik van Big Data. Hergebruik dreigde toen zelfs bijna volledig juridisch te worden dichtgespijkerd.

Het was vanaf het begin duidelijk dat het opstellen van de profielen hoge eisen stelde aan zowel de beschikbaarheid als aan de interpretatie van de data. Ten eerste waren er zeer grote datasets nodig (om zo precies mogelijk te kunnen matchen op de achtergrondkenmerken van leerlingen). Ten tweede vereiste de analyse scherpe en kritische datacompetenties (om de onvermijdelijke verdachte verbanden en misleidende patronen in de enorme datasets te herkennen) mét een grondige kennis van het onderwijsdomein (om de bevindingen in de juiste context te kunnen plaatsen). Gegeven het grote maatschappelijke belang van de onderwijsprofielen – ze bepaalden immers in principe in hoge mate de progressie van een leerling – stelde de overheid vanaf het begin de profielen in eigen beheer op. Zij borgde niet alleen de strikte onafhankelijkheid van de adviezen en de transparantie van de procedures, maar investeerde ook mee in het opstellen. Daardoor konden de beste dataspecialisten worden aangetrokken en waren de analyses van hoge kwaliteit.

Het beleid had een stevig economisch fundament: het 'Dutch model' leidde tot structureel minder maatschappelijke kosten vanwege lagere dropout en churn rates, en tot hogere opbrengsten vanwege een betere matching tussen vraag en aanbod op de arbeidsmarkt. Vanaf medio 2020 vond het internationaal allerwege navolging. De concurrentie om de dataspecialisten was in de laatste jaren daarvoor toegenomen, maar vanwege het first mover advantage van Nederland leek het land een aantrekkelijke vestigingsplaats te blijven voor *learning analytics* specialisten en bedrijven. Dit was een voorsprong die we nog zeker tien jaar konden vasthouden.

4 Inventarisatie en essay wetenschap

4.1 Inleiding

Dit hoofdstuk is een beschouwing op het gebruik en de impact van Big Data op de wetenschap. Het hoofdstuk begint met een kwalitatieve beschrijving (en waar mogelijk kwantitatief) van het huidige gebruik van Big Data in de wetenschap (en ontwikkelingen daarbinnen). Vervolgens presenteren we de dystopie van Big Data in de wetenschap in 2025. Dit is een scenarioschets waaruit blijkt dat Big Data niet heeft gebracht wat we ervan verwachtten. Ergens tussen 2015 en 2025 is een 'verkeerde' beslissing gevallen of hebben negatieve krachten de positieve krachten van Big Data gedomineerd. In het laatste deel van dit hoofdstuk presenteren we de utopie van Big Data in de wetenschap in 2025. Dit is een scenarioschets waaruit blijkt dat Big Data wel heeft gebracht wat we ervan verwachtten. Ergens tussen 2015 en 2025 is een 'goede' beslissing gevallen of hebben positieve krachten de negatieve krachten van Big Data gedomineerd.

4.2 Inventarisatie datagebruik in de wetenschap

In tegenstelling tot het onderwijs heeft Big Data tot nu toe de meeste invloed gehad op het primaire proces (het 'doen' van wetenschap), [nog] nauwelijks op het secundaire proces ([aan]'sturen' van wetenschap). Alle trends die we hierna beschrijven hebben dan ook betrekking op het primaire proces. In onze inventarisatie beschrijven we een aantal stimulerende en afremmende krachten voor het huidige en toekomstige gebruik van Big Data in de wetenschap.

4.2.1 Stimulerende krachten voor gebruik (Big) Data in wetenschap

Hieronder benoemen we de belangrijkste stimulerende krachten voor intensiever gebruik van data in de wetenschap binnen Nederland.

Toename beschikbaarheid van data

Internationaal gezien loopt Nederland voorop bij het gebruik van Big Data in het primaire proces van de wetenschap. De goede uitgangspositie van Nederland geldt voor technische disciplines (zoals astronomie en klimaatwetenschapen), maar vooral in de toepassing van data science in geesteswetenschappen en sociale wetenschappen.[51]

Disciplines zoals astronomie en klimaatwetenschappen hebben al een lange traditie in Big Data. Al decennialang worden daar grote hoeveelheden data verzameld, gedeeld en

[51] Nederland is vooral sterk in publiek-private samenwerkingen. In astronomie heeft dat geleid tot sterke clusters rond ASTRON (Target/Astro-WISE project), in klimaatwetenschappen tot clusters in waterbouw (Digitale Delta, Predictive dikes, IJdijk) en klimaatmodellering (groep rond prof. Henk Dijkstra, eSALSA). Andere leidende programmaas zijn CTMM/TraIT en GENALICE (beiden in life science & health). In academisch onderzoek heeft Nederland specifieke sterktes in beeldherkenning (UvA Intelligent Systems Lab) en business process mining (TU/e, groep rond prof. Wil van der Aalst). Het nationale big-data-onderzoeksinstituut (NLeSC – zie hierna) richt zich vooral op de toepassing van big-data-technologie in sociale wetenschappen en humanities. Het CBS loopt wereldwijd voorop in het gebruik van big data ('Internet als Databron', zie onder andere Dialogic (2007) Go with the dataflow! Analysing the Internet as a data source (IaD). http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2008/05/13/go-with-the-dataflow-main-report%5B2%5D.html).

geanalyseerd. Nederland blinkt juist uit in deze disciplines en is daarom uitstekend gepositioneerd voor het toepassen van Big Data.

Nederland had van oudsher al een sterke traditie in 'digital social sciences'. Dat momentum is bij de bredere verspreiding van computers verloren gegaan (ook leken konden nu digitale toepassingen gebruiken – dus het onderscheidende vermogen van wetenschappers verdween), maar het is nu opnieuw aan een opmars bezig.[52] Deze is anders van aard en niet gestoeld op het gebruik van IT (hardware), maar op het ontsluiten van grote hoeveelheden ongestructureerde data. Dat laatste komt dan weer voort uit de voortschrijdende digitalisering en daarmee beschikbaarheid van nieuwe data:

- Steeds meer informatiebronnen (zoals historische bronnen: krantenarchieven vanaf 1800) kunnen worden gedigitaliseerd. Dat geldt evenzeer voor andere typen van informatiebronnen (audio, video) die voorheen niet goed konden worden ontsloten.
- Een tweede grote nieuwe bron voor sociale wetenschappers is de ongestructureerde data die wordt gegenereerd in 'social media' (Facebook, Twitter, LinkedIn, …) en op internet.
    - Daarmee samenhangend is de opkomst van *citizen science*: het op grote schaal laten uitvoeren van onderzoeksgerelateerde taken zoals observaties, metingen of berekeningen (ook hier komt deze beweging voort uit de astronomie). Deze beweging bestond al decennia, maar door de opkomst van de 'social media' is het nu mogelijk om op veel grotere schaal leken te mobiliseren.
- Een derde bron is overheidsinformatie zoals data uit registers en administraties. Dit is data die van oudsher al werd verzameld door overheden maar voor een specifiek administratief doel. Nieuw is dat deze data vanaf 2010 in toenemende mate voor *hergebruik publiek ter beschikking wordt gesteld*. Aangevoerd door landen als het VK, VS en Australië is het (her)gebruik van overheidsinformatie ('Open Data') wereldwijd hoog op de politieke agenda terecht gekomen.[53] Dit wordt enerzijds gestimuleerd door de gedachte dat openheid en transparantie bijdragen aan de bevordering van de democratie en de efficiëntie en effectiviteit van de overheid. Uitgangspunt daarbij is dat de overheid informatie proactief vrijgeeft.[54] De open-data-beweging heeft nu zoveel momentum gekregen dat ook bedrijven steeds vaker data publiek voor hergebruik ter beschikking stellen ('corporate data sharing').[55]
- Een vierde nieuwe bron is de veelheid aan data die een directe neerslag is van menselijk gedrag (call detail records etc.). Deze ontwikkeling is in een stroomversnelling geraakt door de opkomst van het 'Internet of Things': steeds meer

[52] Interview met prof. Van Eijnatten (Universiteit Utrecht, Onderzoekinstituut voor Geschiedenis en Kunstgeschiedenis), mimeo.

[53] Bongers, F., J. Veldkamp, T. van der Vorst, F. Verschoor, M. de Vries (2012). Open Data Open Doel. Verkenning van de hergebruikkansen van EL&I datasets. Utrecht: Dialogic / The Green Land. http://www.rijksoverheid.nl/bestanden/documenten-en-publicaties/rapporten/2013/02/28/open-data-open-doel-verkenning-van-de-hergebruikkansen-van-el-i-datasets-bijlagen/dialogic-2012-034-eli-inventarisatie-datasets-eli-bijlagen.pdf

[54] Zie onder andere De Digitale Agenda EL&I 2011 en de brief van het Ministerie van Binnenlandse Zaken & Koninkrijksrelaties over Hergebruik en Open Data (30 mei 2011) en meer recent de brief van het Ministerie van Onderwijs, Cultuur & Wetenschappen over Open Access van publicaties (15 november 2013).

[55] Zie onder andere het Global Pulse programma van de VN; http://unglobalpulse.org/mapping-corporate-data-sharing

apparaten worden aan het internet gekoppeld.[56] Er komen dus steeds meer sensoren in de maatschappij waarmee fysieke processen ('de slimme dijk') en sociale processen near-real time kunnen worden gevolgd. Anders dan bij citizen science spelen burgers hier dus een passieve rol. Hun gedrag wordt geregistreerd zonder dat ze het zelf vaak door hebben (en zonder de zekerheid dat er ook daadwerkelijk iets met deze gegevens zal gaan gebeuren). Tot nu toe waren de meeste 'things' in eerste instantie niet bedoeld om menselijk gedrag te observeren maar daar worden ze wel in toenemende mate expliciet (of zelfs speciaal) voor gebruikt (zoals webcams).

- Een vijfde nieuwe bron van data voor sociale wetenschappen – die tot nu toe nog nauwelijks wordt ontgonnen, is de directe neerslag van menselijk gedrag in virtuele omgevingen (zoals online games). In principe zouden deze uitstekend kunnen worden gebruikt om sociaal gedrag te bestuderen, en zelfs te toetsen in pure experimentele research designs. In afwezigheid van een laboratorium (een gecontroleerde omgeving) is het hoogst haalbare in sociologisch en bestuurskundig onderzoek nu een quasi-experimentele research design[57].

Juist vanwege de variëteit in het koppelen van al deze typen ongestructureerde data heeft Big Data waarschijnlijk een relatief grote impact op sociale en geesteswetenschappen. Big Data kan daar worden gebruikt om het gedrag van mensen in kaart te brengen (voltooiing van de 'probabilistische revolutie'). Nederland is, als drukbevolkt en dicht genetwerkt land, de ideale 'testing ground' voor eScience (cf. Medical trials). In technische wetenschappen geldt de wet van de grote aantallen al veel langer. Bovendien is de data daar meer gestructureerd, meer homogeen en minder verknoopt met andere bronnen (meer 'stand alone').

Sterke infrastructuur voor data-intensief onderzoek

Nederland heeft een goede uitgangspositie aan de technisch infrastructurele kant. Debet daaraan is de centrale positie van SURFsara, die van oudsher (SARA) internationaal al sterk is. Zo is er bijna tien jaar geleden al begonnen aan de uitrol van nationale dataservers (biG Grid).[58]

SURFsara ondersteunt wetenschappelijk onderzoek door het aanbieden van een state-of-the-art geïntegreerde ICT infrastructuur. Bovenop die infrastructuur worden ook diensten aangeboden ten behoeve van computing, data-opslag, clouddiensten en E-science. Hiermee biedt SURFsara een stevig fundament en de voorwaarden om data-intensief onderzoek te doen.

De hoogwaardige infrastructuur wordt alom geprezen en komt veelvuldig terug als een van de sterke enablers van het Nederlandse innovatieland. In de onlangs verschenen Wetenschapsvisie 2025[59] van het ministerie van OCW wordt nogmaals gerefereerd aan deze sterkte en wordt tegelijkertijd de ambitie uitgesproken om deze sterke positie te behouden door het vernieuwen van deze infrastructuur.

---

[56] Zie voetnoot 7.

[57] Cook, T.D. en D. T. Campbell (1979). Quasi-Experimentation: Design & Analysis Issues for Field Settings. Boston MA: Houghton Mifflin en Shadish, W.R., T.D. Cook, D. T. Campbell (2001). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston MA: Cengage Learning.

[58] Cook, G (2010) Building A national Knowledge Infrastructure. How Dutch Pragmatism nurtures a 21st Century Economy. Utrecht: SURF http://www.cookreport.com/pdfs/knowledge.pdf.

[59] Ministerie van OCW (2014), Wetenschapsvisie 2025. Keuzes voor de toekomst, Den Haag.

Kruisbestuiving tussen data-intensieve disciplines

Er vindt in Nederland op dit moment een interessante kruisbestuiving plaats tussen de technische disciplines die van oudsher al veel ervaring hadden in het omgaan met grote hoeveelheden data en geesteswetenschappen en sociale wetenschappen. Het Netherlands eScience Center (NLeSC) speelt hierin een instrumentele rol. Technische wetenschappers leren ook omgekeerd van sociale wetenschappers hoe ze om moeten gaan met 'messy' rijke, ongestructureerde data.

In technische wetenschappen ontstaan er ook steeds meer en grotere samenwerkingsverbanden om data uit te wisselen. Een extreem voorbeeld is het virtuele astronomische observatorium IVOA dat meer dan 3.000 grote databases van over de hele wereld met elkaar verbindt. De opkomst van Big Data leidt ook in geesteswetenschappen en sociale wetenschappen tot steeds meer samenwerkingsverbanden. Met name voor de geesteswetenschappen is dat een breuk met de traditie van geïsoleerd werk en monografieën.

Er lijkt zich een symbiose af te tekenen tussen het datagedreven potentieel (automatische patroonherkenning etc.) om sociale fenomenen te registreren en de traditionele sociale wetenschappen en geesteswetenschappen. Autonome computerprogramma's ('agents') van data science zorgen voor het brede overzicht en maken het vervolgens mogelijk voor sociologen en andere wetenschappers om interessante thema's te vinden en via grote hoeveelheden data daar dan dieper op in te gaan. Een indirect gevolg van het werken met grote hoeveelheden data is het ontstaan van meer samenwerking en betere afstemming en arbeidsverdeling tussen sociale wetenschappers.

De hierboven beschreven voordelen van kruisbestuiving leveren nieuwe inzichten op en geven een positieve impuls aan innovatief onderzoek.

Methodieken die tot aantoonbaar beter onderzoek leiden.

Een laatste kracht die we hier beschrijven is het vooruitzicht / de potentie om met behulp van data aantoonbaar beter (betrouwbaarder, meer valide) onderzoek uit te voeren.

In geesteswetenschappen is het bijvoorbeeld nu mogelijk om het gehele corpus exploratief te analyseren, zodat het 'rijke handwerk' (close reading; hermeneutiek) veel gerichter kan worden uitgevoerd. Voorheen kon men in de geesteswetenschappen wegkomen met unieke (niet representatieve) resultaten. Dat is nu niet meer het geval.

Het belang van het gebruik van 'real life' empirische data is de laatste jaren binnen tal van disciplines sterk toegenomen. Konden wetenschappers aanvankelijk nog publiceren op basis van louter theoretische modellen, en later met prototypes gevuld met dummy data, tegenwoordig zal men toch met een proof of concept moeten komen dat is uitgetest met 'echte' empirische data. Dit noodzaakt wetenschappers steeds meer om met Big Data te gaan werken.

Haaks op deze vraag staat een meer fundamentele discussie van breedte versus diepte die bij vele Big-Data-gerelateerde onderzoeken opkomt. Big Data heeft de potentie om beide dimensies te verrijken, maar die belofte van diepte (een adequate duiding) moet vaak nog worden ingelost. IBM's Watson kan de informatie die in circa 1 miljoen boeken staat in 1 seconde verwerken[60], maar mist diepte in de interpretatie van de 'gelezen' boeken (een

[60] Rennie, John (2011-02-14), How IBM's Watson Computer Excels at Jeopardy!, PLoS blogs http://blogs.plos.org/retort/2011/02/14/how-ibm%E2%80%99s-watson-computer-will-excel-at-jeopardy/.

sterkte van het hermeneutische vermogen van de mens). Het is een open vraag in hoeverre het laatste echt een uniek onderscheidend vermogen van de mens blijft of dat massa toch deels voor een gebrek aan diepte kan compenseren. In dat geval kan de human agent worden vervangen door een (software) agent. Als dat niet zo is, wordt het unieke hermeneutische vermogen van de human agent juist extra belangrijk.[61]

Feit is in elk geval dat de toolbox van datagerelateerde methoden en toegang tot data die voorheen niet voorhanden was in veel gevallen tot betere onderzoeksresultaten leiden en de wetenschap verder kunnen brengen.

4.2.2 Afremmende krachten voor gebruik (Big) Data in de wetenschap

Toch is het gebruik van (Big) Data in de wetenschap geen vanzelfsprekendheid. Hieronder passeert een aantal afremmende krachten de revue. We putten in deze beschouwing vooral uit interviews in het veld en deskstudie.

Afbreuk aan kern van wetenschappelijk onderzoek

De verhouding tussen data science en traditioneel onderzoek is de inzet van een methodenstrijd die zich op dit moment in veel disciplines afspeelt. Protagonisten van Big Data stellen dat onderzoek vanuit de data zou moeten beginnen: exploratief onderzoek op basis van data mining die vervolgens als input dient voor vraaggestuurd onderzoek. Andersom stellen tegenstanders dat het onderzoek altijd vanuit een a priori veronderstelling van de wetenschapper zou moeten uitgaan. *Data does NOT speak for itself* en resultaten van data mining zijn altijd gebiased (e.g, filter bubble[62]); de patronen die men vindt zijn niet anders dan de patronen die men er in stopt.

Ceteris paribus speelt hier de verhouding tussen de data scientist en de domeinexpert. Ofwel, in hoeverre is een informatiesysteem in staat om zelf, nadat de kennis uit de domeinexpert eenmalig is gecodificeerd en in algoritmes is vertaald, rijke en fuzzy datasets steeds beter te analyseren ('self learning systems')?

Een verlichte visie is dat data-science-methoden op zichzelf neutraal zijn en dat biases het gevolg zijn van onjuist gebruik van de methoden ('bad science'). Zulk niet-kritisch gebruik

---

[61] De angst dat de opkomst van nieuwe technologie leidt tot grootschalige werkloosheid ('technological unemployment') speelt al sinds het begin van de 19e eeuw (opstand van de Luddieten) en komt in golven telkens weer naar boven. Ook op dit moment wordt er weer veel aandacht besteed aan de mogelijke negatieve gevolgen van robotisering voor de werkgelegenheid. De angst dat deze nieuwe technologieën massale werkloosheid tot gevolg zouden hebben is echter (tot zover) onterecht gebleken, althans op de langere termijn. Door deze vinding is 'technologische werkloosheid' ook wel bestempeld met de 'Luddite fallacy': "If the Luddite fallacy were true we would all be out of work because productivity has been increasing for two centuries" (Tabarrok, A. (2003). Productivity and unemployment. *Marginal Revolution*.) In algemene zin lijkt het erop, ook gezien de historie van de mens, dat de werkzaamheden van de mens zich verschuiven van onderste lagen van de Maslow-piramide (primaire levensbehoeften, productie) naar hogere lagen van deze piramide. Een tweede hoofdargument is dat het niet technologie is die ongelijkheid introduceert, maar sociale en politieke keuzes. Als machines het werk van mensen kunnen overnemen, betekent dit dat de waarde (in termen van productie) gelijk is, en dat mensen meer vrije tijd hebben. In deze vrije tijd kunnen mensen ander werk verrichten, of simpelweg genieten van hun vrije tijd. Het probleem is dat de opbrengsten van automatisering dan wel op een breed gedragen acceptabele manier moeten worden verdeeld. Bron: Brennenraedts, R., A. Vankan, R.A. te Velde, B. Minne, J. Veldkamp, B. Kaashoek (mimeo). De impact van ICT op de Nederlandse economie. Dialogic: Utrecht. In opdracht van het Ministerie van Economische Zaken.

[62] Pariser, E. The Filter Bubble: What the Internet Is Hiding from You, New York: Penguin Press.

vindt in de huidige hype-fase echter zeker plaats. De kans op Type 2-fouten in wetenschappelijk onderzoek wordt vergroot.

Angst voor privacy-issues

Privacy wordt steeds meer een issue bij het gebruik van data voor onderzoeksdoeleinden. In bepaalde domeinen (zoals in de medische wetenschap) waren de privacybepalingen altijd al streng. Omdat data nu op steeds grotere schaal wordt gebruikt – en vooral wordt *her*gebruikt – gaan de belemmeringen echter steeds meer knellen. De belofte van Big Data zit juist in het gebruik van gegevens voor toepassingen waar de data oorspronkelijk nooit voor was bedoeld c.q. voor was verzameld. Strikte privacybepalingen zoals doelbinding (Wet bescherming persoonsgegevens, Artikel 9) betekenen in principe dat persoonsgegevens nooit mogen worden *her*gebruikt – ook niet voor bijvoorbeeld medisch onderzoek dat ten goede komt aan de samenleving als geheel. Hier botst het individuele belang (bescherming privacy) met het publieke belang. Vanwege het toenemende aantal privacy-incidenten in de private sector rond het gebruik van Big Data (met de ING als meest pregnante voorbeeld), zal de privacywetgeving eerder worden aangescherpt dan verruimd. Indirect wordt het gebruik van (Big) Data in de wetenschap hierdoor belemmerd.

De huidige wetgeving rond de bescherming van persoonsgegevens gaat uit van het principe dat microdata – data op het niveau van individuele personen (of bedrijven) - niet wordt vrijgegeven. Onder stringente voorwaarden mag microdata (zoals het Sociaal Statistisch Bestand van het CBS) soms wel door bepaalde derde partijen (zoals academische onderzoekers) worden geanalyseerd, maar dan alleen als de microdata vooraf worden geanonimiseerd.[63] Het grote nadeel daarvan is dat de kracht van Big-Data-analyses nu juist zit in het koppelen van data op microniveau. Daarnaast is anonimisering in multidimensionale datasets niet waterdicht: als er maar voldoende velden/kolommen beschikbaar zijn, zullen er altijd unieke combinaties ontstaan die alsnog verwijzen naar een uniek persoon of bedrijf.[64] Een praktische uitweg om de analyse nog steeds op microniveau uit te kunnen voeren is door de koppeling door een derde partij (Trusted Third Party of TTP) te laten maken. De TTP maakt dan intern de koppeling, maar zal naar buiten toe alleen geaggregeerde resultaten bekendmaken.[65] Dit is een uitweg voor wetenschappelijk onderzoek omdat het doel meestal immers niet is om individuen te traceren (zoals in commerciële Big-Data-analyses vaak wel het geval is – profiling) maar om patronen te vinden over de gehele dataset. Alleen om die patronen te kunnen vinden is vaak – *als tussenstap* – het koppelen op microniveau een vereiste.

Protectionistisch gedrag

Bedrijven beseffen steeds meer de strategische waarde van data en schermen hun data in toenemende mate af. In termen van privacy staan het bedrijfsbelang (geheimhouding) en

---

[63] Dat wordt meestal gedaan door cellen te aggregreren totdat ze een minimale grootte hebben (zeg n>5) en/of door de waarden in bepaalde velden te vervangen door één en dezelfde anonieme waarde (in het veld GESLACHT worden de waarden M of V vervangen door *).

[64] Charu C. Aggarwal (2005). On k-Anonymity and the Curse of Dimensionality. http://www.vldb2005.org/program/paper/fri/p901-aggarwal.pdf
Zie ook http://datascience.berkeley.edu/anonymous-data/.

[65] Zie voor een toepassing in het sociale domein bijvoorbeeld Dialogic (2014), Handreiking Zelfredzaamheid en ICT: inventarisatie Digitale Informele Zorgdiensten en de potentie van Big Data en Open data. In opdracht van Kwaliteitsinstituut Nederlandse Gemeenten (KING). Utrecht, augustus 2014. https://www.visd.nl/sites/visd/files/Handreiking-Zelfredzaamheid-en-ICT-Dialogic-augustus-2014.pdf.

het wetenschappelijke belang (openbaarheid, reproduceerbaarheid) haaks op elkaar. Wetenschappers krijgen vaak alleen maar onder zeer strenge restricties toegang tot data van bedrijven. In de meeste gevallen komen de data nooit buiten het bedrijf (er moet dus altijd on site op de data worden gewerkt).

Op dit moment is in de academische gemeenschap een strijd gaande tussen puristen en pragmatisten wat betreft de vereiste openheid van de data waar in wetenschappelijke publicaties naar wordt verwezen. De eerste groep stelt dat alle data in wetenschappelijke publicaties zonder restricties toegankelijk zou moeten zijn. Bij het gebruik van data uit de private sector (maar zelfs uit de publieke sector) is dat vaak niet mogelijk. De tweede groep stelt dat het beter is om uit te gaan van de maximaal haalbare openheid dan om helemaal geen data te hebben.[66]

Het onder gecontroleerde omstandigheden toegang geven tot vertrouwelijke gegevens kan een geheel andere route opleveren om met privacykwesties om te gaan. Dat kan namelijk als achteraf altijd precies bekend is wie wanneer en op welke wijze toegang heeft gehad tot bepaalde (micro)data. Het nauwkeurig registreren van alle bewerkingsslagen op data zodat de herkomst – *provenance* – van de eindresultaten altijd kan worden gecontroleerd is een belangrijke opkomende trend in de wetenschap. Deze trend wordt gevoed door twee parallelle ontwikkelingen: het alsmaar toenemen van de hoeveelheid data waardoor wetenschappers het overzicht dreigen te verliezen over hun eigen data (en/of van de andere partijen waarmee ze samenwerken) en de toename van het aantal gevallen van wetenschapsfraude. [67] Persoonsgegevens en andere gevoelige data kunnen worden beschermd door provenance omdat altijd precies kan worden achterhaald wie toegang heeft gehad tot welke data. Dan heeft de toegang uiteraard al plaatsgevonden (de controle vindt *achteraf* plaats) maar het idee is dat potentiële gebruikers van de data *vooraf* de afweging wel zullen maken om geen oneigenlijk gebruik te maken omdat ze daardoor altijd ter verantwoording kunnen en zullen worden geroepen.

4.2.3 Interactie tussen stimulerende en afremmende krachten datagebruik

De hierboven beschreven krachten werken op elkaar in (zie Figuur 8). De toegang tot nieuwe data leidt bijvoorbeeld tot verdieping in methodieken om het meeste uit deze data te halen om tot aansprekende onderzoeksresultaten te komen. De resulterende toename van gebruik van Big Data in wetenschap kan er juist toe leiden dat de afremmende krachten worden aangezwengeld.

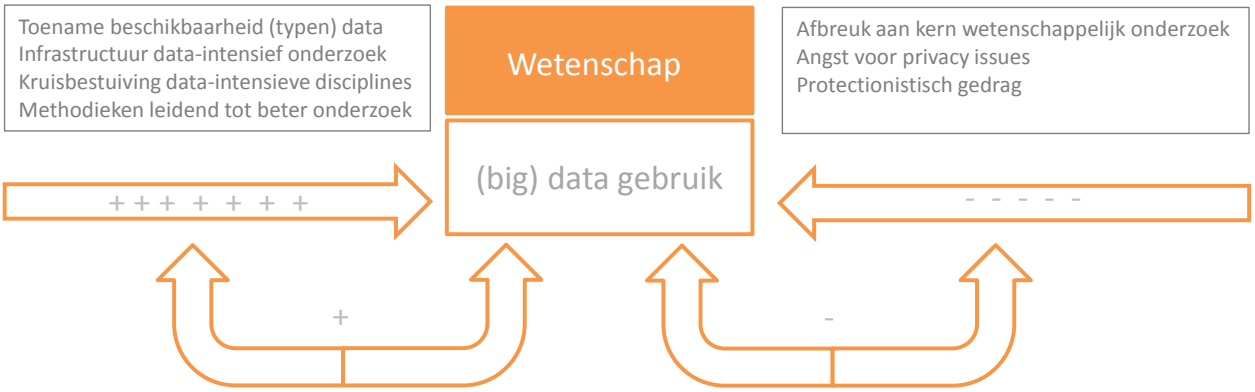
Zo is het goed denkbaar dat een toenemend gebruik de 'sense of urgency' van privacyissues die met de toegang tot deze data gepaard gaat versterkt. In het algemeen zullen de 'eigenaren'/'verzamelaars' van de data ook in toenemende mate het strategisch belang

---

[66] In de praktijk worden er door bedrijven vele soorten eisen aan de toegang tot hun data gesteld. Meestal moet er on site worden gewerkt en/of wordt er alleen na een bepaalde periode toegang verleend. In sommige gevallen is toegang alleen mogelijk voor medewerkers en moet de wetenschapper (tijdelijk) in dienst treden van het bedrijf (dit geldt bijvoorbeeld voor promovendi die willen werken met data van de grote Amerikaanse social mediasites en zoekmachines). Een uitweg is soms om alleen met verouderde data te werken. Die is vanuit commercieel oogpunt vaak niet of nauwelijks meer iets waard maar is voor wetenschappelijk onderzoek meestal nog steeds heel bruikbaar.

[67] Het blijft een open vraag of het *feitelijke* aantal gevallen van wetenschapsfraude is toegenomen, of alleen het *waargenomen* aantal. In het laatste geval is de toename grotendeels toe te schrijven aan het feit dat we nu beter in staat zijn om fraude op te sporen dan voorheen – bijvoorbeeld door de toepassing van softwaretools zoals eTBLAST.

inzien van deze gegevens, hetgeen kan leiden tot protectionistisch gedrag of het onnodig uitbuiten van wetenschappers die deze gegevens voor onderzoeksdoeleinden willen inzetten.



Figuur 8: Interactie tussen krachtenvelden ten aanzien van gebruik (Big) Data in de wetenschap.

In de verdere ontwikkeling van de rol van Big Data in de wetenschap staan drie opeenvolgende processen centraal:

1. Het verzamelen en verkrijgen van de data.
2. Het interpreteren en analyseren van de data.
3. Het toepassen van de resultaten van de analyses.

In alle drie processen gaat het om een vraagstuk van machtsverdeling tussen wetenschap, bedrijfsleven, burgers en overheid.

We beschrijven deze ontwikkelingen eerst aan de hand van een ongunstig scenario (*dystopie*) en daarna van een gunstig scenario (*utopie*) in de hierop volgende paragrafen.

4.3 Dystopie Wetenschap 2025

In deze paragraaf kijken we vanuit het jaar 2025 terug op de impact van Big Data op de wetenschap. De veronderstelling in dit essay is dat Big Data in de wetenschap niet gebracht heeft van wat wij ervan verwachtten. Het gaat dus om een dystopie.

4.3.1 De strijd om de data

Big Data leek aanvankelijk de nieuwe olie en dat was het ook. Overal in de wereld stimuleerden overheden het ontstaan van 'de Big Data economy'. Niemand wilde de boot missen. Bang om de hausse in de kiem te smoren, lieten overheden het belang van bedrijven prevaleren boven het belang van haar burgers. Data was immers de core asset van bedrijven. Dat gold vooral voor de 'Big Data born' bedrijven die hun hele business hadden gebouwd op en rondom bepaalde datasets. Het intellectueel eigendomsregime werd sluipenderwijs aangescherpt in plaats van versoepeld – steeds meer data werd proprietary en de geheimhouding van bedrijfsgegevens ging altijd voor het belang van publieke openbaarheid. In een tijdsbestek van een paar jaar kwam bijna alle data in private handen.

Alleen bleek in de ontnuchterende nadagen van de hype dat het eigenlijk heel onduidelijk was hoe je geld kon verdienen met Big Data. Alle aandacht was uitgegaan naar de handvol zeer succesvolle Amerikaanse Big-Data-multinationals maar nu bleek dat er buiten deze groep om nauwelijks winst werd gemaakt. Deze bedrijven hadden alle belangrijke centrale datasets (de 'sleutelsets') in handen. Er was nog steeds data in overvloed maar zonder deze sleutelsets was het vrijwel onmogelijk om data te koppelen aan individuele burgers. Deze

bedrijven hadden de beste intelligence. Alleen zij wisten near-real time van bijna alle mensen in de wereld waar zij waren, wat zij deden of ooit hadden gedaan, hoe ze hun geld verdienden, hoe gezond ze waren. En wat ze dachten.

De concentratie deed zich ook in geografische termen voor. Pas toen in 2018 al 90% van alle data in de Amerikaanse cloud of in de databases van Amerikaanse bedrijven stond werd de Europese Commissie wakker. Er ontstond een openlijke handelsoorlog tussen Europa en de VS, maar Europa trok uiteindelijk aan het kortste eind. Binnen Europa werd de positie van Nederland steeds marginaler. Dat lag niet aan de Nederlandse wetenschap – dat had aanvankelijk een uitstekende uitgangspositie – maar aan het Nederlandse bedrijfsleven, dat vanaf het begin internationaal nauwelijks een rol van betekenis had gespeeld. En de wetenschap was voor het uitvoeren van onderzoek in toenemende mate afhankelijk van proprietary data uit de private sector.

Data was een strategisch goed en de dataprijs bleef maar stijgen. Ook werden de kosten van de opslag van data steeds meer een bottleneck. Lange tijd leek, vanwege de exponentiële technologische vooruitgang, opslag nauwelijks iets te kosten. Maar toen die vooruitgang begon te vertragen bleek alras dat ook opslag geen 'free lunch' was. De kosten van opslag in de cloud stegen opeens exorbitant. Overal ter wereld moesten faculteiten en overheden keuzes maken. Big Data werd een Grote Kostenpost en veel universiteiten en faculteiten konden niet langer de toegang tot relevante datasets betalen of nog langer al hun datasets beheren of alle data opslaan.

De voortschrijdende vermarkting van data had een relatief sterke negatieve impact op de Nederlandse wetenschap. De meeste Nederlandse data scientists kwamen voort uit disciplines zoals natuurkunde en astronomie, die een cultuur kenden van delen en openheid – ze waren niet gewend om te opereren in commerciële omgevingen en verloren in hun naïviteit en goedbedoelde idealisme al snel de slag met het bedrijfsleven. De kerngroep van het Nederlandse Big-Data-cluster was niet bestand tegen de alsmaar toenemende commercialisatie. De beste jonge wetenschappers werden direct opgekocht of gingen in arren moede in dienst van Big-Data-bedrijven omdat dat de enige manier was om nog toegang te krijgen tot de data die ze nodig hadden.

Ondertussen werd de kleine groep van Big-Data-multinationals steeds sterker. Dat kwam onder andere omdat de toegevoegde waarde van Big Data – en zeker voor toepassingen in de markt en in het sociale domein – vooral bleek te liggen in het koppelen van bestaande datasets in plaats van het verzamelen van nieuwe datasets. De grote bedrijven sloten steeds meer exclusieve uitruildeals (net als patenten) maar ze deden dat alleen onderling, binnen hun eigen groep, in besloten *data federations*. Dit was een open-data-concept dat ze van de technische wetenschappen hadden overgenomen – maar dat in feite een 'datakartel' was. Zo stapelde de macht zich op: de partijen die al toegang tot de strategische datasets hadden kregen, op basis van de data die ze al hadden, toegang tot nog meer datasets. *The rich are getting richer and the poor are getting poorer*.

De overheid en de wetenschap werden steeds meer afhankelijk van private data. De overheid kon niet langer toegang tot private data afdwingen – de Wet op de CBS was al jaren geleden opengebroken en daarna afgeschaft. Er waren nooit checks & balances ingebouwd in het dataverzamelingsproces in de private sector. Bedrijven brachten soms voor de vorm of voor marketingtechnische redenen nog wel data naar buiten maar ze gaven geen toelichting bij de wijze waarop de data was verzameld en was opgeschoond en verwerkt. Dit was een vorm van pseudo-openbaarheid: zonder deze toelichtingen konden buitenstaanders zoals wetenschappers, burgers en ambtenaren niet zoveel met de data.

De wetenschap, die van oudsher de contra-expertise gaf voor de overheid en burgers vis-à-vis het bedrijfsleven, raakte (of werd) steeds meer afgesneden van toegang tot data. Hierdoor werd de positie van de wetenschap zo ernstig verzwakt dat de wetenschap haar countervailing power voor de overheid en de burgers niet meer kon waarmaken. Ze had geen toegang tot data meer, geen geld voor dure IT-infrastructuur en daardoor werd haar autoriteit ondermijnd. Alle specialisten uit de wetenschap verdwenen naar het bedrijfsleven. Even heeft de wetenschap nog geprobeerd via het alternatieve spoor van *citizen science* tegenwicht te bieden. Het mobiliseren van burgers bleek echter ook geen oplossing meer te bieden omdat de burgers allang de controle over hun eigen data waren kwijtgeraakt. Ze kregen geen toegang tot de data die ze, ironisch genoeg, zelf genereerden. Die ontwikkeling was sluipenderwijs gegaan. Ze *hoefden* hun data niet ter beschikking te stellen aan bedrijven, maar dan werden ze uitgesloten van de diensten die deze bedrijven – op basis van data van burgers – aanboden. Met de groeiende invloed van de Big-Data-multinationals konden burgers niet meer om deze diensten heen op straffe van sociale uitsluiting. Welbeschouwd hadden ze weinig te kiezen.

Ondertussen verslechterden de sociale verhoudingen binnen de gemarginaliseerde groep van academische dataspecialisten zienderogen. Terwijl het belang van empirische data in publicaties steeds verder toenam, was er een groeiende schaarste aan data en middelen. Dit zorgde ervoor dat onderzoeksgroepen steeds bezitteriger werden over hun datasets. Internationale academische datafederaties vielen uit elkaar. Ook binnen de groepen zaten individuele wetenschappers steeds krampachtiger op hun data, die ze meestal via eigen lijntjes met een bedrijf hadden verkregen. Het was nu ieder voor zich. *Homo homini lupus*.

4.3.2 De nieuwe Orde uit chaos

Aan het begin van de jaren ’10, op het hoogtepunt van de Big-Data-hype, was er een grenzeloos geloof in de mogelijkheden van wetenschappelijk onderzoek en gebruik van (Big) Data. Met name ten aanzien van het doen van voorspellingen (*predictive statistics*) waren de verwachtingen hooggespannen. In de apotheose van de probabilistische revolutie was het mogelijk om perfecte voorspellingen te doen – er kon eindelijk definitief orde in de chaos worden geschapen. In naam beleden de meeste wetenschappers wel dat “data does *not* speak for itself” maar in de praktijk voeren ze blind op de resultaten van data mining-exercities en lieten ze de richting van hun onderzoek daardoor bepalen. Debet daaraan was ook het feit dat zelfs basale Big-Data-vaardigheden nooit integraal onderdeel werden van disciplines buiten het kernvakgebied. Die expertise bleef geconcentreerd in een relatief kleine gemeenschap van data scientists.

Het ongebreidelde geloof in Big Data leidde tot een vorm van pseudowetenschap die langzamerhand de standaard werd; in publicaties kwam de nadruk steeds meer te liggen op data en steeds minder op duiding. Deze trend werd verder versterkt vanuit de samenleving, waar burgers en ambtenaren vol ontzag waren over de duizelingwekkend grote datasets en betoverd werden door een niet aflatende stroom van gelikte infographics. Net als in de begindagen van de reclame slikten ze de boodschappen als zoete koek.

De ruimte voor menselijke waarneming en interpretatie werd steeds kleiner. Domeinexperts waren eerst nog nodig om agents te ontwikkelen waarmee datasets konden worden geclassificeerd. Was hun kennis echter eenmaal geëxtraheerd en gecodificeerd in algoritmes en heuristieken dan waren de menselijke experts (de human agents) niet meer nodig. De zelflerende systemen verbeterden zichzelf daarna met de data die ze binnenkregen op basis van de directe neerslag van de menselijke gebruikers van de informatiesystemen. Zo pasten dit soort systemen (zoals zoekmachines, game engines, decision support systemen) hun aanbod automatisch op maat aan aan individuele gebruikers. Er was niet of nauwelijks ruimte

in dit soort recursieve processen voor spontaniteit, grilligheid en creativiteit. Er ontstonden conceptuele trechters (*filter bubbles*) waarin gebruikers gevangen raakten zonder dat ze het zelf beseften. In wetenschappelijke analyses werden ongebruikelijke waarnemingen automatisch verwijderd – terwijl de wetenschapsgeschiedenis heeft laten zien dat juist deze extreme waarden vaak aan de grondslag lagen van radicale doorbraken en wetenschappelijke revoluties.

Tegelijkertijd ging de oude ambacht van *close reading* verloren. In de samenleving voer men steeds meer blind op grote sets data uit directe neerslag. Het unieke hermeneutische vermogen van de mens werd niet meer erkend. Diepte en rijkheid werden gesubstitueerd door breedte en massa.

Er was maar een kleine groep mensen die voldoende kennis had van Big-Data-technieken om dit soort structurele biases te doorgronden. Er kwam in eerste instantie nog wel tegenwicht uit deze groep, maar de mainstream liet steeds minder ruimte voor reflectie. Al te veel kritiek op de beloften van Big Data werd niet op prijs gesteld. Daarvoor was het geloof bij de meerderheid op dat moment nog te sterk en werd het commerciële belang alleen maar voortdurend groter.

In het algemeen zette de trend in de wetenschap zich voort dat er steeds minder ruimte was voor vrij onderzoek, waarin bijvoorbeeld maatschappijkritische vraagstellingen konden worden onderzocht. De mogelijkheid om vrij onderzoek te doen was nu juist decennialang het verkoopargument voor universiteiten aan bedrijven. Vanaf 2010 draaide dit om. Juist grote bedrijven gaven jonge veelbelovende onderzoekers de ruimte voor eigen onderzoek. Alle creatieve talenten weken uit naar de Big-Data-multinationals.

Ironisch genoeg heeft het vrije onderzoek in Nederland nog een tijdje kunnen blijven drijven op de kurk van Aziatische promovendi. De overheden uit hun land van herkomst betaalden carte blanche hun onderzoek – zolang ze maar in het buitenland bij prestigieuze vakgroepen gingen promoveren. Zo konden Nederlandse hoogleraren deze promovendi – die zelf nu niet bepaald vrij denken – inzetten voor hun eigen onderzoeksagenda op het terrein van Big Data. Zo promoveerden er in 2017 en 2018 twee Chinese studenten op de inherente biases van populaire recommender systems. Het merendeel van deze studenten vertrok na hun promotie weer naar het buitenland. Tegelijkertijd werden de beste datawetenschappers sowieso door het bedrijfsleven opgekocht. Na 2016 waren er geen postdocs meer over. In het tweede en derde geldstroomonderzoek was steeds minder ruimte voor reflexief onderzoek op het terrein van Big Data.

De groep van verlichte academische data scientists kromp en kromp. Dit leidde zowel in termen van kwaliteit als kwantiteit tot een achteruitgang in het Big-Data-onderzoek in Nederland. Door het gebrek aan kritische massa en het verlies aan internationaal prestige bleven de Aziatische promovendi nu ook weg. Er was weer minder ruimte voor vrij, kritisch onderzoek. De beste hoogleraren vertrokken naar het buitenland. De tweederangs wetenschappers die overbleven konden niet al te veel uitrichten. Ze kregen geen toegang tot de dichtgetimmerde datasets van de grote bedrijven of ze kregen alleen onder zeer stringente randvoorwaarden toegang tot de data. Ze hadden noch de statuur en invloed om veel tegenwicht te bieden noch de vaardigheden om de analyses van de bedrijven te reverse engineeren.

Ondertussen raakten steeds meer wetenschappers uit andere disciplines ondergesneeuwd in de grote hoeveelheden data waarvoor ze feitelijk niet de skills hadden om er mee om te gaan (*stack overflow*). Anderen kwamen eindelijk tot het inzicht dat meer data niet zonder meer leidde tot meer inzichten. Er trad een algeheel gevoel van teleurstelling op – de Big-Data-trend gaat door de beruchte *through of disillusionment* (zie Figuur 4). Debet daaraan was de

aanvankelijke sterke focus op voorspellende statistieken. Het was echter veel moeilijker om het ziektebeeld van één patiënt te voorspellen dan processen te beschrijven binnen een heel ziekenhuis. De Big-Data-tools en -technieken waren jarenlang overstretched.

De kundige en kritische wetenschappers wisten dit allang. Alleen waren er nog maar bitter weinig van dit soort kritische experts over. Omdat er nu alleen nog maar prescriptieve analyses werden uitgevoerd in de private sector verdampte de kennisbasis in het academisch onderzoek langzamerhand. Er waren nauwelijks nog academische wetenschappers die zelf dit soort analyses hadden uitgevoerd of nog konden uitvoeren. De wetenschap had niet de ervaring of kennis meer om de Big-Data-exercities van grote bedrijven kritisch te evalueren terwijl de maatschappelijke impact van deze voorspellingen alleen maar verder toenam (zie hierna, paragraaf 4.3.3).

4.3.3 De data dictatuur

Alle processen genereerden data. Of het nu fysieke of sociale processen waren. En zelfs mentale processen konden in de loop van de periode 2015-2025 steeds beter worden uitgelezen. Wie toegang had tot deze data, en in staat was om deze data te analyseren en de onderliggende mechanismes te doorgronden, kon de oorspronkelijke processen bijsturen en optimaliseren. Wie dat beter kon dan een ander, had dus ook meer macht over de fysieke, sociale en mentale processen dan de ander.

Dat was geen nieuwe ontwikkeling – zij is begonnen met het beheersen van natuurkundige processen in de 18de eeuw, met de opkomst van de ingenieur als professie. Het beheersen van *sociale* processen komt in het midden van de 19e eeuw in zwang – de *moral statistics* en *physique social* (*social physics*) van Quetelet – maar ze zetten nooit door. Men had in die tijd simpelweg niet genoeg data noch voldoende analytisch vermogen om de complexe sociale processen te doorgronden. Dat nam niet weg dat het denken in termen van risico's en waarschijnlijkheden – van regelmatigheden die op macroniveau ontstaan op basis van talrijke toevalsprocessen op microniveau – sluipenderwijs in belang toenam. Het gebruik van statistiek en kansmodellen produceerde onderscheidingen die vervolgens 'sociale realiteit' worden.[68] Op basis van statistische analyses werden er specifieke groepen of categorieën van individuen (of bedrijven) geïdentificeerd en vervolgens werden op deze classificaties strategieën van bedrijven en beleid van overheden gebaseerd. Dat gebeurde zonder dat deze groepen politiek waren vertegenwoordigd. Welbeschouwd waren ze overgeleverd aan de willekeur van bedrijven en overheden die hen als (risico)groep hadden gedefinieerd.

Al deze processen speelden al tientallen jaren maar in de periode 2015-2025 kwamen de ontwikkelingen, door de exponentiële toename van de hoeveelheid data en het vermogen om die data te analyseren en toe te passen in allerhande sociale processen, in een stroomversnelling. Voor het eerst in de menselijke geschiedenis was het nu mogelijk om near-real time collectief te leren van individuele gevallen.

De maatschappelijke gevolgen van het probabilistische, datagedreven denken kwamen nu pas aan het (volle) licht – en die gevolgen waren niet altijd even gunstig. Althans, voor de overgrote meerderheid van de burgers die – *for the greater good* – verplicht waren hun data voor analyse af te staan. In de hoogtijdagen van de Big-Data-hype – rond 2015 – waren veel mensen nog erg makkelijk in het verstrekken van hun persoonsgegevens. Het gebruik van tracking en tracing devices (zoals health apps), dat sterk werd gestimuleerd door de Big-Data-multinationals, werd in deze fase – met name door jongeren – nog enthousiast omarmd (*data exhibitionisme*). In de jaren daarna werden mensen voorzichtiger – ze gaven

[68] Gerard de Vries (1999). Op.cit.

in ieder geval niet meer zonder meer hun data gratis weg. De heersende gedachte was nu dat de markt de privacy-issues wel zou oplossen. Mensen die veel waarde hechtten aan privacy – en dus bereid waren om daar geld voor te betalen – konden die privacy immers 'kopen'. In de praktijk bleek dat laatste alleen weggelegd voor de elite van de allerrijksten. Alleen zij konden de transparantieverplichtingen omzeilen door de beste juridische en IT-experts in te huren en de randen van de wet op te zoeken. Die wisten haarfijn het verschil tussen transparantie-ontwijking (legaal) en transparantie-ontduiking (strafbaar) te duiden.

De leidende bedrijven – met de zieltogende overheid in haar kielzog – waren steeds beter in staat om het kansprofiel van mensen te berekenen. Er kwam steeds minder ruimte voor gedrag dat sterk afweek van het gemiddelde: 'outliers' werden 'outcasts'.[69] Als er een verhoogde kans bestond dat het vertonen van bepaald gedrag (of juist het nalaten ervan, zoals achterstallig onderhoud op het eigen lichaam) maatschappelijke kosten met zich mee zou brengen, dan zou de groep of het individu in kwestie daarvoor aansprakelijk worden gesteld. Dit was niet meer dan logisch in de publieke opinie, en de bedrijven uit het Big-Data-kartel konden dit ook met kwantitatieve analyses onderbouwen. Wie een ongezonde levensstijl had, betaalde meer verzekeringspremie. Andersom gold ook, dat wie voldoende gezond leefde, minder premie betaalde.[70] Ceteris paribus voor studenten die een studie kozen waarvan bekend was dat hun slaagkans laag was.

In 2019 werd nog bij wet vastgelegd dat eindbeslissingen over processen die natuurlijke personen betroffen, ook altijd door natuurlijke personen moesten worden genomen. Conclusies die door (zelflerende) kennissystemen werden genomen, hadden daarmee hoogstens de status van een aanbeveling. In de praktijk bleek dit al snel een wassen neus en was de discretionaire bevoegdheid van de natuurlijke personen waaraan beslisbevoegdheid was toegewezen (zoals artsen, schoolhoofden, decanen, belastinginspecteurs, uitkeringsambtenaren enzovoort) zeer beperkt. Aan de ene kant zagen zij zich geplaatst voor exacte quota die op basis van complexe kansmodellen van bovenhand werden opgelegd ("dit jaar komt maximaal 16% van de populatie natuurlijke personen waarover u beslist in aanmerking voor Y"). Aan de andere kant werden ze persoonlijk aansprakelijk gesteld voor de gevallen waarbij ze de aanbevelingen van de kennissystemen naast zich neerlegden – en het kennissysteem het toch bij het rechte eind bleek te hebben. Dat laatste gold onverkort voor de natuurlijke persoon waar de uitspraak betrekking op had. Als die het besluit van de beslisbevoegde niet volgde kon hij of zij daar persoonlijk aansprakelijk voor worden gesteld. Omdat de kennissystemen steeds beter werden (de term *near unfailing* kwam rond 2020 in zwang) en de samenleving tegelijkertijd meer risicomijdend werd (omdat er immers steeds beter rekening kon en werd gehouden met risico's) nam de beslisbevoegde natuurlijke persoon in 2022 in 97,3% van de gevallen de aanbevelingen van het expertsysteem over. In 2025 was dit percentage verder gestegen tot 99,4%.

De sociale wetenschappen hadden in het primaat van kansmodellen een dubieuze rol gespeeld. Door de opkomst van Big Data beleefde de *social physics* een revival – de sociale wetenschappen beschikten nu eindelijk over dezelfde 'harde' kwantitatieve instrumenten en het voorspellende vermogen als de natuurwetenschappen waar ze al die tijd tegenop hadden

---

[69] En daarmee was de cirkel rond want Quetelet had in 1835 niet alleen het concept van de 'gemiddelde mens' (*l'homme moyen*) geïntroduceerd maar ook van een morele meerwaarde voorzien: de homme moyen representeerde volgens hem het ideaalbeeld vanuit de natuur. Bron: Porter, T. (1986). *The Rise of Statistical Thinking, 1820-1900*. Princeton: Princeton University Press.

[70] Bijvoorbeeld als je fitness tracker minder dan 10.000 stappen per dag heeft geregistreerd gedurende twee aaneengesloten weken, of als je cortisolplasma-waarden meer dan een maand achtereen bovengemiddeld hoog of laag zijn.

gekeken.[71] *C.P. Snow was right after all*. De datagedreven sociale wetenschappen kregen alras een steeds groter economisch belang vanwege hun vermogen om patronen in sociaal gedrag te detecteren en menselijk gedrag te voorspellen. Sociale wetenschappers omarmden massaal hun nieuwe rol als *social engineers* en lieten zich – achteraf misschien enigszins naïef – voor het karretje spannen van de Big-Data-multinationals.[72] Veel viel ze niet te verwijten – bij die bedrijven zat ook het grote geld.

De positie van de geesteswetenschappen was veel minder rooskleurig. Ze wees tevergeefs op de teloorgang van het interpretatieve vermogen van de intelligentsia. Het doemdenken van de geesteswetenschappen, waarin wordt gewezen op de gevaren van het automatiseren en van het rücksichtlos uitselecteren van creatieve en afwijkende personen werd door de mainstream weggehoond. Na 2020 waren de geesteswetenschappen aan een achterhoedegevecht bezig – de Methodenstreit leek eindelijk gestreden. Op het gevaar af om zelf als *anomalisten* te worden beschouwd, stonden ze onder toenemende druk om juist de maatschappelijke kosten van afwijkend gedrag in de samenleving te laten zien. Zoals Quetelet al meer dan 150 jaar eerder had betoogd, bereikt de natuur in de Gemiddelde Mens zijn perfectie in mentale en morele termen. Dat gemiddelde was de maat waarop sociaal beleid zou moeten worden gebaseerd.

4.4 Utopie Wetenschap 2025

In deze paragraaf kijken we vanuit het jaar 2025 terug op de impact van Big Data op de wetenschap. De veronderstelling in dit essay is dat Big Data in de wetenschap gebracht heeft van wat wij ervan verwachtten. Het gaat dus om een utopie.

4.4.1 Data manna

Rond 2015 zette de opgang van Big Data definitief door. Allerwege kwam het inzicht dat data de zuurstof was van de moderne samenleving. Dat besef werd breed gedeeld, van markt tot civil society en staat. Big Data was niet alleen van belang voor een selecte groep van gebruikers – de impact van Big Data reikte veel verder en strekte zich tot alle domeinen van de samenleving uit. *Alle* processen genereerden immers data, en wie toegang had tot deze data, en in staat was om deze data te analyseren en de onderliggende mechanismes te doorgronden, kon de oorspronkelijke processen bijsturen en optimaliseren. Ook in de politiek en het maatschappelijk middenveld werden de mogelijkheden van Big Data duidelijk herkend. Voor het eerst in de menselijke geschiedenis was het mogelijk om near-real time collectief te leren van individuele gevallen. Dat vermogen was niet alleen voorbehouden aan het bedrijfsleven – als data de zuurstof was van de moderne samenleving dan mag niemand daarvan worden afgesloten. *Data werd een publiek goed*.

---

[71] Alex Pentland, een onderzoeker bij het Media Lab op het prestigieuze MIT, schrijft in 2014 een zeer invloedrijk boek met de veelzeggende titel: 'Social Physics: How Good Ideas Spread – The Lessons from a New Science'. Een jaar later komt er al een tweede boek: 'Social Physics: How Social Networks Can Make Us Smarter'. Er komt een aparte onderzoeksgroep binnen het MIT Media Lab onder de naam 'Social Physics'.

[72] Zo kunnen onderzoekers van het Psychometric Centre van de universiteit van Cambridge al in 2012 op basis van een analyse van tienduizenden Facebook likes met 95% zekerheid het ras van de (gepseudonimiseerde) Facebook gebruiker is, met 80% zekerheid wat de religie is, en met >80% wat de seksuele geaardheid is (Kosinkski, M., D. Stillwell, T. Graepel (2013) Private traits and attributes are predictable from digital records of human behavior, *PNAS 110* (15), pp. 5802-5805. Het onderzoek is gesponsord door Microsoft.

In die tijd was er een handjevol voornamelijk Amerikaanse bedrijven die veel geld verdienden met het exploiteren van hun proprietary datasets. Terwijl in de VS de nadruk lag op de ontwikkeling van dit verdienmodel vond in Europa een omkering van het perspectief plaats. Men kwam daar tot het inzicht dat de (indirecte) maatschappelijke en economische kosten van het niet vrijgeven van data vele malen hoger waren dan de winsten die met het exploiteren van rechten op datasets waren gemoeid.[73] Bovendien werden de laatste inkomsten ook nog eens scheef verdeeld binnen de samenleving. Met de Free Data Directive uit 2018 verbood de Europese Commissie het exclusieve hergebruik van data door bedrijven. Dit was een radicale breuk met het traditionele *proprietary* IPR-regime. Dat betekende onder andere dat bedrijven alleen nog data mochten verzamelen als ze die data ook publiek toegankelijk maakten.

Vrije toegang tot data werd breed gedragen door de samenleving. Die gedijde bij een sterke rol van de overheid, die de correcte omgang met data garandeerde. In 2017 werd vanuit een burgerinitiatief het 'datacodicil' gelanceerd. Daarin verklaarden burgers dat ze hun persoonlijke data ter beschikking stelden aan de wetenschap zolang deze de gegevens gebruikte voor maatschappelijke doeleinden. Initiatiefnemer was een genezen kankerpatiënt die haar genetische data ter beschikking stelde voor medisch onderzoek. Het initiatief vond aanvankelijk aarzelend opgang maar nadat de Nederlandse overheid hard ingreep na het privacy-debacle rond de Amerikaanse pharmareus Pferck en de verantwoordelijke individuen binnen het bedrijf strafrechtelijk vervolgde, nam het geloof in het op *provenance*-gebaseerde privacyregime hand over hand toe. Alle burgers kregen hun eigen persoonlijke data-dashboard waarin ze precies konden zien welke data er over hun verzameld was, en wie toegang had tot die data. Ze kregen automatisch berichten als er iets op het dashboard – en dus in de status van hun data – veranderde. Zo hielden ze de controle over hun eigen data.

In 2020 nam de Europese Commissie het data-codicil-principe als opt-out-bepaling in de Free Data Directive op. Dat wil zeggen dat alle burgers binnen de EU hun data ter beschikking stelden aan de wetenschap tenzij ze zich daar expliciet tegen verzetten. De implementatie van de regeling werd geborgd door het Europese Hof van Data Protectie dat streng optrad tegen gevallen van oneigenlijk gebruik van data. Onder het motto *given enough eyeballs, all thugs are traceable* maakte het Hof in de controle op de naleving van de out-out-bepaling gebruik van een Europees-breed meldingensysteem waaraan miljoenen Europese burgers deelnamen.

De Free Data Directive was de aanleiding tot een handelsoorlog met de VS (waar bedrijven het alleenrecht over hun data claimden) en met China (waar de overheid dat deed). Uiteindelijk gaven de consumenten de doorslag. In Europa lieten ze massaal de bedrijven vallen die de free-data-principes met de voeten traden. Voor bedrijven die in Europa wilden blijven opereren werd het correcte (her)gebruik van data al snel een kwestie van overleven. Na een korte periode van terugval leidde dit uiteindelijk tot een enorme boost voor het Europese bedrijfsleven. De open-data-beweging verloor haar wollige en vrijblijvende imago – open data werd big business. Na 2020 stapten ook steeds meer Amerikanen over naar Europese bedrijven omdat daar de transparantie van hun data wél was gegarandeerd. Knarsetandend nam het door Republikeinen gedomineerde Congres in 2023 een nieuwe wetgeving aan waarin grosso modo de free-data-principes werden erkend. Alleen China bleef een besloten markt, tot grote frustratie van de leidende Chinese producenten van smartphones zoals Huawei en Xiaomi, die hun verkoop van telefoons in de VS en met name

[73] Zie voor een vergelijkbare redenering: R.A. te Velde (2009). Public Sector Information: Why Bother? in: P. Uhlir (ed.) The Socio-economic Effects of Public Sector Information on Digital Networks. Towards a better understanding of different access and reuse policies. Washington DC: National Academies Press (Ch.6).

in Europa steeds verder zagen dalen vanwege de angst van de consumenten dat de data op Chinese telefoons niet veilig was voor de Chinese overheid.

Een tweede stimulans voor de Europese IT-bedrijven was het feit dat de Europese overheden de infrastructuur die nodig was om grote hoeveelheden data op te slaan en te verwerken, tot vitale infrastructuren bestempelen. Er werd zowel door overheden als door Europese bedrijven op grote schaal geïnvesteerd in de Big Data-kennisinfrastructuur van Europese universiteiten en onderzoeksinstituten. Dit gold niet alleen voor investeringen in hardware en software maar ook in mensen. Alleen al in Nederland kwamen er binnen een tijdsbestek van vijf jaar tien nieuwe masteropleidingen op het gebied van Big Data bij, variërend van hardware (Big-Data-computing) en software (machine learning en Big-Data-algoritmes) tot data entrepreneurship en *social computing*.[74] Daarnaast investeerde de Nederlandse overheid ook miljoenen in data zelf. Het aantal datasets dat door DANS werd beheerd nam in de periode 2015-2020 exponentieel toe. Dat kwam niet alleen omdat onderzoekers verplicht waren om hun datasets – die ze immers met publieke gelden hadden verzameld – bij DANS te deponeren maar ook omdat steeds meer bedrijven hun datasets – via DANS – voor publiek onderzoek ter beschikking stelden.

De Nederlandse wetenschap voer wel bij de hausse aan publiek-private investeringen in Big Data. Een voordeel hierbij was dat de Nederlandse wetenschappers een sterke traditie hadden in het delen van informatie en van oudsher gewend waren om samen te werken in omvangrijke internationale *data federations*. In het buitenland werd met afgunst gekeken naar het onderzoeksklimaat in Nederland, en met name naar het Netherlands eScience Center (NLeSC) dat een belangrijke brugfunctie speelde tussen de traditionele data-intensieve disciplines (zoals astronomie en klimaatwetenschappen) en sociale wetenschappen en geesteswetenschappen.

Het in een vroeg stadium van het onderzoek delen van data was rond 2015 nog zeker geen gemeengoed in andere disciplines zoals chemie en moleculaire biologie. Dat kwam omdat de wetenschappelijke credits exclusief gingen naar de auteurs die op basis van de data resultaten in toptijdschriften publiceerden – niet naar de wetenschappers en dataspecialisten die de brondata hadden verzameld en verwerkt. Onder druk van wetenschapsfinanciers kwam de traditie van geheimhouding steeds meer onder druk te staan. Bij publiek gefinancierd onderzoek waren wetenschappers sowieso verplicht alle data – en de algoritmes die ze hebben gebruikt om de data te bewerken en te analyseren – te ontsluiten voor derde partijen. In 2019 werd de verplichting uitgebreid naar privaat gefinancierd onderzoek. De (veelal Amerikaanse) bedrijven die zich verzetten tegen deze bepaling ("het Trojaanse paard van transparantie") trokken zich terug uit publiek-privaat onderzoek. Ze werden niet gemist – hun plaats werd geruisloos ingenomen door de nieuwe Europese Big-Data-bedrijven.

Tegelijkertijd vond er op universiteiten en publieke onderzoeksinstellingen een stille revolutie plaats: discipline-overstijgende functionele specialisten zoals data scientists kregen een eigen, gelijkwaardig carrièrespoor, naast de traditionele inhoudelijke disciplinegebonden sporen.

4.4.2 De tweede Verlichting

Het vrij circuleren van data was een groot goed maar dat leidde niet zonder meer tot gelijke kansen in de samenleving. Zonder de middelen om gebruik te maken van de vrijheid, dat

[74] In de geest van de Collective Awareness Platforms for Sustainability and Social Innovation (CAPS) van de Europese Commissie, https://ec.europa.eu/digital-agenda/en/collective-awareness-platforms-sustainability-and-social-innovation.

wil zeggen zonder de vaardigheden om de data te kunnen analyseren en te interpreteren, hadden burgers weinig aan data. De Nederlandse overheid zag al in een vroeg stadium het belang van het aanleren van deze vaardigheden in. De investeringen in human resources aan universiteiten en onderzoeksinstellingen werden al snel uitgebreid naar de basis. Data-vaardigheid werd een verplicht vak op het middelbaar beroepsonderwijs en datawetenschap werd een keuzevak op het VWO. De focus bij deze vakken lag in het leren stellen van de juiste (onderzoeks)vragen: waar komt informatie vandaan, op welke bronnen is ze gebaseerd en hoe is ze tot stand gekomen (dat wil zeggen op welke aannames beslissingen van informatiesystemen gebaseerd zijn)? Het ging dus steeds om het kunnen interpreteren van resultaten (van Big-Data-analyses) door ze te plaatsen in een bredere context.[75]

In de sociale wetenschappen vond een belangrijke methodologische verschuiving plaats; de traditionele surveys met relatief kleine steekproeven werden vervangen door analyses van de directe neerslag van menselijk gedrag (de *social physics* van Pentland et al.). Resultaten van sociaalwetenschappelijke onderzoeken werden dus niet langer gebaseerd op subjectieve meningen ('woorden') maar op objectief gedrag ('daden'). De uitdaging voor de (amateur)wetenschapper kwam nu – net als voor leerlingen en studenten – te liggen in de 'rijke' interpretatie van de steriele 'arme' patronen die werden gevonden in de Big-Data-analyses. Die 'rijke' interpretatie van grote hoeveelheden data vereiste een combinatie van typische alfa-, bèta- en gamma-vaardigheden. Omdat voor het kunnen plaatsen van de resultaten in een bredere context altijd grondige domeinkennis nodig was, ontstonden er nu op steeds meer plaatsen in de wereld domein-specifieke onderzoeksgroepen waarbinnen sociale wetenschappers de overkoepelende onderzoeksvragen stelden, data scientists de data verzamelden en classificeerden, en geesteswetenschappers de afwijkende cases door middel van close reading verder analyseerden. Veel interessante bevindingen bleken juist uit deze outliers voort te komen, en de data-analyses werden op basis van deze anomalieën vaak bijgesteld. Het was de ideale combinatie van onvermoeibare maar rechtlijnige agents en feilbare maar creatieve human agents.

Het optimisme over de mogelijkheden van Big Data bleef ondertussen onveranderd hoog. Het vooruitgangsgeloof werd gevoed door een brede maatschappelijke stroming die, net als tijdens de Verlichting, veel nadruk legde op het opleiden van autonome, kritisch denkende individuen. Enter de *homo informans*: de mens die zichzelf informeert en formeert, en die (weer) vrij durft te denken (*saupere aude!*). Het afwerpen van de mythes en het bijgeloof uit de Verlichting werd vervangen door het vermogen om voorbij de eigen vooroordelen en biases te denken (en die van anderen).

De uitkomsten van datagedreven sociaalwetenschappelijk onderzoek (inclusief onderzoeksjournalistiek) bood voor het eerst in de geschiedenis aan grote groepen burgers een meedogenloze maar onpartijdige blik op de werkelijkheid. Deze uitkomsten waren immers niet langer gebaseerd op subjectieve meningen maar op objectieve feiten. De grote uitdaging van de 21e eeuw was om – letterlijk op een verlichte manier – met de stortvloed aan feiten om te leren gaan. De onderzoeksjournalistiek, die rond de eeuwwisseling nog op

[75] Zodat ze bijvoorbeeld door de Simpson-paradox heen kunnen prikken: de reden dat overboord geslagen zeelieden met zwemvest vaker verdrinken dan zeelieden zonder zwemvest is omdat zwemvesten vaker met storm worden gedragen – en bij storm is de kans op het redden van een over boord geslagen zeeman veel lager dan bij rustiger weer, wanneer er geen zwemvesten worden gedragen. De oorzaak van een hogere verdrinkingspercentage is dus de aanwezigheid van storm, niet van het dragen van een zwemvest (voorbeeld ontleend aan: S. Kloos en N. Wielaard (2014). *Wij zijn Big Data. De toekomst van de informatiesamenleving*, p.147-149).

sterven na dood was door de teloorgang van traditionele papieren media, beleefde een revival en raakte steeds nauwer verweven met de datagedreven sociale wetenschappen.

4.4.3 De samenleving als data-ecosysteem

De voortdurend uitdijende stroom aan Big-Data-analyses leverde een onverwacht inzicht op dat een grote impact kreeg op de manier waarop de samenleving werd vormgegeven. Dat inzicht was dat wij allemaal, als autonome individuen, in contact met elkaar staan en in hoge mate van elkaar afhankelijk zijn. Steeds opnieuw bleek namelijk dat voor de duiding van specifieke patronen naar steeds bredere verklaringen moest worden gezocht – alles hangt met alles samen. Al een decennium eerder waren mensen zich steeds meer bewust geworden van het feit dat ze altijd en overal in een virtueel netwerk van concrete relaties leven. Iedereen was een uniek knooppunt in een virtuele wereld van relaties.[76] Maar er werd toen nog steeds vooral vanuit zichzelf geredeneerd: men stond vooral in het centrum van zijn eigen belangstelling.[77] Verder onderschatte men de individuele invloed om op kleine schaal iets aan bredere, maatschappelijke kwesties te doen.

Uit Big-Data-analyses kwam echter steeds duidelijker het beeld naar voren dat op macroniveau alle grote patronen uiteindelijk gebaseerd zijn op de interactie tussen miljoenen min of meer willekeurige beslissingen op microniveau; orde – het berekenbare – is het resultaat van chaos. Die ogenschijnlijke chaos werd steeds verder ontrafeld. Steeds beter waren wetenschappers in staat om te beschrijven hoe kleine veranderingen op microniveau via een exponentiële vermeerdering van netwerken, opeens tot enorme verschuivingen konden leiden (de zogenaamde *tipping points*). De schaal waarop de veranderingen plaatsvonden bleek te worden bepaald door (persoonlijke) netwerken en vooral door de creatieve *koppelingen* tussen al die netwerken.[78] Die angst die rond de jaren 2010 nog bestond dat het alom verspreide gebruik van mobiele devices tot een hyper-individuele samenleving zou leiden (‘eigen scherm eerst’) bleek dus ongegrond. Juist samenwerking kwam steeds centraler te staan in de samenleving – die ook steeds meer een samenleving van autonome individuen werd.

Voor het (her)gebruik van data – de zuurstof van moderne samenleving – had die samenwerking nog wel de nodige voeten in de aarde. Dat kwam omdat er een nieuwe soort van tripartiete samenwerking tussen (Europese) Big-Data-bedrijven, wetenschappers en burgers nodig bleek te zijn met duidelijke verantwoordelijkheden en waarborgen. Pas vanaf 2015 ontstonden er, vanuit de domeinspecifieke interdisciplinaire Big-Data-onderzoeksgroepen, bredere samenwerkingsverbanden met bedrijven, overheden en civic-society-organisaties. Dit soort data-ecosystemen vormden knooppunten waar partijen die data bezitten (burgers, bedrijven, overheden) op een verantwoordelijke manier – zowel financieel als inhoudelijk als juridisch – hun data en hun inzichten konden delen.

In succesvolle ecosystemen bleken niet alleen duidelijke regels gesteld aan de vraag tegen welke vergoeding inzichten mochten worden gegenereerd en gebruikt, maar ook over de manier waarop de inzichten tot stand waren gekomen en voor welke doeleinden ze wel of niet mochten worden gebruikt.[79] Door de jaren heen kristalliseerde zich een ideaaltypische

---

[76] H.A.F. Oosterling (2013). De Tweede Verlichting: verlichten, verheffen, of vervlechten. Het Nut van het Nut in de 21e-eeuwse netwerksamenleving. http://www.nutslezing.nl/uploads/pdf/NUT-2013.pdf.

[77] Of, zoals Marc Zuckerberg, de oprichter van Facebook, het stelt: “[een] eekhoorn die doodgaat in je voortuin kan voor jou op dit moment relevanter zijn dan stervende mensen in Afrika.” (D. Kirkpatrick (2011). *The Facebook Effect*. Geciteerd in Kloos en Wielaard (2014), op.cit.).

[78] H.A.F. Oosterling (2013). Op.cit.

[79] Kloos en Wielaard (2014), op.cit., p.209.

structuur van data-ecosystemen uit. De essentie daarvan was dat de data niet naar een (centrale) plek werd gebracht waar ze werd geanalyseerd, maar dat de analyse zoveel mogelijk naar de data toe werd gebracht. Deze methode kwam voort uit de astronomie en natuurkunde waarin analyses werden opgesplitst in subanalyses die werden uitgevoerd door de eigenaren van de data zelf.[80] In de praktijk bleek dat in veel gevallen slechts voor een klein deel van de analyse afgeleide inzichten nodig waren van andere partijen uit het ecosysteem. Alleen voor dat deel hoefde dus data te worden uitgewisseld.[81] Dat gebeurde in bijna alle subsystemen door middel van een Trusted Third Party die werd beheerd door een overheidsorganisatie uit het ecosysteem.

[80] Ibid, p.210.

[81] Ibid, p.211

5 De volgende stap

5.1 Inleiding

In de vorige twee hoofdstukken hebben we enkele Big-Data-scenario's geschetst voor het onderwijs en de wetenschap. Hoewel we niet pretenderen met deze scenario's een waarheidsgetrouw beeld van de toekomst (2025) te hebben geschetst - dat was ook niet de bedoeling - kunnen we onszelf wel afvragen wat we van deze exercitie kunnen leren. Welke keuzes kunnen we tegenkomen op de weg naar 2025? In dit hoofdstuk formuleren we daarom een aantal keuzemomenten die zich de komende jaren *kunnen* aandienen. Aan deze keuzes kan het ministerie van OCW (met partners) aandacht schenken bij vervolgstappen in de domeinen van Big Data in het onderwijs en in de wetenschap.

5.2 Keuzemomenten

We hebben een vijftal hoofdkeuzes geïdentificeerd. We stellen niet voor in welke richting deze (politieke) keuzes gemaakt moeten worden, maar wel dat er onderwerpen en ontwikkelingen zijn rondom Big Data die een keuze vragen van het ministerie (en haar partners) en waar de keuze zich op zou kunnen toespitsen. Deze keuze kan de vorm hebben van bijvoorbeeld het aangaan van een dialoog met betrokken partijen over de impact van Big Data, het aanpassen van wetgeving, of het stimuleren van Big-Data-vaardigheden onder de bevolking. Wij benoemen deze keuzes kort en geven een beknopte handreiking waar op gelet moet worden. Soms zullen we refereren aan de scenario's uit de voorgaande hoofdstukken. Het is immers mogelijk dat daar een keuze uit voortvloeit voor het ministerie.

5.2.1 Keuzemoment 1: Identificeren van rol(len) die het ministerie van OCW moet spelen in ontwikkelingen rondom Big Data in het onderwijs en in de wetenschap

Het ministerie van OCW zou – in navolging van de dialoog over de impact van robotisering op de arbeidsmarkt [82] – een maatschappelijke dialoog kunnen initiëren (of deelnemen aan een lopend debat) over het gebruik en de impact van Big Data in het onderwijs en de wetenschap.[83] Welke invloed heeft Big Data op onze scholen en leerlingen? Welke grenzen willen we stellen aan het gebruik van Big Data voor wetenschappelijk onderzoek? Het is nodig om een genuanceerd debat te voeren over de mogelijkheden en onmogelijkheden van Big Data. We kunnen nu reeds constateren dat het gebruik van Big Data gekoppeld raakt aan het vertrouwen dat burgers hebben in de overheid, in het onderwijs en in de wetenschap. Het ministerie van OCW kan de uitkomsten van een dergelijke dialoog gebruiken als input voor een visie op Big Data die niet als te repressief, maar ook niet als overmatig stimulerend

---

[82] Op 29 september 2014 heeft de minister van Sociale Zaken en Werkgelegenheid (SZW) op het SZW congres een toespraak gehouden waarin hij "robotisering van arbeid" agendeerde als een ontwikkeling die grote impact kan hebben op de Nederlandse arbeidsmarkt. Bron: http://www.rijksoverheid.nl/documenten-en-publicaties/toespraken/2014/09/29/robotisering-kansen-voor-morgen-toespraak-van-minister-asscher-tijdens-het-szw-congres-op-29-9-2014.html.

[83] Het onderwijs en de wetenschap schenken inmiddels volop aandacht aan de kansen en impact van Big Data. Kennisnet ('Big Data, van hype naar actie. Op zoek naar waardevolle inzichten voor het vergroten van studiesucces'), Het Expertisecentrum Mediawijsheid.net (www.mediawijsheid.nl/big-data/) en het Expertisecentrum Big Data van de Fontys Hogescholen (fontys.nl/Over-Fontys/Fontys-Hogeschool-ICT/Expertisecentrum-Big-Data.htm) besteden expliciete aandacht aan Big Data. Ook wetenschappelijke instellingen tonen groeiende belangstelling voor Big Data, bijvoorbeeld NWO (zie voetnoot 3) en het KNAW dat minisymposia over dit fenomeen organiseert.

wordt ervaren. Het is volgens ons weinig realistisch te denken dat een dialoog de maatschappelijke tegenstellingen zal overbruggen, maar er kan vertrouwen worden gewonnen wanneer OCW aantoont hoe zij zich verbindt aan de verschillende opvattingen in het onderwijs en in de wetenschap, hoe deze verbintenis doorwerkt in haar organisatie en beleid en welke acties zij met partners (in onderwijs en wetenschap) onderneemt om eventueel gestold wantrouwen over het gebruik van Big Data in bijvoorbeeld onderwijs (zie paragraaf 3.2.2) te verminderen.[84]

Wij onderscheiden drie rollen die het ministerie van OCW kan vervullen in het domein van Big Data in onderwijs en wetenschap: (1) kaderstellen en controleren; (2) aanjagen; en (3) produceren, verzamelen en verstrekken. Deze rollen staan op volgorde van de mate waarin het ministerie Big Data adopteert.

- **Kaderstellen en controleren:** het ministerie is een centrale speler in de voorbereiding, uitvoering en handhaving van beleid en wetgeving in de domeinen onderwijs en wetenschap. De algemene bescherming van privacy is geregeld in de Wet bescherming persoonsgegevens. Wanneer onderwijs- en onderzoeksinstellingen persoonlijke gegevens over ons verzamelen, delen en (her)gebruiken – bijvoorbeeld voor profilering – om onze slaagkans voor een opleiding te bepalen, kan dat privacy aantasten. In de dystopieën over het onderwijs hebben we aangeven waartoe dat kan leiden. Dit kan leiden tot betere en op maat gemaakte diensten, maar de keerzijde is dat er ook persoonlijke informatie kan worden gebruikt ten behoeve van ongewenste profilering. Het ministerie van OCW kan dus nagaan of wetgeving in beide domeinen aanpassing behoeft of dat bijvoorbeeld de Onderwijsinspectie meer moet gaan controleren op datagebruik. Zie ook de volgende keuze.
- **Aanjagen:** het ministerie kan het gebruik van Big Data aanjagen. Voorbeelden van het aanjagen door het ministerie zijn:
    - Agenderen van het belang van Big Data door een antennefunctie te vervullen (waarvan dit rapport een voorbeeld is) of door dialoog aan te gaan met (maatschappelijke) partners over de impact van Big Data op onderwijs en wetenschap.
    - Stimuleren van onderzoek naar de rol van Big Data in het onderwijs en in de wetenschap, bijvoorbeeld door geld beschikbaar te stellen aan onderzoeksinstellingen. Hieronder valt ook het creëren van experimenteerruimte (zie het kader).

[84] Onderwijspartijen waarschuwen dat OCW niet alleen moet besluiten en afrekenen met sturingsinformatie op basis van kengetallen en rendementen (hetgeen nog eens versterkt wordt met Big Data). Big Data gaat voorbij aan de vraag wat kwaliteit van onderwijs is en vergt dus ook ander soort informatie dan alleen middels informatiesystemen verzameld wordt.

Experimenteerruimte Big Data

Big Data kan leiden tot innovaties die voor een eventuele opschaling eerst getest kunnen worden in een veilige omgeving. Dit is een experimenteerruimte (digitaal / fysiek) waar Big-Data-toepassingen ontworpen, ontwikkeld en getest kunnen worden voordat zij hun weg vinden in de 'echte' wereld. Veel wetenschappelijke disciplines zijn uiteraard gewend om in laboratoria onderzoek te doen. Zelfs in de gedragswetenschappen wordt met living labs en simulaties gewerkt. Voor het onderwijs is dit minder evident. Ook hier is echter denkbaar dat Big-Data-toepassingen kleinschalig getest en onderzocht worden. Dit kan soms heel praktisch, bijvoorbeeld middels 'learning analytics' in een MOOC (Massive Open Online Course).

Een ander voorbeeld zou kunnen zijn dat studenten op vrijwillige basis deelnemen aan een experiment waarin een onderwijsinstelling op basis van allerhande gegevens die zij over een student binnen en buiten de instelling verzameld heeft een advies geeft over het vervolg van een opleiding (stoppen, specialisatie, stage, …).

Voor deze digitale of fysieke omgeving (denk aan een 'living lab') gelden andere, mindere strenge restricties die toelaten om uitgebreid te testen en te onderzoeken welke voor- en nadelen Big-Data-toepassingen hebben, hoe de eindgebruiker tegen een dienst of product aankijkt en onder welke voorwaarden de innovatie opgeschaald kan worden naar de 'echte' wereld. Deze testruimte kan zowel door overheden, bedrijven, onderwijsinstellingen als burgers worden gebruikt. In een dergelijke gesimuleerde omgeving kan dus ook nagegaan worden of een beveiligingstechniek werkt en welke partijen mogelijk interesse zullen hebben in de data.

Dit geldt bijvoorbeeld sterk voor het onderwijs. Het ministerie zou vooral moeten durven experimenteren met verbeteringen in het onderwijs door inzet van Big Data. Daarbij kan de expertise van professionals goed worden ingezet. Vanuit die experimenten ontstaan langzamerhand de contouren van het speelveld van inzet van Big Data in het onderwijs. Zonder die experimenten blijft het te veel gissen naar toegevoegde waarde en zal deze moeizaam worden gerealiseerd.

    - Verzamelen en verspreiden van goede voorbeelden over het (positief) gebruik van Big Data in het onderwijs en de wetenschap en deze actief verspreiden. Dit kan onderdeel zijn van een bredere benadering waarin onderwijs en wetenschap – voor zover nodig – bewust wordt gemaakt van kansen en bedreigingen die Big Data bieden.

- **Produceren, verzamelen en verstrekken**: Het ministerie van OCW kan het goede voorbeeld geven bij Big Data, zodat partijen in het onderwijs en de wetenschap dit voorbeeld kunnen volgen waardoor het vertrouwen van leerlingen, docenten, respondenten over Big Data in het onderwijs en de wetenschap kan toenemen. Ministeries zijn zelf een grote partij waar het gaat om het verzamelen, uitwisselen en bewerken van gegevens. Dit positioneert ook het ministerie van OCW als een vragende en aanbiedende partij, maar maakt haar meteen ook kwetsbaar. In de afgelopen periode zijn er binnen en buiten de overheid enkele incidenten geweest (denk aan de ING) die duidelijk maken dat datagebruik grote zorgvuldigheid vraagt van betrokken partijen. Er zijn ook mogelijkheden om Big Data intern effectief in te zetten, enerzijds gericht op 'evidence-based policy-making' en anderzijds voor het verbeteren van bedrijfs- en uitvoeringsprocessen (bijv. fraudebestrijding op basis van risicoprofielen en bestandskoppeling van overheden). Ook kan OCW gebruik

aanmoedigen middels een open-data-strategie waarin gegevens – binnen de juridische kaders – openbaar worden gemaakt voor hergebruik door derden en het delen van 'best practices'. Wanneer het ministerie van OCW Big Data toepast om 'slimme' diensten en producten voor de eindgebruiker te ontwikkelen, zou het vertrouwen van deze eindgebruiker (burger, bedrijf) in de overheid en Big Data wel eens kunnen toenemen, mits hierover duidelijk en transparant gecommuniceerd wordt. Hiertoe zou het ministerie een interne verkenning kunnen laten uitvoeren naar de datasets die zij 'op de plank' heeft liggen en onder welke voorwaarden deze als open data ontsloten kunnen worden.[85]

5.2.2 *Keuzemoment 2: Stimuleren van het gebruik van Big Data binnen de kaders van de bescherming van privacy enerzijds en de stimulering van het benutten van kansen anderzijds*

De mogelijke aantasting van privacy is één van de grootste zorgpunten die tegenstanders (maar ook voorstanders) opwerpen tegen Big Data. Deze zou kunnen uitmonden in 'Big Brother'. Dit is niet alleen van toepassing in voor de hand liggende disciplines als marketing, maar net zo goed in onder andere het onderwijsdomein. Denk bijvoorbeeld aan het gebruik van data analytics bij digitale leerondersteuning in klaslokalen.[86] Tegelijkertijd constateren – vooral voorstanders van Big Data – dat te stringente wetgeving rondom Big Data een rem zet op relevante en innovatieve ontwikkelingen, bijvoorbeeld in de geneeskunde of het onderwijs. Er lijkt wat betreft bij het tegemoet komen aan privacywensen sprake te zijn van voortdurend balanceren op een dun koord.

Overigens zou – op basis van de geschetste scenario's – een privacyissue zeer actueel kunnen worden: De Amerikaanse overheid zou met het argument van nationale veiligheid altijd toegang kunnen eisen tot data die op de servers van Amerikaanse bedrijven staan. Dat geldt dus ook voor Nederlandse data op Amerikaanse servers. De opkomst van cloud computing en de dominantie van Amerikaanse bedrijven daarin is een groot privacyrisico. In dit verband kan het ministerie Nederlandse onderwijsinstellingen en universiteiten mogelijk verplichten om data op Nederlands grondgebied op te slaan, door Nederlandse bedrijven en in Nederlandse data centers. Mogelijk is hier een Europese actie gewenst.

De Amerikaanse Big-Data-expert Pentland stelt dat het realiseren van de mogelijkheden van een datagedreven samenleving om een beleidsontwerp (een "New Deal") voor data vraagt. Dit ontwerp moet werkbare garanties geven zodat betrokkenen kunnen beschikken over de data die nodig is voor het algemeen belang, terwijl tegelijkertijd de burgers worden beschermd. Het handhaven van de bescherming van de persoonlijke privacy en vrijheid is essentieel voor een succesvolle samenleving, aldus Pentland.[87] Misschien dat een dergelijk omvattend beleidsontwerp een stap te ver is, maar het is wellicht mogelijk te overwegen of een dergelijk ontwerp in het onderwijs of de wetenschap toegevoegde waarde biedt. Het is

---

[85] Open data is (1) uit publieke middelen bekostigd; (2) gegenereerd bij of voor de uitvoering van een publieke taak, (3) openbaar; (4) vrij van auteursrechten of andere rechten van derden; (5) computer-leesbaar; (6) voldoen bij voorkeur aan de 'open standaarden' (geen pdf, wel xml of csv); en zijn voor hergebruik beschikbaar zonder beperkingen, zoals kosten of verplichte registratie Algemene Rekenkamer (2014), Trendrapport Open Data, Den Haag.

[86] Podesta, J., Pritzker, P., Moniz, E.J., Holdren, J. & Zients, J. (2014), *Big data: Seizing opportunities, preserving values*, Executive Office of the President, Washington.

[87] Pentland spreek uitdrukkelijk over toegang tot data voor het *algemeen* belang (en niet voor een *particulier* belang). Pentland, A. (2014), *Sociale Big Data. Opkomst van de datagedreven samenleving*, Maven Publishing BV, Amsterdam, p. 35.

dus mogelijk dat het beschermen van privacy en het benutten van kansen naast elkaar bestaan.

Het ministerie van OCW *kan* inzake de bescherming van persoonsgegevens in het kader van Big Data de volgende activiteiten ondernemen, namelijk:

- Ga na of de huidige wetgeving voor onderwijs en wetenschap in het licht van Big Data voldoende waarborgen biedt tegen een aantasting van de bescherming van persoonsgegevens en tegelijkertijd geen onnodige drempels opwerpt voor het gebruik en de toepassing van Big Data in beide sectoren.[88] Op basis van deze analyse kunnen wetten en regels aangepast worden.[89] Blijf daarbij in lijn met andere Europese landen (met het oog op level playing field).

- Instrumenten (laten) ontwikkelen waardoor belanghebbenden in het onderwijs en in de wetenschap meer en betere sturing kunnen geven op welke publieke en private partijen welke data over hun verzamelen, waar deze data opgeslagen wordt en waar deze data voor gebruikt (mag) worden.[90] De leerling, docent, proefpersoon en respondent worden dus meer in positie gebracht als eigenaar c.q. beheerder van de data.[91]

- Het – waar nodig – uitrusten van toezichthouders (bijvoorbeeld de Onderwijsinspectie) op de naleving van relevante wetgeving ten aanzien van datagebruik en –bescherming in beide domeinen.

- Het stimuleren van de ontwikkeling van zelfregulering, gedragscodes en -regels in het onderwijs en de wetenschap gericht op het correct gebruik van Big Data (en transparantie daarover), eventueel gepaard met voorlichting en begeleiding van scholen en kennisinstellingen. Ook het stimuleren van het gebruik van een Big-Data-keurmerk zou hieronder kunnen vallen, zoals het Nederlands Privacy Keurmerk | Big Data.[92]

---

[88] Hier speelt nu en in de komende jaren het vraagstuk dat privacywetgeving ten aanzien van persoonsgegevens achterhaald kan zijn, omdat personen bijvoorbeeld altijd te traceren, ook uit 'indirecte' gegevens. Dat wordt alleen maar erger met opkomst Internet of Things (zie voetnoot 7). Bovendien is doelbinding achterhaald, want dat blokkeert het hergebruik data voor maatschappelijke doeleinden. Een verbod op het verzamelen persoonsgegevens lijkt dus dweilen met de kraan open. Het ministerie van OCW zou in het onderwijs en de wetenschap kunnen aansturen op het *gebruik* van persoonsgegevens in laats van louter regulering en daarbij het targeten van individuen kunnen verbieden tenzij daar een urgente maatschappelijke noodzaak voor is.

[89] Een adviesrapport aan het Witte Huis stelt dat Big Data de potentie heeft "to eclipse longstanding civil rights protection" op het gebied van persoonlijke informatie, gevolgd door een advies voor uitbreiding van huidige wetgeving. Een duidelijke indicatie dat een dergelijke toetsing ook in Nederland niet overbodig is. (Podesta, J., Pritzker, P., Moniz, E.J., Holdren, J. & Zients, J. (2014). *Big data: Seizing opportunities, preserving values*. Executive Office of the President). Een voorbeeld van een artikel dat voor een dergelijke toets in aanmerking komt, is Artikel 40b van de Wet op het primair onderwijs. Dit artikel bepaalt welke gegevens over een leerling verstrekt moeten worden bij toelating.

[90] Denk aan 'opt-in' en 'opt-out' constructies die vergelijkbaar zijn met het Bel-me-niet Register.

[91] Dit zal zijn grenzen kennen, omdat leerlingen bijvoorbeeld verplicht zijn om bepaalde gegevens te verstrekken, bijvoorbeeld in het kader van een aanvraag voor een studielening. Desondanks kunnen gebruikers middels zogenaamde 'data lockers' (virtuele kluizen) persoonsgegevens beveiligd opslaan en beheren.

[92] http://stichtingprivacy.nl/nederlands-privacy-keurmerk.

- Altijd transparant zijn over de afwegingen die het ministerie van OCW maakt bij het (niet) gebruiken van Big Data in eigen bedrijfs- en beleidsvoering.

5.2.3 *Keuzemoment 3: In welke mate moeten het ministerie en het onderwijsveld investeren in generieke en specifieke competentie-ontwikkeling om met Big Data om te kunnen gaan*

Het is belangrijk dat de bevolking – en meer specifiek leerlingen, studenten, docenten, onderzoekers, … - de vaardigheden ontwikkelt om met Big Data te kunnen omgaan. Of het nu gaat om bewustzijn van de 'kruimels' die digitaal gedrag achterlaten, sturing geven aan data die derden mogen gebruiken (en controleren), het toepassen van Big Data in onderwijs of het aanbieden van Data-Science-opleidingen. Het ontwikkelen van de vaardigheden om met Big Data om te kunnen gaan, is dus een breed thema.

Democratiseer via onderwijs, opleidingen, gedragscodes en desnoods wetgeving de specialistische organisaties en nieuwe professies die op basis van (Big) Data politieke vraagstukken in ogenschijnlijk neutrale statistische problemen vertalen (denk in dit verband aan de probabilistische revolutie). Deze vertalingen onttrekken zich voor een belangrijk deel aan het publieke oog. De politiek verplaatst zich naar de kantoren van modellenbouwers en statistische bureaus.[93] Bij de meer traditionele beroepsgroepen en organisaties die tussen staat en burger staan (zoals juristen, notarissen, politiemensen) zijn de taken en verantwoordelijkheden nauwkeurig in de wet geregeld. Voor deze nieuwe professies (zoals accountants, statistici, verzekeringsdeskundigen, epidemiologen en risico-analisten) bestaan er nog nauwelijks publieke verantwoordingskaders of checks and balances.[94] Ook zou educatie van deze groepen bij kunnen dragen aan kennis van en bewustzijn over de privacyimplicaties van hun werk. Onderwijsinstellingen kunnen hier een grote rol in spelen, zoals nu in de VS lijkt te gebeuren.[95]

Dit kan aangevuld worden met een meer specifieke benadering, waarin de overheid in overleg met partners (onderwijs, wetenschap, bedrijfsleven) investeert in de ontwikkeling van specialistische Big-Data-vaardigheden (zgn. 'hard skills') die bijdragen aan een goede (internationale) positionering van Nederland in het domein van Big Data. Het gaat dan om investeringen in het opleiden van Data Scientists, het verrichten van fundamenteel en toegepast Big-Data-onderzoek in voor Nederland relevante wetenschapsdomeinen, het ontwikkelen van valorisatiemechanismen om opgedane kennis te vertalen naar commerciële toepassingen. In een onderzoek van e-skills bleek dat het aantal Big-Data-specialisten dat in het VK bij grote bedrijven werkt in de komende vijf jaar waarschijnlijk meer dan 240% zal stijgen.[96] Gezien de relevantie van Big Data voor de (Nederlandse) publieke sector kan men hier niet in achter blijven.

Redenerend vanuit de scenario's kan gewezen worden op een issue dat nu reeds speelt, namelijk de opkomst van digital social sciences. We hebben eerder gewezen dat de opkomst Big Data vooral in de sociale en geesteswetenschappen leidt tot grote verschuivingen. Het ministerie kan met het onderwijs en de universiteiten een plan trekken om *in den brede* in de data analytic skills van (sociale) wetenschappers te investeren. Daarbij moet begonnen worden aan de basis. In het primair onderwijs kan extra aandacht worden gegeven aan kwantitatieve vaardigheden. Ongecijferdheid zal alleen nog maar een nog kritischer

---

[93] Gerard de Vries, op.cit.

[94] ibid.

[95] PCAST (2014). Big data and privacy: A technological perspective.

[96] Zie voetnoot 15.

probleem worden dan het al was (dus leve de rekentoets …). Besteed in wiskundeonderwijs op de middelbare school (veel) meer aandacht aan data-analyse.

Dit kan gezien worden als onderdeel van een meer algemeen transitiemanagement waarin het ministerie van OCW met partners oplossingen definieert en aan verwachtingenmanagement doet om onze kennissamenleving te begeleiden in de transitie naar een meer datagedreven samenleving en ons daar internationaal – ook in onderwijs en wetenschap – goed te positioneren.

5.2.4 Keuzemoment 4: Stimuleren dat partijen in het onderwijs en in de wetenschap actief data gaan delen

Naarmate de economische, maatschappelijke en wetenschappelijke waarde toeneemt, zal competitief gedrag groeien bij partijen om data af te schermen voor derden (zie ook paragraaf 4.2.2). Op zich is dit geen nieuw fenomeen, want private partijen gebruiken nu ook unieke marktkennis om hun positie te verbeteren. Ook de overheid ontwikkelt soms beleid op basis van een kennisvoorsprong ten opzichte van de omgeving. Big Data betekent – zoals we in hoofdstuk 2 aangaven – juist dat kansen benut worden door het slim combineren van uiteenlopende datasets (die vaak door verschillende partijen – publiek en privaat – beheerd worden). Ook het onderwijs en de wetenschap beschikken over datasets die mogelijk interessant zijn voor hergebruik door derden. Het ministerie van OCW zou dus kunnen stimuleren dat publieke en private partijen – binnen de wettelijke kaders – gegevens gaan uitwisselen. Hieronder valt ook de lopende open-access-discussie met commerciële wetenschappelijke uitgevers. Nederland loopt internationaal voorop in de 'open access' en het ministerie kan voet bij stuk houden om 'open access' in de wetenschap zoveel mogelijk te stimuleren.

Wanneer er goede wettelijke afspraken bestaan over het gebruik en de toepassing van Big Data kan het ministerie gericht beleid voeren om het onderwijs en de wetenschap aan te sporen gegevens uit te wisselen, zodat de voordelen van Big Data benut gaan worden. Dat kan nog een hele uitdaging zijn, omdat het delen van gegevens nog geen gemeengoed is (niet in de wetenschap, niet in het onderwijs en ook niet in het bedrijfsleven); zeker waar het gegevens betreft die economische voordelen bieden aan individuele partijen. Niet voor niets spoorde de Research Data Alliance-Europe in december 2014 nog de Europese Commissie aan om het delen van (wetenschappelijke) data te stimuleren en faciliteren.[97] Nu is het nog zo dat het eigen belang ertoe leidt dat iedereen op haar of zijn data gaat zitten, met het gevolg dat er nauwelijks data beschikbaar komt voor hergebruik en verdere analyse. Dit geldt vooral voor grote bedrijven – en dan met name voor de nieuwe soort ('Big Data') bedrijven die data als basis hebben voor hun economische activiteiten (zoals Google, Facebook, Microsoft etc.). Ook maatschappelijke weerstand gevoed door wantrouwen belemmert (her)gebruik van grote gegevensbestanden.[98]

Het eenvoudig uitwisselen van gegevens vergt niet alleen een goed beheer van data, maar eveneens afspraken over technische standaarden, gemeenschappelijke investeringen,

---

[97] Research Data Alliance Europe (2014). The Data Harvest: How sharing research data can yield knowledge, jobs and growth.

[98] Volgens de wet zijn ziekenhuizen verplicht om op een gestandaardiseerde manier hun activiteiten te rapporteren aan verzekeraars (anders krijgen ze de behandelingen niet vergoed). Deze gegevens worden centraal opgeslagen. Terwijl deze wetgeving voor diagnosebehandeling (DBC) geruisloos is ingevoerd en ziekenhuizen binnen een paar jaar al hun data op een uniforme manier uitwisselen, is het elektronisch patiëntendossier (EPD) vanwege de grote maatschappelijke weerstand uiteindelijk afgevoerd.

mogelijk ook onderliggende verdienmodellen en waarvoor data wel en niet mag worden gebruikt.[99] Dat is een grote uitdaging, want van deze standaardisatie is in het onderwijs en de wetenschap nog geen sprake. Hier is een rol voor het ministerie weggelegd, bijvoorbeeld bij het faciliteren van een grootschalige IT-infrastructuur. In dit verband is het verstandig dat het ministerie een meer proactieve houding aanneemt en relevante sectoren – onderwijs en wetenschap – uitnodigt om een strategie te gaan vormen omtrent uitwisseling en (her)gebruik van data.

Parallel aan deze stimulerende activiteiten is het ministerie aan zet om het brede publiek en eigenaars van data continu te informeren welke stappen gezet worden door verschillende partijen omtrent (her)gebruik van data (of deze partijen te verplichten derden daarover te informeren).

5.2.5 Keuzemoment 5: Het onderwijs en de wetenschap actief stimuleren om aansluiting te zoeken bij internationale ontwikkelingen omtrent Big Data

Alle scenario's hebben vermeld dat Big Data per definitie een internationale ontwikkeling is. Hoe goed de Nederlandse wetenschap en het Nederlandse onderwijs aangesloten zijn bij deze ontwikkeling is medebepalend voor het succes waarmee we Big Data toepassen, zowel economisch, maatschappelijk als wetenschappelijk.

Big Data kent zowel 'harde' kanten (ontwikkeling van hardware, software en netwerken) als 'zachte' kanten (toepassingsdomeinen). Het is voor ons land onmogelijk om bij elke internationale ontwikkeling aan te sluiten laat staan de eerste viool te spelen. Daarom valt te overwegen dat het ministerie van OCW (samen met EZ) na laat gaan op welke aspecten van Big Data ons land internationaal een onderscheid kan maken (niche) en daar met relevante partners afspraken over maakt hoe deze internationale kansen te verzilveren (ambitie, rollen, middelen, samenwerking, …). Voor de wetenschap geldt bijvoorbeeld dat Nederland een unieke positie heeft wat betreft de kwaliteit van de wetenschappelijke IT-infrastructuur (netwerken en opslag) of vakgebieden die internationaal vooroplopen met gebruik van grote datasets (bijv. astronomie en geneeskunde) (zie ook paragraaf 4.2.1). Zo heeft de TU/e in 2013 het Data Science Center Eindhoven geopend. Op het vlak van onderwijs zou nagegaan kunnen worden of Nederland toonaangevende Data-Science-opleidingen en specialisaties kan aanbieden.

Het is dan de kunst om die niches te vinden waar Nederland op het vlak van onderwijs en wetenschap zou kunnen excelleren, maar ook die bedrijven, onderwijs en kennisinstellingen te identificeren die moeten samenwerken om deze kansen te verzilveren. Nederland zou bijvoorbeeld kunnen aansluiten bij landen en industriële sectoren waar een evidente rol voor Nederlandse kennisinstellingen en bedrijven is weggelegd of eigen sterkte sectoren als uitgangspunt kunnen nemen voor verdere toepassingen van Big Data, bijvoorbeeld een van de Topsectoren. Het onderbenutten van Big-Data-kansen zal ongetwijfeld leiden tot verlies aan concurrentiekracht.[100]

---

[99] Dit geldt vooral bij de vermenging van publieke en private data. Persoonsgegevens over een leerling kunnen alleen gebruikt worden voor het leerproces in de onderwijsinstelling en – zonder persoonlijke en juridische toestemming – gebruikt worden door educatieve uitgevers.

[100] De Vlaamse Raad voor Wetenschap en Innovatie (VRWI) heeft in samenspraak met bedrijven, onderwijs en kennisinstellingen Big Data geïdentificeerd als één van de sleutelthema's in de Vlaamse Digital Society 2025. VRWI (2014), *VRWI Toekomstverkenningen 2025*, Studiereeks 26, Brussel, p. 48-51.

Bijlage 1: Geïnterviewde personen

Wil van der Aalst, TU/e.

Mary Berkhout, Mediawijzer.net.

Remond Fijneman, VU / CTMM / TraIT.

Erik Fleur, Dienst Uitvoering Onderwijs

Lex Freund, Hogeschool Rotterdam.

Dennis Groot, KPN.

René Hageman, VSNU.

Floris Kreiken, Bits for Freedom.

Han van der Maas, Oefenweb.

Anneke Mathijssen, Radboud Universiteit / bestuurslid DAIR.

Maarten de Rijke, UvA.

Marten Roorda, CITO.

Eric van Tol, Fontys.

Albert Vlaardingerbroek, Noorderpoort College.

Rein Willems, STT.

Guido van Wingen, AMC.

Marius van Zandwijk, Kennisnet.

Contact:
Dialogic
Hooghiemstraplein 33-36
3514 AX Utrecht
Tel. +31 (0)30 215 05 80
Fax +31 (0)30 215 05 95
www.dialogic.nl