De datagedreven samenleving

Achtergrondstudie

Linda Kool, Jelte Timmer & Rinie van Est

Rathenau Instituut

Het **Rathenau Instituut** stimuleert de publieke en politieke meningsvorming over wetenschap en technologie. Daartoe doet het instituut onderzoek naar de organisatie en ontwikkeling van het wetenschapsysteem, publiceert het over maatschappelijke effecten van nieuwe technologieën, en organiseert het debatten over vraagstukken en dilemma's op het gebied van wetenschap en technologie.

De datagedreven samenleving

Achtergrondstudie

Rathenau Instituut
Anna van Saksenlaan 51

Postadres:
Postbus 95366
2509 CJ Den Haag

Telefoon: 070-342 15 42
E-mail: info@rathenau.nl
Website: www.rathenau.nl

Uitgever: Rathenau Instituut
Opmaak: Boven de Bank, Zeist
Redactie: Duidelijke Taal tekstproducties
Coverbeeld: Hollandse Hoogte
Beeld pagina 19: Hollandse Hoogte

ISBN/EAN: 978-90-77364-71-0

Deze publicatie kan als volgt worden aangehaald/ Preferred citation:
Kool, L, J. Timmer & R. van Est, De datagedreven samenleving - Achtergrondstudie, Den Haag, Rathenau Instituut 2015

De datagedreven samenleving

Achtergrondstudie

Linda Kool, Jelte Timmer & Rinie van Est

Bestuur Rathenau Instituut

mw. G.A. Verbeet (voorzitter)
prof. dr. E.H.L. Aarts
prof. dr. ir. W.E. Bijker
prof. dr. R. Cools
dr. H.J.M. Dröge
drs. E.J.F.B. van Huis
prof. dr. ir. H.W. Lintsen
prof. mr. J.E.J. Prins
prof. dr. M.C. van der Wende
dr. ir. M.M.C.G. Peters (secretaris)

Voorwoord

Als ik praat met politici, beleidsmakers, onderzoekers en mijn buren, dan zie ik een groot vertrouwen waar in big data. Het ideaal van de maakbare samenleving bloeit bij velen op. Big data, denken zij, gaan onze maatschappelijke problemen oplossen. Zo kunnen dijken dankzij sensoren zelf alarm slaan als onderhoud nodig is. Twitterberichten signaleren over opkomende werkloosheid. En politie hoopt met slimme camera's crimineel of gewelddadig gedrag op te sporen en de buurt veiliger te maken. De datagedreven samenleving biedt inderdaad kansen voor een betere dienstverlening en een efficiëntere bedrijfsvoering.

Tegelijkertijd rijzen er ook steeds meer vragen over het gebruik van big data. Vragen over privacy, gelijke behandeling en eigenaarschap, maar ook over de kwaliteit en de betrouwbaarheid van gegevens. Hoe accuraat zijn de gegevens en bijbehorende analyses eigenlijk? Wie weet nog precies hoe een data-analyse in elkaar steekt en welke keuzes daarbij worden gemaakt? Zijn we er op tijd bij als er fouten in het systeem sluipen en weten we hoe dat te corrigeren?

Het Rathenau Instituut bestudeerde hoe het potentieel van big data op een verantwoorde manier kan worden gerealiseerd. Deze studie maakt deel uit van het onderzoek naar de hyperconnectieve consument binnen het thema 'de Meetbare Mens' in ons werkprogramma.

Deze publicatie is bedoeld voor beslissers en beleidsmakers die meer willen weten over wat big data eigenlijk zijn, en welke maatschappelijke en economische kwesties samenhangen met de inzet van big data. Op basis van literatuurstudie en gesprekken met stakeholders brachten we het nationale en internationale debat hieromtrent in kaart.

Het Rathenau Instituut hoopt met deze studie een bijdrage te leveren aan de verdere gedachtevorming over verantwoord datagebruik. Dat krijgt echter pas echt vorm in de praktijk. Ik zie deze publicatie daarom als startpunt voor een gezamenlijke dialoog tussen overheden, bedrijven, maatschappelijk middenveld en mijn buren, waarin de voorwaarden voor verantwoord datagebruik nader invulling krijgen.

Dr. ir. Melanie Peters
Directeur Rathenau Instituut

Inhoudsopgave

1 Inleiding

> *If you asked me to describe the rising philosophy of the day, I'd say it is data-ism. We now have the ability to gather huge amounts of data. This ability seems to carry with it certain cultural assumptions - that everything that can be measured should be measured; that data is a transparent and reliable lens that allows us to filter out emotionalism and ideology.*
> (David Brooks 2013)

De datagedreven samenleving is volop in ontwikkeling. De explosief groeiende hoeveelheid digitale gegevens in onze samenleving biedt een nieuwe grondstof voor innovatie, en brengt nieuwe economische en maatschappelijke kansen met zich mee. De verwachting is dat *big data* ons de 'macroscoop' leveren: een instrument waarmee we het 'geheel' in beeld krijgen en dat ons daardoor in staat stelt om sneller, efficiënter of goedkoper te organiseren. Steeds meer organisaties ontdekken dat ze over enorme hoeveelheden data beschikken die mogelijk waardevol zijn voor hun bedrijfsproces of hun maatschappelijke doelstelling. Er valt een groeiend geloof en vertrouwen in big data te bespeuren, en een sterke drang om te geloven dat elk fenomeen te kwantificeren is en de problemen van de wereld op te lossen zijn met big data. Big data bevinden zich in een 'goudkoortsfase': veel beslissers en investeerders hebben hoge verwachtingen van big data. Tegelijkertijd is echter onduidelijk wat nu precies big data zijn, en hoe waardevolle big data op een verantwoorde manier ontgonnen kunnen worden.

Het gebruik van data en slimme software roept inmiddels ook steeds meer vragen op. Bijvoorbeeld over de kwaliteit van de gebruikte data, het eigenaarschap en de beveiliging van gegevens, maar ook de privacybescherming, en de controle en het toezicht op complexe softwaresystemen, die een steeds grotere rol spelen bij het maken van beslissingen. Internationaal, en ook in Nederland, worden studies of werkgroepen gestart om de implicaties van big data te onderzoeken (zie kader 1.1.). In deze achtergrondstudie brengen we de stand van zaken van dit debat rondom big data in kaart. Waarover gaat de discussie, nationaal en internationaal? Welke (contrasterende) visies spelen daarbij een rol?

Deze achtergrondstudie heeft een algemeen karakter; het brengt algemene kenmerken van big data in kaart. Maar het is belangrijk om te beseffen dat big data een rol spelen in vele domeinen van de samenleving en dat in die verschillende domeinen telkens andere, specifieke vragen opdoemen. Denk aan de zorg of de financiële sector, waarin verschillende omstandigheden en regulerende kaders verschillende vragen met zich mee brengen over innovatiekansen en -risico's. In deze achtergrondstudie concentreren we ons op de overkoepelende thema's die in elk domein spelen. Om meer zicht te krijgen op de specifieke dynamiek rondom innovatie met big data is er parallel aan deze studie een onderzoek gedaan naar de impact van big data in de verzekeringssector (Timmer, Kool en van Est 2015).

Deze achtergrondstudie is tot stand gekomen op basis van literatuuronderzoek, bezochte nationale en internationale bijeenkomsten en daar gesproken stakeholders (zie bijlage 1).

De studie is bedoeld voor zowel beslissers als beleidsmakers die meer willen weten over wat big data nu precies zijn, en welke maatschappelijke en economische kwesties samenhangen met het gebruik en de inzet van big data. Vijf thema's staan daarbij centraal:

1. *Kwaliteit en veiligheid van datasharing*: wat is de kwaliteit van de gebruikte datasets, en welke garanties over kwaliteit en beveiliging kunnen organisaties elkaar geven? Is de meerwaarde van delen voor alle partijen aanwezig? Denk bijvoorbeeld aan een gemeente, die waardevolle inzichten kan verkrijgen uit data van celmasten over mensenstromen op verschillende locaties en tijden in de stad. Het is echter niet op voorhand duidelijk hoe de telecomproviders kunnen profiteren van het vrijgeven van die data, terwijl hieraan voor hen wel risico's kleven: onder meer een negatieve publieke opinie, het mogelijk (onbedoeld of onbewust) delen van bedrijfsgeheimen en/of het lekken van data.

2. *Privacy, autonomie en gelijke behandeling*: big data gaan uit van zoveel mogelijk gegevens verzamelen en combineren, om er waarde uit te kunnen halen. Hoe verhoudt dat zich tot het dataprotectierecht, waarin begrippen als dataminimalisatie (zo min mogelijk gegevens verzamelen), doelbinding (alleen voor specifiek omschreven doel) en proportionaliteit (alleen in verhouding tot dat doel) centraal staan? Met het gebruik van big data intensifeert ook het gebruik van (groeps)profielen, waarmee organisaties beslissen over de behandeling van specifieke individuen, bijvoorbeeld de prijs of toegang tot een dienst of product. Als er essentiële verschillen in behandeling ontstaan, kan dat discriminerend zijn of leiden tot ongerechtvaardigde uitsluiting. Burgers en consumenten hebben slechts beperkt zicht op de groeiende berg data die overheden en bedrijfsleven over hen verzamelen en gebruiken en welke gevolgen zij daarvan ondervinden. Dat maakt bezwaar maken lastig. Hun digitale autonomie staat op het spel.

3. *De grenzen van en keuzes in big data*: hoe goed is slimme software eigenlijk? Wat representeert een dataset eigenlijk precies? Hoe kan het toezicht op systemen waarin software soms automatisch beslissingen neemt, in microseconden, worden georganiseerd? Denk aan de automatische handel op de financiële beurzen, waarbij bedrijven soms grote financiële verliezen leiden door, bijvoorbeeld een 'softwarebug'. Of aan Amerikaanse kredietbeoordelingssystemen waarin software iemands kredietwaardigheid bepaalt. Hoe bepaalde scores tot stand komen, is onduidelijk. Als klanten een lening geweigerd wordt, wordt niet meege-

deeld waarom. Ook als het systeem een 'denkfout' maakt of verkeerde data gebruikt, is het moeilijk daartegen bezwaar te maken. Hoewel een kredietscore een subjectieve inschatting is van iemands kredietwaardigheid, wordt hieraan in de praktijk een onwrikbare waarde gehecht (Citron & Pasquale 2014).

4. *Benodigde competenties en vaardigheden*: tekortschietende kennis over het kunnen toepassen van big dat wordt gezien als belemmering voor innovatie. Het gaat ten eerste om een tekort aan datascientists die zich ook bewust zijn van de beperkingen van data en slimme software. Een datascientist moet meer zijn dan alleen een goede computerwetenschapper of mathematicus, hij of zij moet ook oog hebben voor de sociale context van een vraagstuk of bepaalde dataset. Ten tweede gaat het om een tekort aan gespecialiseerde managers en beleidsmakers. Naarmate organisatieprocessen meer datagedreven worden, komen ook meer managers en beleidsmakers hiermee direct in aanraking. Dat vraagt ook van hen speciale vaardigheden: kennis en kunde om een verantwoorde inschatting te kunnen maken hoe data-analyses het beste worden ingezet bijvoorbeeld, en welke beslissingen hierop kunnen worden genomen.

5. *Veranderende machtsverhoudingen*: met de opkomst van een datagedreven economie ontstaan nieuwe verdienmodellen, waarin het verwerken van gegevens centraal staan. Deze modellen zijn een bron van innovatie en economische groei. Wie bezit de data? Wie kan er van profiteren? Voor bedrijven en overheden speelt de vraag welke partijen data en software bezitten, en wie daar in economisch opzicht het meeste voordeel uit haalt. Wat zijn de sterktes en zwaktes van Nederland, Europa, de Verenigde Staten en andere delen van de wereld?

Kader 1.1 Overzicht van nationale en internationale beleidsstudies

De potentie van big data is beschreven vanuit veel verschillende invalshoeken. Een vaak aangehaald rapport is dat van consultancybedrijf McKinsey (2011), Big data: The next frontier for innovation, competition, and productivity. McKinsey stelt dat big data onmisbaar zijn voor toekomstige waardecreatie in de publieke en private sector. Ook IT-bedrijven zoals IBM en Microsoft publiceerden rapporten over de toepassing en kansen van big data (Hey, Tansley & Tolle 2009; Microsoft 2011; IBM 2012a, 2012b).

Het World Economic Forum (WEF) publiceerde in 2012 een studie waarin persoonlijke gegevens een waardevol bezit worden genoemd, en een zorgvuldige omgang met deze gegevens een belangrijke voorwaarde voor het gebruik van big data. In het rapport Rethinking Personal Data: A New Lens for Strengthening Trust signaleerde het WEF dat het gebrek aan vertrouwen in het data-ecosysteem een remmend effect kan hebben op de ontwikkeling van big data (WEF 2014).

De Organisatie voor Economische Samenwerking en Ontwikkeling (OESO) leverde in 2008 een belangrijke bijdrage aan de ontwikkeling van *open data*, met een serie aanbevelingen hoe informatie uit publieke sector vrij beschikbaar moet worden gemaakt om nieuwe waardecreatie mogelijk te maken (OESO 2008). Het Europees Parlement wijzigde in 2013 haar Open-datarichtlijn (2013/37/EU) en stimuleert door de oprichting van een Europees Open Data Portal de ontwikkeling van open data.

De OESO signaleerde in 2013 diverse maatschappelijke en bestuurlijke kwesties rondom big data (2013b), onder andere over privacy en consumentenbescherming, open data, veiligheidsrisico's, benodigde vaardigheden en de benodigde infrastructuur. In de Verenigde Staten riep het Witte Huis een werkgroep in het leven om de problemen rondom big data en privacy in kaart te brengen. Begin 2014 kwam die werkgroep met het rapport Big Data: Seizing Opportunities, Preserving Values, waarin ze beschrijft hoe big data de economie en samenleving veranderen en waarin ze aanbevelingen doet om de privacy van de burger te beschermen. In Europa worden eveneens veel discussies gevoerd over de gevolgen van big data voor de privacy onder andere naar aanleiding van de nieuwe Europese Dataprotectiewetgeving (bijvoorbeeld Cate, Cullen & Mayer-Schönberger 2013; Cavoukian, Dix & Emam 2014). De Europese Commissie publiceerde in 2014 een strategische visie voor een datagedreven economie (EC 2014).

Ook Nederland denkt na over de mogelijkheden en de vraagstukken rondom big data. De kabinetsvisie op e-privacy behandelt de impact van big data op de bescherming van de persoonlijke levenssfeer (Kamerstukken II 2012-13). Het Centraal Planbureau (CPB) bracht in april 2014 een beleidsbrief uit over de markt voor persoonsgegevens. Het ministerie van Economische Zaken publiceerde eind 2014 een Kamerbrief over big data en privacy in de private sector (EZ 2014).Daarin wordt aangegeven aan dat een strikt juridische benadering niet voldoende zal zijn om vertrouwen te creëren tussen burgers, consumenten, overheden en bedrijfsleven en zoekt de oplossing in de invulling van drie randvoorwaarden voor vertrouwen: controle van de burger over zijn eigen gegevens, meer transparantie en verantwoordelijkheid van bedrijven. Het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties zal inventariseren welke impact big data heeft op de uitoefening van diverse grondrechten zoals het recht op bescherming van de persoonlijke levenssfeer, het recht op gelijke behandeling, de vrijheid van meningsuiting en de vrijheid van vereniging. Het ministerie van Economische Zaken stelt een high level expert groep in. De opdracht voor de expert groep is de relatie tussen big data en profilering en de bescherming van grondrechten verder te verkennen en oplossingsrichtingen uit te werken voor het verenigen van twee doelen: het benutten van de mogelijkheden van big data enerzijds en het behoud van vertrouwen van de samenleving in het internet anderzijds.

Naar aanleiding van de motie van Tweede Kamerlid Segers is de Wetenschappelijke Raad voor het Regeringsbeleid (WRR) door de (toenmalige) ministers Opstelten (VenJ) en Plasterk (BZK) gevraagd om advies te geven over de beleidsvraagstukken rondom big-datatoepassingen (Kamerstukken II 2013-14b).

2 Wat zijn big data?

De datagedreven samenleving is volop in ontwikkeling. Er komen steeds meer digitale gegevens beschikbaar. Aansprekende voorbeelden maken duidelijk dat die data van grote waarde kunnen zijn voor organisaties. Zo gebruikt de politie in Los Angeles de data van eerdere inbraakmeldingen om effectievere patrouilleroutes uit te stippelen. Op deze manier kunnen ze politieagenten effectiever inzetten en de misdaad beter bestrijden.[1] Rijkswaterstaat wil op grote schaal sensoren gaan inzetten om de toestand van de dijken in de gaten te houden.[2] Door die beter en vooral preciezer te monitoren verwacht Rijkswaterstaat de dijken goedkoper en beter te kunnen onderhouden. Zoektermen bij Google vormen mogelijk een alternatieve manier om een opkomende griepepidemie te volgen. Bepaalde opgezochte termen blijken te correleren met het verloop van een griepepidemie.[3] Slimme software op de beurs tot slot bepaalt automatisch welke aandelen worden verhandeld:[4] in fracties van seconden worden aandelen gekocht en verkocht, en wordt gespeculeerd op kleine wijzigingen in valuta- en aandelenkoersen. Door snel te kopen, te verkopen en weer te kopen kunnen bij kleine koersverwachtingen grote winsten gemaakt worden.

De voortschrijdende digitalisering van de samenleving zorgt voor een continue stroom en groei van gegevens. De data zijn bijvoorbeeld afkomstig van camera's, smartphones, tablets en draagbare apparaten, browsers en sociale netwerken. Maar ook van clouddiensten, en van de toenemende toepassing van sensoren in producten en machines, die non-stop in verbinding staan met het internet, waardoor men spreekt van het Internet of Things en het Industrial Internet (Evans & Annunziata 2012). Met name de stroom van gegevens die internetbedrijven verwerken, is indrukwekkend. Zo verwerkten de servers van Google in 2013 meer dan 24 miljoen gigabytes aan data per dag – dat is duizend keer de hoeveelheid van het gedrukt materiaal in de bibliotheek van het Amerikaanse congres (Mayer-Schonberger & Cukier 2013a). De Amerikaanse supermarkt Walmart handelde in 2012 meer dan 1 miljoen klanttransacties af per uur. Elk dag worden er worden meer dan 10 miljoen nieuwe foto's geüpload naar Facebook. En dit zijn allemaal cijfers van enkele jaren oud, de versnelling en de vooruitgang gaan nog steeds door.

---

1 http://www.theguardian.com/cities/2014/jun/25/predicting-crime-lapd-los-angeles-police-data-analysis-algorithm-minority-report

2 http://publicaties.minienm.nl/documenten/macrostabiliteit-ijkdijk-sensor-en-meettechnologie-samenwerking

3 In 2008 geeft Google zelf aan dat het Flu Trends een griepepidemie kan voorspellen met 97% accuraatheid. Maar wetenschappers rapporteren in de daaropvolgende jaren problemen met de accuraatheid van Flu Trends, zie bijvoorbeeld Lazer et al. (2014).

4 AFM (2010) In Nederland bepalen slimme algoritmes zo'n 40-50% van de handel. In de Verenigde Staten ligt dit hoger op ca 70%.

De totale hoeveelheid data in de wereld wordt inmiddels gemeten in exabytes (1 exabyte is 1 miljard gigabytes) en lijkt zijn eigen 'wet van Moore' te kennen: elke twee tot drie jaar verdubbelt de hoeveelheid data (Mayer-Schonberger & Cukier 2013b; IDC 2014). Organisaties verzamelen grote hoeveelheden data over vrijwel elk aspect van hun bedrijfsproces. Die data kunnen op van alles betrekking hebben, variërend van de status van een machine, communicatie-patronen, gedrag van mensen (gemeten in gps, kliks, browsergegevens of zoektermen) en relaties (sociale netwerken) tot financiën, gegevens over gezondheid en ziekteverzuim. Simpele gebeurtenissen kunnen zeer diverse data opleveren. Een enkel twitterbericht bijvoorbeeld bevat verschillende gegevens: het bericht zelf, met of zonder link met tekst, foto of video, gegevens over de browser en het mobiele apparaat waarmee het bericht verzonden is, de locatie waar het apparaat zich bevindt, informatie over mensen die het binnen hun socialenetwerk doorsturen (retweeten), de aard en de hoeveelheid van de reacties die het bericht oproept en de reacties die het bericht in de bredere media oproept.

Kader 2.1 Veiligheid op straat door big data

Het project CityPulse heeft als doel de veiligheid te verbeteren in het populaire uitgaansgebied Stratumseind in Eindhoven. Samen met Dutch Institute of Safety & Security (DITSS) en Intel heeft Atos CityPulse ontwikkeld. Grote aantallen slimme camera's houden doorlopend in de gaten wat er op straat gebeurt. Waar dat nodig is, kan de politie snel ingrijpen. Met behulp van big-data-intelligence – het snel analyseren van zeer grote hoeveelheden gegevens – lukt het de overheid dikwijls om mogelijke incidenten een stap voor te zijn.

Zowel big data als realtime-data-analyses worden benut. Op straat registreren de slimme camera's waar individuen zich bevinden, hoe snel ze bewegen en voor welk café de meeste mensen staan. Data-analisten destilleren patronen en inzichten uit de verzamelde data, en vullen die aan met gegevens van sociale media. Die informatie wordt vervolgens doorgespeeld aan bestuur en politie. Die bepaalt dan of specifieke acties nodig zijn. De privacy blijft daarbij gewaarborgd: de meetgegevens zijn niet te herleiden tot individuele personen.

Het systeem heeft het niet altijd bij het rechte eind: een ingestudeerde dans voor de opening van een nieuw café – waarbij twee groepen zich tegenover elkaar opstelden en in elkaars richting bewogen – werd door de slimme camera's geregistreerd als het begin van een vechtpartij. Hoewel dat gebeurde in de beginfase van het project, en de lerende software inmiddels beter is geworden, valt niet uit te sluiten dat het

systeem onschuldige burgers als verdachte aanmerkt bij afwijkend gedrag; bijvoorbeeld als iemand in zijn eentje rondjes loopt om zijn vrienden te zoeken.

Bron:
Kist, R. en Noort, van R. (2015) Het misdrijf is al ontdekt voor het gepleegd is. NRC, 22 augustus 2015.
Keizer, R. (2015) Eindhoven blijft veilig met Big Data. De automatiseringsgids, 15 juli 2015. http://www.automatiseringgids.nl/nieuws/2015/30/eindhoven-veilig-met-big-data

2.1 De uitdaging van grote datavolumes

De gegevensbestanden zijn dusdanig groot en divers dat er nieuwe of andere IT-architecturen nodig zijn om deze gegevens te kunnen opslaan en te analyseren. De 'oude' databasestructuur voldoet niet meer. Daarin wordt van te voren bepaald welke eigenschappen van de data in velden ('records') worden opgeslagen. Andere – nog te ontdekken – eigenschappen worden zo op voorhand aan het gezicht onttrokken. Nieuwe IT-architecturen moeten rekening houden met wat het Amerikaanse adviesbureau Gartner (2011) aanduidde als de onderscheidende eigenschappen van big data: *volume*, *variety* en *velocity* (de 3 V's). Ze moeten grote volumes van zeer variërende data aankunnen, die met een hoge snelheid continu (realtime) worden aangevuld met nieuwe data.

Betrouwbaarheid van datavolumes

Hoewel de drie V's van Gartner de kenmerken van de nieuwe IT-architecturen goed illustreren, raken ze de kern van big data nog niet. Een groot probleem bij de grote hoeveelheden te verwerken data is de betrouwbaarheid. Daarom heeft IBM (2013) een vierde V als kenmerk van big data toegevoegd: *veracity* oftewel 'waarachtigheid'.[5] De data en de analyses zijn niet altijd betrouwbaar en de betekenis van de gevonden correlaties is niet altijd duidelijk (zie hoofdstuk 5). Tegelijkertijd is er ook een groot vertrouwen – of zelfs geloof – in de resultaten van big data. Soms lijkt het erop dat mensen denken dat big data een nieuwe objectieve kennisbron vormen. Boyd en Crawford (2011) noemen dit de 'mythe' van big data (zie hoofdstuk 5).

Waardecreatie uit grote datavolumes

Het gaat dus niet alleen om 'technische' kenmerken van data, maar ook om het doel van de analyses. Big data creëren goudkoorts, omdat de data en de mogelijke nieuwe informatie die daaruit verworven kan worden, grote economische en maatschappelijke waarde hebben. De OESO (2013b) kent dan ook nog een vijfde V toe: *value*. Door slimme dataminingtechnieken en toegenomen rekenkracht kunnen deze enorme databestanden inmiddels snel doorzocht worden en nieuwe, waardevolle verbanden en patronen ontdekt worden. Big data worden dan ook niet gekenmerkt door grote hoeveelheden data alleen, maar door de combinatie van die datahoeveelheden en software die bijvoorbeeld op basis van patroonherkenning en correlaties onverwachte – waardevolle – verbanden kan vinden.

Kader 2.2 De vijf V's van big data: volume, variety, velocity, veracity en value

Volume: het volume van de datasets wordt steeds groter

Variety: het gaat om diverse data, zoals databases, documenten, e-mail, video, beelden, audio, logs, financiële transacties e.d.

Velocity: betreft de snelheid waarmee data worden geproduceerd, en de verwerkingssnelheid (bijna realtime)

Veracity: de datasets zijn niet altijd 'schoon' of zonder fouten – het is belangrijk om datasets en analyses te testen op hun 'waarachtigheid'

Value: uit de datasets kan informatie gehaald worden door de data te analyseren, via correlaties en patroonherkenning

5 http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Bronnen:
Voor volume, variety & velocity, zie Gartner (2011). 'Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data', persbericht, www.gartner.com/it/page.jsp?id=1731916
Voor veracity, zie IBM (2013)
Voor value, zie OESO 2013b

Samenvattend kunnen we stellen dat big data verwijzen naar zowel een technologische als een economische en een maatschappelijke ontwikkeling: een (nabije)toekomstperspectief waarin het verzamelen en doorzoeken van data met slimme algoritmen nieuwe waardevolle kennis opleveren. De belofte van big data houdt een verschuiving in naar een door datagedreven economie en maatschappij.

3 Kwaliteit en veiligheid van datasharing

'*Data [are] the starting point.*' Zonder gegevens geen analyse. Ondanks de verhalen over het exponentieel uitdijende data-universum, zijn goede data in veel gevallen nog schaars. Data zijn niet zozeer waardevol omdat er veel data zijn, maar wanneer er relevante data zijn. Het verkrijgen van toegang tot waardevolle en kwalitatief goede gegevens om data-analyses op uit te voeren blijkt voor veel bedrijven die zich bezighouden met big data, niet eenvoudig.[6] De OESO (2013b) noemt 'toegang tot data' een van de grote uitdagingen bij het benutten van data-driven innovation. Een cruciale stap om dit te bereiken is het creëren van open-datasystemen. Open data zijn 'vrij' toegankelijke informatie; er berust geen auteursrecht of andere rechten van derden op en de data zijn leesbaar door de computer (*machine-readable*). De informatie is (online) beschikbaar, zodat anderen deze data kunnen hergebruiken. Denk bijvoorbeeld aan gegevens van het KNMI, waarmee Buienradar wordt gemaakt, of de gegevens van het Kadaster, waarmee websites en apps worden gemaakt om informatie over huizenprijzen te vinden. Voorwaarden en licenties beschrijven hoe de data (her)gebruikt mogen worden.

Er zijn diverse internationale initiatieven om open data te realiseren. Zo deed de OESO in 2008 aanbevelingen over het beschikbaar stellen van 'Publieke Sector Informatie' als open data (OESO 2008). Het Witte Huis lanceerde in 2013 een aantal open-data-initatieven.[7] Ook de Europese Unie werkt aan open data. Op grond van de Open-datarichtlijn 2013/37/EU is het beschikbaar stellen van publieke informatie als open data sinds juli 2015 verplicht. Om dit te faciliteren heeft de EU platformen gelanceerd, zoals de European Union Open Data Portal, waarin alle open datasets van Europese instanties te vinden zijn. Ook de Nederlandse overheid werkt actief aan het beschikbaar stellen van de eigen data, zie bijvoorbeeld het Actieplan open overheid (BZK 2013).

Dankzij deze initiatieven komen steeds meer gegevens beschikbaar. Een inventarisatie van het ministerie van Binnenlandse Zaken en Koninkrijksrelaties van juli 2015 toont bijvoorbeeld dat er inmiddels 550 datasets beschikbaar zijn voor hergebruik. Circa 300 meer zijn gepland of wordt onderzoek naar gedaan (BZK 2015). Toch is het grootste deel van de gegevens in het digitale universum nog onbenut of ontoegankelijk. In 2013 schatte marktbureau IDC in dat slechts een half procent van de 25 procent potentieel waardevolle data in het digitale universum op dit moment wordt benut voor analyses (IDC 2013). Het delen en beter benutten van data kent een aantal technische en organisatorische

6 Bijeenkomst Data Science Center, Eindhoven, DCS/e, 08-04-2014
7 http://www.whitehouse.gov/sites/default/files/microsites/ostp/2013opendata.pdf

beperkingen: de kwaliteit van gegevens en de manier waarop gegevens beschikbaar worden gesteld, het garanderen van de veiligheid van gegevens, en het creëren van vertrouwen en bereidheid om gegevens te delen. Deze beperkingen lichten we nader toe in deze paragraaf.

3.1 Datakwaliteit

Een belangrijke voorwaarde voor het uitvoeren van een goede data-analyse, is de beschikking over kwalitatief goede data. Daarvoor moeten data op een goede manier worden verzameld, 'getagd' (gemerkt) en worden opgeslagen in een machine-readable format. Het opbouwen van grote datasets van kwalitatief goede data stuit op problemen, wanneer de data van verschillende instanties, maar ook van verschillende afdelingen binnen bedrijven, bijeengebracht moeten worden. Doordat veel werkprocessen in afzonderlijke afdelingen zijn geoptimaliseerd, zijn gegevens uit die verschillende afdelingen niet zomaar met elkaar te combineren; ze komen bijvoorbeeld uit verschillende systemen.[8] Zo kunnen gegevens over de openbare ruimte soms aan postcodes gekoppeld zijn, soms aan telefoonnummers. Het is goed mogelijk om die gegevens te 'vertalen', maar alleen als bekend is welk bestand precies in welk format gebruikt wordt. Een ander voorbeeld komt uit de verzekeringsbranche. Veel verzekeraars hebben in het verleden verschillende polissen op verschillende manieren georganiseerd. De data van de autoverzekeringen en de brandverzekeringen van een en dezelfde verzekeraar – en zelfs de verzekeringen van een en dezelfde klant – zijn in verschillende formats opgeslagen en kunnen niet zomaar met elkaar worden gecombineerd (Timmer et al. 2015).

Datakwaliteit wordt een nog groter vraagstuk, wanneer organisaties gebruik maken van externe (data)bronnen. Het is moeilijk om te beoordelen of gegevens van een externe bron op een geschikte manier zijn verzameld en verwerkt, doordat elk datacentrum en iedere eigenaar van data hun eigen beleid hebben. Organisaties die met deze data aan de slag willen, kunnen niet anders dan vertrouwen op geboden garanties wat de kwaliteit van de gegevens betreft (zie ook paragraaf 3.2 over veiligheid).

Ook voor open data zijn garanties over de kwaliteit van gegevens wenselijk. Het simpelweg beschikbaar stellen van gegevens is niet genoeg. Het open-databeleid van de Verenigde Naties werd bijvoorbeeld door *The Guardian* bekritiseerd, omdat veel bestanden over ontwikkelingswerk alleen in pdf-format beschikbaar gesteld werden.[9] Pdf-bestanden zijn niet – of niet zomaar – machine-readable. De Nederlandse, Europese en Amerikaanse open-datarichtlijnen stellen daarom

8 Bijeenkomst Data Science Center, Eindhoven, DCS/e, 08-04-2014

9 http://www.theguardian.com/global-development-professionals-network/2013/oct/21/development-open-data-action

dat gegevens waar mogelijk in machine-readable format beschikbaar moeten worden gesteld.[10]

Ten slotte heeft datakwaliteit ook betrekking op de daadwerkelijke toegankelijkheid van de data. Om effectief gebruik te kunnen maken van de betrouwbare open data moeten gebruikers weten waar ze welke gegevens kunnen vinden en moeten ze zich bewust zijn van wat ze ermee kunnen (Esmeijer, Bakker & Munck 2013). Kortom, gegevens moeten op een zinvolle manier worden ontsloten en organisaties die open data willen gebruiken, moeten weten welke data er zijn en in hoeverre de gegevens betrouwbaar zijn.

3.2 Veiligheid

De waarde van big data zit vaak in het combineren en uitwisselen van gegevens. De toekomst van big data staat of valt met een goede uitwisseling. Tegelijkertijd roept de uitwisseling van gegevens ook veiligheidsvragen op. Dataveiligheid en dataprotectie zijn cruciale vraagstukken bij big data. En daar liggen nog grote uitdagingen. Marktbureau IDC berekende in 2013 dat terwijl voor 40 procent van alle gegevens in het digitale universum een vorm van dataprotectie of –beveiliging noodzakelijk is, slechts de helft van die gegevens daadwerkelijk beschermd is (IDC 2014).

De kwetsbaarheid is niet alleen het gevolg van eventueel tekortschietende beveiliging of gegevensbescherming bij de opslag van data, maar komt ook door het combineren van verschillende databronnen, afkomstig van meerdere partijen of infrastructuren. Een organisatie kan haar eigen informatiebeveiliging op orde hebben, maar is meestal ook afhankelijk van de gegevens van anderen. Volgens het WEF (2014) is het bijna onmogelijk om zicht te houden op verschillende datastromen en op de wijze waarop de data vervolgens wordt gebruikt. Dit is geen onwil, maar komt voort uit het gebrek aan gemeenschappelijke kaders. Er zijn wereldwijd geen gezamenlijke meetinstrumenten en normen over wat goed '*data stewardship*' inhoudt.

De ontwikkeling van een nationaal en internationaal kader voor *data stewardship* heeft dan ook een hoge prioriteit. Pas als er zo'n kader bestaat, is het mogelijk om zicht te krijgen op wie waar verantwoordelijk voor is in de keten van dataverzameling, opslag, verwerking en analyse. Een belangrijk onderdeel van zo'n kader is een vast systeem voor data 'taggen' en toegangsmonitoring. Op die manier wordt inzichtelijk wie toegang heeft tot welke gegevens en onder welke voorwaarden. Daarmee kan de toegang tot data worden gestuurd.

10 Zie richtlijn 2013/37/EU, en Executive Order 13642-Making Open and Machine Readable the New Default for Government Information.

'It gives them the ability to be almost like the GPS for data', aldus Peter Guerra, vice-president bij Booz Allen Hamilton.[11]

Beveiliging en gegevensbescherming zijn daarnaast belangrijk, omdat ook criminelen geïnteresseerd kunnen zijn in gegevens en analyses. Datalekken en andere beveiligingsincidenten zijn aan de orde van de dag. In 2014 werden bijvoorbeeld Target – met 70 miljoen records met namen, adressen en creditcardgegevens – en eBay – 140 miljoen accounts van klanten – getroffen door criminele hacks. De ontwikkeling van gecentraliseerde opslag en de verwerking van data in gespecialiseerde datawarehouses stellen bijzonder hoge eisen aan de gegevensbeveiliging om ongeautoriseerde toegang te voorkomen. Maar niet alleen criminelen zoeken toegang tot data. Ook inlichtingendiensten zijn geïnteresseerd in data en data-analyse. De onthullingen van Snowden laten zien dat inlichtingendiensten soms bewust beveiligingsmaatregelen verzwakken om hun toegang tot de data te vergemakkelijken (Greenwald 2013). Vertrouwelijk-heid van gegevens (confidentialiteit) op computers die in verbinding staan met het internet, lijkt moeilijk te garanderen (EP 2014). Dat vormt de achilleshiel van big data.

3.3 Vertrouwen

Voor het delen van gegevens tussen consumenten, bedrijven en overheidsorganisaties is vertrouwen een noodzakelijke voorwaarde. Zonder vertrouwen, zijn partijen niet bereid zijn om gegevens beschikbaar te stellen. Hetzelfde geldt voor consumenten. Ook zij zullen minder geneigd zijn bedrijven en organisaties toestemming te geven voor het gebruik van hun gegevens, wanneer ze die niet vertrouwen. Het WEF signaleert een gebrek aan vertrouwen in het data-ecosysteem en ziet dit als een fundamenteel probleem voor de economische ontwikkeling van big data (WEF 2014). Garanties op het gebied van datakwaliteit en gegevensbeveiliging dragen bij aan vertrouwen, maar daarnaast is er duidelijkheid nodig over hoe gegevens worden verwerkt en wie daarvoor verantwoordelijk is. Dit speelt op het niveau van de consument, bij interne gegevensuitwisseling in bedrijven, en bij uitwisseling van gegevens tussen bedrijven onderling.

Voor de consument is het onduidelijk of en aan wie zijn data worden doorverkocht of hoe die precies worden gebruikt in data-analyses, wanneer hij die data afstaat. Uit een survey van telecombedrijf Orange (2014) blijkt dat 78 procent van de consumenten bedrijven niet vertrouwt op dit punt. Dit gebrek aan vertrouwen kan bepalend zijn voor de publieke opinie over big data en de noodzakelijke gegevensuitwisseling, en kan zo een remmend effect hebben op de ontwikkeling van big data. De voorbeelden van ING bank en pintransactieverwerker Equens, die beide hun plannen om gegevens over transacties te

11 Gigaom's Structure Data conference https://gigaom.com/2014/03/19/how-to-make-big-data-more-secure-less-creepy/

vermarkten, schielijk terugtrokken, onderstrepen de kracht van de publieke opinie tegenover bedrijven die werkzaam zijn op de consumentenmarkt.

Bij het uitwisselen van gegevens binnen bedrijven kunnen er problemen ontstaan doordat verschillende betrokkenen niet bereid zijn om gegevens met elkaar te delen. Zelfs binnen één bedrijf laat de praktijk soms zien dat personen 'op de data gaan zitten', waardoor er verschillende datasilo's ontstaan.[12] Dit gebeurt bijvoorbeeld wanneer verschillende afdelingen binnen bedrijven conflicterende belangen hebben. Op het moment dat er geen inzicht is in hoe anderen data willen inzetten en of dat mogelijk strijdig is met de belangen van degenen die de data beschikbaar moeten stellen, zal de bereidheid om data te delen laag zijn.

Deze dynamiek speelt nog sterker tussen verschillende organisaties. Het combineren van gegevens uit verschillende bronnen of vanuit verschillende sectoren kan veel economische voordelen opleveren. De vraag is echter voor wie: *'the best thing to do with your data will be thought of by someone else'* (Rufus Pollock in Esmeijer, Bakker & Munck 2013, p. 37). Daarom is het delen van data in de praktijk vaak een moeizaam proces. De angst om bepaalde bedrijfsgevoelige informatie bloot te geven of een competitief voordeel te verliezen, speelt een rol. Maar ook het ontbreken van een duidelijke meerwaarde voor de partij die haar data beschikbaar stelt, is ook een belangrijke factor. *'Many sources of third-party data do not yet consider sharing their data, and economic incentives may not be aligned to do so.'* (OESO 2013b, p. 24.) Een voorbeeld uit de praktijk van Havenbedrijf Rotterdam illustreert dit. Het havenbedrijf beschrijft hoe het zijn logistieke processen wil verbeteren door gebruik te maken van bepaalde gegevens over binnenkomende schepen, maar die weigeren de geanonimiseerde gegevens die hiervoor nodig zijn, te delen (Esmeijer, Bakker & Munck 2013, p30).

Er is behoefte aan een robuust raamwerk dat delen van gegevens tussen bedrijven en organisaties mogelijk maakt (OESO 2013b). Volgens Alex Pentland, auteur van het boek *Sociale Big Data - Opkomst van de data-gedreven samenleving*, is een grote barrière voor toepassingen van big data het ontbreken van een licentiemodel en daarbij horende voorwaarden voor het gebruik van data. De partij die gegevens beschikbaar stelt, moet niet kunnen vertrouwen op een goede omgang met die gegevens, het moet ook duidelijk zijn dat zij zelf baat heeft bij het delen ervan.

3.4 Datamarkten

Op verschillende manieren wordt er aan oplossingen gewerkt om het delen en verkrijgen van toegang tot gegevens te verbeteren. Opkomende 'datamarkten'

12 Bijeenkomst Data Science Center, Eindhoven, DCS/e, 08-04-2014

kunnen volgens internetgoeroe Tim O'Reilly de infrastructuur bieden die het delen van data aantrekkelijker maakt. Onlinemarktplaatsen voor data zorgen namelijk niet alleen voor een makkelijke toegang, ze maken ook vergelijkbaarheid en kwaliteitsgaranties mogelijk, en ze kunnen er daarnaast voor zorgen dat de data worden opgeschoond en in de juiste formats worden aangeleverd (Dumbill 2012). Partijen die data bezitten, kunnen die tegen vergoeding op datamarkten aanbieden. Het verlenen van toegang tot data via een Application Programming Interface (API), waarbij softwareprogramma's met elkaar communiceren en gegevens uitwisselen, is een trend waarmee bedrijven meerwaarde aan hun gegevens geven of mogelijk maken.[13] Maar niet alleen bedrijven profiteren van data-uitwisseling: via start-ups zoals Personal.com en andere zogenoemde datakluisjes kunnen individuele consumenten hun persoonlijke gegevens bijeenbrengen, toegang tot deze gegevens managen en tegen een financiële vergoeding aanbieden aan commerciële partijen (Tene & Polonetsky 2012, p. 29).

Voor de uitwisseling en het combineren van gevoelige gegevens wordt ook naar oplossingen gezocht. Onderzoekers van het Intel Research Lab werken aan een systeem waarin datasets gecombineerd en geanalyseerd kunnen worden zonder dat verschillende partijen elkaars gegevens te zien krijgen. '*This is a neutral environment where parties can place their data and derive an answer without revealing their data to one another.*' (Sridhar Iyengar, directeur security research bij Intel Labs). Op deze manier kan een zogenoemde '*trusted third party*' zorgen voor vertrouwen en veiligheid bij datasharing. Tegelijkertijd signaleren kleinere bedrijven dat in een dergelijk 'big-dataspeelveld' wel het risico bestaat dat enkele grote IT-dienstverleners een centrale positie verwerven, en controle krijgen over heel veel data. Monopolisering van datatoegang kan de lokale innovatiekracht belemmeren.[14]

3.5 Discussiepunten

Uit het voorgaande blijkt dat datakwaliteit, adequate bescherming van data en vertrouwen van bedrijven en consumenten onontbeerlijk zijn om big data verder te ontwikkelen. Datamarkten kunnen een structuur bieden waarmee aan deze behoeften van consumenten en bedrijven kan worden voldaan. Een belangrijke vraag die rest, is hoe dergelijke marktplaatsen precies vorm moeten krijgen. Die vraag roept weer andere vragen op: Is hier nationale of internationale regelgeving voor nodig? Wie kan de belangen van verschillende stakeholders waarborgen? Welke verdere eisen moeten aan de betrouwbare data-uitwisseling gesteld worden?

13 Interview Microsoft, 16-04-2014
14 Bijeenkomst Data Science Center, Eindhoven, DCS/e, 08-04-2014

4 Privacy, autonomie en gelijke behandeling

Onder big data vallen veel verschillende soorten data: data uit dijksensoren voor een betere dijkbewaking of financiële data om beter te kunnen handelen op de beurs. Niet alle data zijn dus gerelateerd aan mensen of aan menselijk gedrag. Een groot deel van de digitale gegevens die nu beschikbaar zijn, hebben echter wel direct of indirect betrekking op individuen (IDC 2014). Analyse van die data kan commercieel heel interessant zijn, omdat bedrijven op basis van big data onder meer de mogelijkheid krijgen specifieke profielen over individuen op stellen waarmee ze 'passende' diensten kunnen aanbieden. Maar big-data-analyse beperkt zich niet tot het opstellen van klantprofielen. Met big data neemt ook het gebruik van *predictive analytics* neemt toe. Software probeert patronen te herkennen in huidig en historisch gedrag, om zo toekomstig gedrag te voorspellen. Voorspellende software wordt bijvoorbeeld gebruikt om mogelijke fraude bij verzekeringen, overheidsuitkeringen, creditcards of belastingaangiftes op te sporen. Het systeem berekent een (waarschijnlijkheids) score voor ieder individu om te bepalen of een bepaald gedrag zal optreden; bijvoorbeeld over de waarschijnlijkheid dat iemand een lening kan terugbetalen. De politie gebruikt voorspellende software om bijvoorbeeld vuurwapengevaarlijke personen te herkennen of om te bepalen waar agenten efficiënt op straat kunnen worden ingezet.[15] Nog een andere toepassing is *learning analytics,* waarmee in het onderwijs op persoonlijke behoeften en leerstijlen afgestemde onderwijsbegeleiding mogelijk gemaakt wordt.

Het gebruik van persoonsgebonden data roept vragen op privacybescherming. Zeker als door het slim combineren en analyseren van data uit verschillende bronnen steeds meer informatie over iemand ontsloten kan worden. Maar het gaat niet alleen om mogelijke aantasting van de privacy. Verregaande profilering op basis van bestaande gegevens en op basis van verwacht toekomstig gedrag raakt ook aan de autonomie van individuen (Hildebrandt 2015). Profilering kan ertoe leiden dat personen handelingsopties worden ontnomen ('u bent een kredietrisico en krijgt dus geen creditcard') of dat personen een andere behandeling krijgen ('dit telefoonnummer is van een kleine klant, die sluit dus achteraan in de wachtrij van de helpdesk'). Dit roept vragen op over ongewenste of onterechte uitsluiting van diensten, en het recht op gelijke behandeling. Bieden de bestaande juridische kaders voldoende waarborgen om het individu te beschermen? In deze paragraaf kijken we eerst naar hoe de 'logica' van big data en digitale innovatie zich verhoudt tot dataprotectie. Vervolgens kijken we naar de risico's van profilering.

15 Zie bijvoorbeeld http://www.sentient.nl/?misdaad

Kader 4.1. Gedetailleerde profilering op basis van sociale media

Persoonlijksheidsanalyse door 'likes'
Data-analyse maakt het mogelijk om op basis van 'likes' op Facebook uitspraken en voorspellingen te doen onder andere over iemands seksuele voorkeur, religieuze en politieke oriëntaties, persoonlijke karaktereigenschappen en gebruik van verslavende middelen én of hij of zij is opgegroeid in een gebroken gezin.

Een dataset van profielen van meer dan 58.000 personen in de Verenigde Staten die vrijwillig gedetailleerde demografische gegevens, de resultaten van verscheidene psychologische tests en hun 'likes' op Facebook aanleverden, vormde hiervoor de basis. Met speciale patroonherkenningstechnieken kan daardoor uit onder meer likes persoonlijke informatie worden vergaard. Zo kon het programma in 88% van de gevallen correct aangeven of iemand heteroseksueel of homoseksueel is, in 95% van de gevallen Afro-Amerikanen herkennen, en in 95% van de gevallen Amerikanen van Kaukasische origine. In 85% van de gevallen geeft het correct aan of iemand meer affiniteit heeft met de Republikeinen of juist de Democraten.

Bron: Kosinski, M., Stillwell D., & Graepel. T (2012) Private traits and attributes are predictable from digital records of human behavior In PNAS vol. 110 no. 15

Ethisch witwassen
Stiekem experiment van Facebook brengt aan het licht dat het tonen van positieve of juist negatieve informatie de stemming van facebookgebruiker beïnvloedt. Hij gaat hierdoor zelf ook positievere, of negatievere, posts plaatsen. Nieuws over dit experiment leidde tot grote ophef onder de facebookgebruikers.

Facebook voerde dit experiment in 2012 uit onder 680.000 Engelstalige facebookgebruikers, in samenwerking met onderzoekers van de Cornell University (Adam et al. 2014). Het bedrijf stelde de gebruikers hiervan echter niet op de hoogte en vroeg hun zodoende ook niet expliciet om toestemming. Ook nadat de studie was afgerond en de resultaten geanalyseerd waren, werd dit niet aan de 'proefpersonen' meegedeeld. Facebook voert hiervoor als reden aan dat gebruikers al toestemming hadden gegeven voor het experiment door akkoord te gaan met de algemene voorwaarden van het bedrijf.

De ophef onder de facebookgebruikers was groot. Hoewel het bedrijf vanuit juridisch oogpunt geen regels heeft overschreden, blijkt dat er

ondanks de impliciete toestemming toch de nodige vraagtekens geplaatst kunnen worden bij van de werkwijze van Facebook. Voor vergelijkbare experimenten bij universiteiten gelden striktere regels, zoals het onderzoekvoorstel voorleggen aan een speciale ethische commissie ter beoordeling, en expliciete toestemming van deelnemers aan het experiment. Deze studie was daar ook voorgelegd, en zonder veel discussie goedgekeurd, de onderzoekers van Cornell kregen namelijk zelf geen toegang tot de ruwe data; ze leverden alleen analysemethoden aan. In feite omzeilde Facebook door de gekozen constructie deze striktere ethische regels, hetgeen Jeroen van der Ham beschrijft als een vorm van ethisch witwassen.

Bronnen:

Ham, J. van der (2014). Ethisch witwassen, Blog Datadenkers Rathenau Instituut, https://datadenkers.wordpress.com/2014/11/27/ethisch-witwassen/#more-229

Adam, D. Kramer, I. Guillory, J., Hancock, J. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences of the United States of America, Vol 111, No 24. http://www.pnas.org/content/111/24/8788.full

4.1 Digitale innovatie en de Europese dataprotectie

Op Europees niveau wordt met een wijziging van de dataprotectiewetgeving getracht om de problemen van de digitale samenleving het hoofd te bieden. Het voorgestelde Europese kader borduurt voort op de bestaande dataprotectierichtlijn (95/46/EC) en voegt daaraan een aantal nieuwe plichten voor dataverwerkers en een aantal nieuwe rechten voor het individu toe. De inhoud van de nieuwe verordening is in grote lijnen vastgesteld.[16] Toch is er discussie over de vraag of de nieuwe voorstellen een adequaat kader schept waarin de ontwikkeling van big data op een verantwoorde manier tot bloei kan komen.

16 Op 15 juni 2015 is er overeenstemming bereikt tussen de Europese Raad van Ministers, het Europees Parlement en de Europese Commissie over de verordening: http://ec.europa.eu/justice/newsroom/data-protection/news/150615_en.htm

Kader 4.2 Veranderingen in dataprotectie

De juridische kaders voor privacy en dataprotectie worden in Europa onder meer gevormd door het Handvest van de Grondrechten van de Europese Unie, waarin zowel het recht op privacy (artikel 7, 'recht op eerbiediging van het privéleven, familie- en gezinsleven, zijn woning en zijn communicatie') als het recht op dataprotectie (artikel 8, 'bescherming van de hem betreffende persoonsgegevens' zijn vastgelegd (EG 2000).

Het grondrecht op privacy heeft een inhoudelijk karakter en beschermt tegen overmatige inmengingen in het privéleven, en beperkingen van de vrijheid en autonomie van individuen (Gutwirth & Gellert 2011). Het recht op dataprotectie is afzonderlijk vastgelegd in de richtlijn 95/46/EG (EC 1995) en wordt nu herzien. Dit meer procedurele recht bepaalt de spelregels waaronder persoonsgegevens verwerkt kunnen worden. Belangrijke principes uit deze wetgeving zijn onder andere:

- *Dataminimalisatie*: er mogen niet meer gegevens dan nodig worden verzameld.
- *Doelbinding*: gegevens mogen alleen voor een vooraf gespecificeerd doel worden verzameld
- *Subsidiariteit*: zijn er alternatieven mogelijk waarbij verwerking van persoonsgegevens niet nodig is, of waarbij minder gegevens nodig zijn?
- *Proportionaliteit*: staat het doel van de gegevensverzameling in relatie tot de middelen, met andere woorden tot de risico's die de gegevensverwerking voor individuen met zich meebrengt?
- *Beschermende maatregelen*: zoals eisen aan de kwaliteit, accuraatheid en beveiliging van gegevensverwerking.
- *Rechten van individuen waarvan persoonsgegevens worden verwerkt('datasubjecten')*: zoals toestemming, inzage- en correctiemogelijkheden.

Deze punten blijven het ook het uitgangspunt voor de voorgestelde verordening. Zie onder een beknopt overzicht van enkele belangrijke wijzigingen en doelen (zie EC 2015 en bijvoorbeeld Reding 2012):

- Het nieuwe kader is geen richtlijn, maar een verordening. Dat betekent dat de regels direct geldig zijn in alle lidstaten en niet eerst in nationale wetgeving geïmplementeerd hoeven te worden. Het doel is om meer uniformiteit en rechtszekerheid over databescherming in de gehele EU te creëren.

- Bereik: de regels moeten gaan gelden voor elke dataverwerker en datasubject in de EU, ook voor organisaties gevestigd buiten de EU, als zij gegevens verwerken van mensen gevestigd in de EU.
- Toezicht: het streven is om bedrijven en organisaties te maken te laten hebben met één dataprotectieautoriteit, in plaats van meerdere toezichthouders in verschillende landen waarin het bedrijf actief is, die verschillend kunnen oordelen. De nieuw op te richten Europese Dataprotection Board coördineert de activiteiten van de nationale toezichthouders. De toezichthouders kunnen bij niet naleving van de regels boetes opleggen van 1 miljoen euro of tot 2% van de wereldwijde omzet.
- Versterking van de rechten van het individu, bijvoorbeeld 1) via verduidelijking van het toestemmingsvereiste (het gaat om 'expliciete' toestemming, en datacontroller moet kunnen aantonen dat toestemming verkregen is), 2) introductie van 'right to erasure', waarbij individuen hun toestemming voor het verwerken van data kunnen intrekken en verzoeken om de data die over hen bekend is te verwijderen, en 3) recht op toegang en portabiliteit van data, waarbij individuen gegevens moeten kunnen verwijderen van de ene onlinedienst en kunnen verhuizen naar een andere onlinedienst.
- Versterking van verantwoordelijkheden en plichten van controllers: 1) bedrijven moeten *compliance* met de regelgeving kunnen aantonen, 2) datalekken moeten gemeld worden, 3) grote bedrijven of bedrijven die veel persoonsgegevens verwerken, moeten een onafhankelijke *data protection officer* aanstellen, 4) *privacy by design* en by default moeten ervoor zorgen dat de privacy in het ontwerpproces wordt meegenomen en privacyinstellingen standaard zo hoog mogelijk staan, 5) *data protection impact assessments* moeten worden uitgevoerd als dataverwerking risico's van datasubjecten met zich meebrengt.

Een van de interessante kenmerken van big data is dat het vooraf vaak niet duidelijk is welke nieuwe inzichten worden verworven (zie hoofdstuk 2). De nieuwe analysemogelijkheden en het alsmaar groeiende volume van data leveren samen immers nieuwe, vooraf onverwachte, verbanden op. Die kenmerken lijken lastig verenigbaar met centrale pijlers van het dataprotectieregime, met name doelbinding en het toestemmingsvereiste. Immers, als van te voren niet bekend is welke verbanden er zullen worden gevonden, zelfs niet naar welke verbanden precies gezocht gaat worden, dan is het onmogelijk om vooraf een specifiek doel voor de dataverzameling en analyse te formuleren. Het is dan dus ook niet mogelijk voor consumenten om vooraf betekenisvol toestemming te geven voor het gebruik van die specifieke data. Bovendien ligt een belangrijk deel van de waarde van big data niet besloten in het verzamelen en gebruiken

van data voor een specifiek doel, maar in het hergebruiken en nader analyseren van bestaande (persoons)gegevens. Een voorbeeld hiervan is Google Flu Trends, waarbij zoektermen van personen worden hergebruikt voor analyses over het uitbreken en het verloop van griepepidemieën. Dataprotectie bemoeilijkt in dit opzicht om voluit de vruchten te plukken van big data.

Van verzameling naar gebruik

Dit probleem wordt ook in Amerikaanse studies onderkend, ondanks de verschillen tussen het Amerikaanse en Europese privacyregime. Zo stellen de adviseurs van het Witte Huis vast dat de huidige manier van toestemming vragen in de VS niet meer werkt vanwege de kenmerken van big data: '*the notice and consent regime is defeated by exactly the positive benefits that big data enables: new, non-obvious, unexpectedly powerful uses of data.*' (Podesta et al. 2014). Zij trekken daaruit de conclusie dat er meer aandacht moet komen voor het daadwerkelijke *gebruik* van data, en minder voor de verzameling en de analyse van data. Daarmee komt volgens hen meer nadruk te liggen op verantwoordingplichten (*accountability*) door bedrijven en organisatie. De vraag zou niet meer moeten zijn 'Waar komen deze gegevens vandaan?', maar 'Is deze wijze van gegevensgebruik te verwantwoorden?' Een andere studie van het Witte Huis over de toekomst van big data en privacy stelt een vergelijkbare oplossingsrichting voor: een 'geen-verrassingenregel'. Volgens deze regel zouden gebruikers van te voren een redelijke verwachting moeten hebben waarvoor hun gegevens wel en niet gebruikt zullen worden (Podesta et al. 2014, p. 56).[17] Deze regel zou dan in de praktijk ondersteund kunnen worden met technische maatregelen. Bijvoorbeeld door persoonsgegevens te voorzien (te 'taggen') van informatie over de context waarin de gegevens verzameld zijn of toestemming voor een bepaald gebruik.

Ook onderzoekers van het Oxford Internet Institute en Microsoft stellen dat de huidige dataprotectiewetgeving innovatie met big data kan belemmeren: '*Avoid surpressing innovation with overly restrictive or inflexible data privacy laws.*' (Cate, Cullen & Mayer-Schönberger 2014). Omdat het echter niet zozeer om specifieke nationale regelgeving gaat, maar om onderliggende internationale afspraken van de OESO, pleiten zij ervoor om de OESO-privacyrichtlijnen te herzien.[18] Inhoudelijk strookt hun boodschap echter met die van de adviseurs van Obama: leg minder nadruk op reguleren van dataverzameling, maar regel vooral het gebruik van data (Cate, Cullen & Mayer-Schönberger 2014). Gegevensverwerkers moeten meer verantwoordelijk gemaakt worden voor het gebruik van de data.

17 Vergelijkbaar met het respect voor contextprincipe dat al in het Amerikaanse privacyrecht is opgenomen (Consumer Privacy Bill of Rights) (Podesta et al. 2014, p. 56).

18 Die overigens in 2013 herzien zijn, waarbij de originele uitgangspunten intact zijn gebleven.

In Nederland constateert onder andere het CPB de spanning tussen het reguleren van privacy enerzijds en de ruimte om te innoveren (op basis van persoonsgegevens) met big data anderzijds (CPB 2014; Roosendaal, Van den Broek & Van Veenstra 2014). De nieuwe Europese dataprotectieverordening tracht datacontrollers meer verantwoordelijk te maken voor de naleving (aantoonbaar) van de regelgeving.

Toestemming

Ook over de effectiviteit van het toestemmingsvereiste wordt discussie gevoerd. Toestemming is een belangrijke pijler van het dataprotectieregime. Volgens critici wordt echter het doel –controle geven aan het individu over hoe zijn gegevens worden gebruikt – in de praktijk vaak niet gehaald. In de praktijk worden burgers en consumenten geconfronteerd met ellenlange privacyverklaringen die dienen als juridische disclaimer. Bijna niemand leest die verklaringen en bijna iedereen gaat zonder te lezen direct akkoord door 'akkoord' aan te vinken. Daadwerkelijk zicht op welke gegevens worden verzameld, voor welk doel en met welke mogelijke gevolgen wordt niet verkregen. En een serieus alternatief voor 'akkoord' aanvinken ontbreekt vaak. Het groeiend aantal big-datatoepassingen versterkt dat probleem.

Een mogelijke oplossing is het creëren van '*layered consent*'. Hierbij wordt op verschillende niveaus uitleg gegeven over de dataverzameling. Dat kan door de kernboodschap weer te geven met symbolen en een korte uitleg (voor meer toelichting en de juridische beschrijving kan worden doorgeklikt), door gebruik te maken van standaard *privacy policies* (zoals nu zijn opgenomen in de nieuwe Europese verordening) of door keurmerken te introduceren. Een andere manier om het toestemmingsprobleem op te lossen is door nieuwe rechten in de dataprotectieverordening op te nemen. Voorbeelden zijn het 'recht om vergeten te worden' en het recht op dataportabiliteit. Met die rechten wordt geprobeerd om de positie van het individu te versterken. Toch blijven er twijfels, gelet op de technische complexiteit van big data en de sterke incentives voor bedrijven en overheden om gegevens te verwerken, of geïnformeerde toestemming houdbaar blijft. Zo gelooft de Amerikaanse rechtwetenschapper Ira Rubinstein hier niet in. Zij stelt: '[…] *the informed choice model is broken beyond any regulatory repair, and the only way to reinvigorate it is by changing the relevant information markets.*' (Rubinstein 2012, p. 6).

Niet iedereen voelt er echter veel voor om de oplossing voor dit probleem te zoeken in het loslaten – c.q. omvormen –van doelbinding en toestemming. De Duitse en Canadese privacy-autoriteiten spreken in een gezamenlijk rapport de vrees uit dat daardoor bedrijven en organisaties (nog) meer macht zullen krijgen (Cavoukian, Dix & Emam 2014). Zonder het principe van doelbinding en toestemming zal het voor individuen en toezichthouders nog moeilijker worden om fouten of schade te herstellen, en om privacyschendingen aan te pakken. Volgens hen is de genoemde spanning goed binnen de huidige regelgevende

kaders op te lossen: de huidige wetgeving biedt datacontrollers ruimte en flexibiliteit voor innovatief hergebruik van gegevens via het begrip '*compatible use*'. Ook zij noemen de 'geen-verrassingenregel' als nadere invulling van het principe. De datajurist Mireille Hildebrandt pleit in haar recente boek (2015) tegen het loslaten van doelbinding. Zij meent dat het loslaten hiervan fundamentele onzekerheid genereert. Zij pleit voor het handhaven van doelbinding, omdat door dit principe

1. de datagekte en data-obesitas wordt beperkt;
2. de burger overzicht en voorzienbaarheid wordt geboden; en
3. de verantwoordelijkheid voor de doelstelling en de verwerking, en daarmee het verdienmodel, wordt bepaald.

4.2 De risico's van profilering

Big data vergroten de mogelijkheden van profilering. De schaal waarop gegevens worden verwerkt, is groter dan ooit. Steeds meer organisaties, bedrijven en overheden gebruiken big data om individuen te profileren, te classificeren en te categoriseren, voor steeds meer verschillende doeleinden. Denk aan verregaande personaliseringsmogelijkheden voor reclame, maar bijvoorbeeld ook aan mogelijkheden voor passend onderwijs en onderwijsbegeleiding. Andere toepassingen zijn risicoanalyses voor verzekeringen, maar ook het opsporen van belasting- of sociale verzekeringsfraude, of criminaliteit- en terrorismebestrijding.

Met de vergroting van het aantal mogelijke toepassingen van big data worden de risico's van profilering groter. Profielen raken niet alleen aan de privacy in strikte zin (gegevensbescherming), zelfs al bevatten de profielen niet per se kenmerken van afzonderlijke individuen of zijn individuele kenmerken niet als zodanig te herleiden. Zo gebruikt vallen dergelijke profielen niet onder de Wet bescherming persoonsgegevens. Op basis van hun 'profilering' kunnen burgers ongelijk behandeld worden, door een ongewenste of een onrechtvaardige uitsluiting van diensten en voorzieningen als werk, verzekeringen, kredieten, wonen, zorg of onderwijs (zie bijvoorbeeld kader 4.3). Deze risico's worden door adviseurs van het Witte Huis in het rapport Big data: seizing opportunities, preserving values als volgt beschreven: '*By serving up different kinds of information or different prices or services to different groups, big data have the potential to cause real harm to individuals, whether they are pursuing a job, purchasing a home, or simply searching for information.*' (Podesta et al. 2014, p. 8).

Kader 4.3 Advertenties en ongelijke behandeling?

Dat ons surfgedrag wordt gevolgd om ons advertenties voor te schotelen, is inmiddels algemeen bekend. Hoe dit precies gebeurt, is echter voor veel mensen volstrekt ondoorzichtig. Onderzoekers bij Carnegie Mellon University in de Verenigde Staten ontwikkelden daarom een tool, AdFisher genaamd. Deze software probeert ze te reconstrueren hoe de gepersonaliseerde advertenties van bedrijven zoals Google werken.

Tijdens een test kwam daarbij aan het licht dat een nep-profiel van een werkzoekende man meer advertenties voor hoge functies voorgelegd kreeg dan een vergelijkbare vrouwelijke tegenspeler. Toch slaagde de software er niet helemaal in om de verschillen te verklaren.

Onlineadvertentiesystemen blijken zeer complex te zijn. Google gebruikt bijvoorbeeld zijn eigen data om te bepalen wie welke advertentie te zien krijgt. Maar bedrijven kunnen ook zelf selectiecriteria opgeven en eigen data toevoegen voor specifiekere targeting. '*I think our findings suggest that there are parts of the ad ecosystem where kinds of discrimination are beginning to emerge and there is a lack of transparency*', aldus Anupam Datta een van de onderzoekers die aan AdFisher werkte.

Bron: http://www.technologyreview.com/news/539021/probing-the-dark-side-of-googles-ad-targeting-system/

Het juridische kader van dataprotectie biedt slechts ten dele bescherming tegen profilering. Als profielen bestaan uit gegevens die juridisch niet als persoonsgegevens gelden, is immers geen toestemming van het individu nodig. Daarnaast gaat het bij dergelijke profilering om groepsprofielen – een enkel individu dat zich terug trekt of geen toestemming geeft, houdt de opbouw en gebruik van het groepsprofiel als zodanig niet tegen (Vedder 1998). Individuen hebben steeds minder grip op de wijze waarop profielen tot stand komen. Ook hebben ze geen invloed op de manier waarop profielen worden gebruikt of op de invloed die profilering op hen heeft. Dat maakt het in de praktijk lastig om bezwaar te maken tegen profilering. Hun digitale autonomie staat op het spel (Munnichs & Kool 2014; Hildebrandt 2015).

De afhankelijkheid van burgers en consumenten van de analyses toont zich vooral wanneer er fouten sluipen in data en daarop gebaseerde (risico)profielen (Munnichs en Kool 2014). Deze fouten kunnen het gevolg zijn van een incorrecte invoer van gegevens, verouderde data, identiteitsdiefstal of een verkeerde match van gegevens. Als gevolg hiervan kan iemand ten onrechte als 'probleemkind', 'wanbetaler' of 'drugscrimineel' worden beschouwd en overeen-

komstig behandeld. De mogelijkheden van burgers en consumenten om zich tegen dit soort fouten te verweren schieten vaak tekort.

Om daadwerkelijk actie tegen misbruik van profilering te kunnen ondernemen, moeten burgers en bedrijven inzicht hebben in de wijze waarop ze geprofileerd worden. Hildebrandt (2011; 2015) pleit daarom bijvoorbeeld voor het toepassen *van transparency enhancing tools* (TET's) of 'profieltransparantie', die het mogelijk maken voor de gebruiker om zicht te krijgen op de profilering. Om effectief te zijn, dient de transparantieverplichting scherper te worden geformuleerd dan in de huidige Europese richtlijn het geval, én standaard in diensten te worden ingebouwd (Ibid). Transparantie moet actief gegeven worden, en niet pas als het individu daarom vraagt. Anders blijft de controle van individuen over profilering beperkt: immers, als je niet weet waar je naar moet vragen, kun je er ook niet naar vragen.

De risico's van ongelijke behandeling en ongewenste uitsluiting van diensten door big data worden deels ook door andere wetten dan privacywetgeving geregeld. Denk aan het verbod op discriminatie, of een belangrijk uitgangspunt dat iedereen recht heeft op zorg, huisvesting en onderwijs. De fijnmazigheid van big-data-analyses en profilering roept vragen op of deze algemene punten voldoende waarborgen bieden, of dat er op termijn regels zullen moeten komen over informatie die niet verzameld of gebruikt mogen worden in een big-data-analyse. Mag in een voorspellende data-analyse over kansen op school en benodigde onderwijsbegeleiding bijvoorbeeld gebruik worden gemaakt van gegevens waar de leerling geen controle over heeft (zoals geboorteplaats of postcode), maar die wel een voorspellende waarde, en 'andere' behandeling kunnen betekenen?

Ook het College bescherming persoonsgegevens (CBP 2014) waarschuwt voor de risico's van profilering. Het college roept organisaties daarom op om persoonsgegevens te anonimiseren. Maar ook anonimisering is niet zonder problemen. Geanonimiseerde persoonsgegevens zijn steeds vaker te de-anonimiseren (re-identificatie) door die te combineren met niet-persoonsgegevens (zie kader 4.4). Daardoor blijven de risico's voor het individu van de gegevensverwerking bestaan. Adviseurs van het Witte Huis zeggen hierover: '*While there are promising research efforts underway to obscure personally identifiable information within large datasets, far more advanced efforts are presently in use to re-identify seemingly "anonymous" data. Collective investment in the capability to fuse data is many times greater than investment in technologies that will enhance privacy.*' (Podesta et al. 2014).

Anderen pleiten voor meer mogelijkheden voor individuen voor toegang én gebruik van hun gegevens om de scheve balans tussen individuen enerzijds en bedrijven en overheden anderzijds te herstellen: een *share the wealth*-strategie (Tene & Polonetsky 2012). Bovendien zou deze toegang nieuwe innovaties en

producten mogelijk kunnen maken, gericht op het hergebruik van persoonlijke data door het individu (ibid).

Het ministerie van Economische Zaken erkent de potentie en de risico's van profilering door middel van big data. Het ministerie geeft aan dat een strikt juridische benadering niet voldoende zal zijn om vertrouwen te creëren tussen burgers, consumenten, overheden en bedrijfsleven (EZ 2014). Het kabinet zoekt de oplossing in de invulling van drie randvoorwaarden voor vertrouwen: meer controle van de burger over zijn eigen gegevens, meer transparantie en meer verantwoordelijkheid van bedrijven.

Kader 4.4 Open data en privacy

Amerikaans voorbeeld toont aan dat een goede versleuteling van privacygevoelige informatie in open data een hoge prioriteit moet hebben. In de praktijk blijkt dat niet makkelijk te realiseren.

In 2013 kwam Chris Whong erachter dat de data van alle taxiritten in New York City opvraagbaar waren door de Amerikaanse variant op de Wet openbaarheid bestuur. Hij ontving een usb-stick met daarop bijna twintig gigabyte aan data over de taxiritten in New York City in de afgelopen jaren. Van elke taxirit waren onder meer begin- en eindtijd, vertrekpunt en eindpunt, en het aantal passagiers geregistreerd. In een apart bestand waren de prijs en de fooi opgenomen. In de originele data waren deze data bovendien gekoppeld aan een specifieke taxi. In een poging die privacygevoelige informatie te verhullen maar toch nuttige data te leveren, werden die data versleuteld meegeleverd.

Ontcijferen
Chris Whong maakte mooie plaatjes van de data die hij kreeg: van populaire plekken waar taxi's rijden, de frequentie waarmee taxi's rijden, et cetera. Hij stelde de data online beschikbaar aan derden. Al snel bleek dat de versleutelde data vrij makkelijk te ontcijferen waren. Toen kon vrij snel worden achterhaald welke taxi waar actief was geweest en wat de chauffeur in een jaar verdiend had, of gevoeligere informatie, zoals afwijkende pauzes. Ook koppelde iemand foto's van beroemdheden die in New York waren gesignaleerd terwijl ze in of uit een taxi stapten aan de data van Chris Wong. Zo werd duidelijk dat de ene beroemdheid beduidend scheutiger met fooien was dan een andere.

Privacy
Het voorbeeld laat zien dat het belangrijk is om privacygevoelige data in open data goed te anonimiseren. In Nederland wordt medische of

statistische data bijvoorbeeld vaak gereduceerd tot de postcode, het geslacht en de geboortedatum. In de praktijk blijkt dat een groot deel van de mensen van wie slechts de postcode, geslacht en geboortedatum bekend zijn, toch uniek identificeerbaar te zijn. Zelfs als alleen de vier cijfers van de postcode en de geboortedatum worden gebruikt, kan nog steeds twee derde van de Nederlanders geïdentificeerd worden.

K-anonimiteit
De structuur van de brondataset moet daarom steeds goed onder de loep worden genomen. Er moet steeds een goede afweging gemaakt worden tussen wat er precies nodig is voor het analyseren van een dataset, en wat er wordt vrijgegeven. Een goed hulpmiddel daarbij is de zogenoemde k-anonimiteit. Die berekent nauwkeurig per set kenmerken tot hoeveel personen de data te reduceren zijn. Al wordt het in de praktijk steeds moeilijker om die afweging te maken, omdat er steeds meer verschillende datasets beschikbaar zijn.

Bron:
Ham, J. van der (2014) Taxi's en regenbogen. Rathenau blog 'Datadenkers'. https://datadenkers.wordpress.com/2014/11/20/taxis-en-regenbogen/#more-220

4.3 Discussiepunten

In het discours over big data en privacy staan twee discussiepunten centraal. Het eerste discussiepunt betreft de logica van big data en big-data-innovatie. Hoe verhoudt die 'logica' (alles verzamelen en later bekijken hoe en waarvoor het gebruikt kan worden) zich tot de bestaande regels van dataprotectie? Met name de eisen van doelbinding en toestemming lijken daarmee op gespannen voet te staan. Ook is het een discussiepunt of de nieuwe privacywetgeving die in Europa in de maak is (EC 2015), voldoende aansluit op de uitdagingen waarvoor big data ons stellen? Of hinderen striktere dataprotectieregels juist innovatie met big data?

Het tweede discussiepunt betreft de sociale gevolgen van profilering. Bij profilering worden de gegevens over individuen, al dan niet op geaggregeerd niveau, verwerkt tot profielen. De impact van deze profielen gaat verder dan strikte privacy-impact. Op basis van profiel worden besluiten genomen over de behandeling van specifieke individuen. Ze krijgen specifieke producten, diensten of prijzen al dan niet aangeboden op basis van hun profiel. Als het om essentiële verschillen in behandeling gaat, kan dat discriminerend zijn of kan dat leiden tot ongerechtvaardigde uitsluiting. Omdat het voor individuen vaak niet duidelijk is dat gegevens over hen verzameld worden en hoe die bij hen 'terug'komen, is het voor hen moeilijk om zich hiertegen te beschermen. Is het

mogelijk en wenselijk om door transparantie-eisen consumenten en burgers in staat te stellen zich daartegen te weren? Is het mogelijk en wenselijk om door regelgeving te voorkomen dat er ongewenste uitsluiting van diensten plaatsvindt of dat individuen ongerechtvaardigd ongelijk worden behandeld?

5 De grenzen van en keuzes in big data

Big data kunnen niet alleen worden gezien als een technologische ontwikkeling, maar ook als een nieuw sociaal-economisch paradigma, waarin het verwerken van grote hoeveelheden gegevens om tot inzichten en beslissingen te komen, centraal staat (Hey, Tansley & Tolle 2009). Deze visie legt de nadruk op het meetbaar maken en analyseren van gegevens, gebaseerd op correlaties in plaats van oorzakelijke verbanden. Mayer-Schonberger en Cukier (2013) betitelen deze drang om elk fenomeen in kwantificeerbare gegevens te vangen als *dataficatie*. *New York Times*-journalist David Brooks beschrijft dit 'data-ism' als de opkomende hedendaagse filosofie (zie het citaat in de inleiding).

Achter de fixatie op data gaat een bepaald geloof in de objectiviteit en verklarende kracht van gegevens schuil, wat Boyd en Crawford (2011) de mythe van big data noemen. Big data beloven nieuwe inzichten en de grote hoeveelheden maken een accurate manier om fenomenen objectief in kaart te brengen, mogelijk. Dat is precies de aantrekkingskracht van big data: *'bedrijven willen data-driven, fact-based solutions'*.[19] Verschillende critici beargumenteren dat het belangrijk is om kritisch naar dit soort ideeën over objectiviteit te kijken (Boyd & Crawford 2011; Gitelman 2013; Gillespie 2013; WEF 2014). Hildebrandt (2015) verwacht dat binnen vier jaar de vraag is hoe we weer afkomen van al die data, vooral van onjuiste en irrelevante data. Hoewel een datagedreven benadering veel voordelen kan opleveren, is het ook van belang te weten waar de beperkingen liggen. Aan iedere dataverzameling, -analyse en –visualisatie liggen keuzes ten grondslag. Welke zijn dat? Ook kennis van de kwaliteit en representativiteit van de dataset is onontbeerlijk om de resultaten op waarde te kunnen schatten. En wanneer vormen correlaties een geschikte basis om uitkomsten of gedrag te voorspellen, en in welke gevallen niet?

5.1 Data is niet neutraal

Microsoft Researcher Kate Crawford waarschuwt voor de gevaren van data-fundamentalisme.[20] Daarmee doelt zij op het onbetwiste vertrouwen en het geloof in de objectiviteit van gegevens. Crawford en anderen (bv. Gitelman 2013; Gillespie 2013) laten zien dat het verzamelen, analyseren en visualiseren van gegevens geen neutraal proces is. Er kunnen fouten in sluipen, bewust bepaalde omissies worden gemaakt of als gevolg van gegevensverzameling kunnen bepaalde gegevens ontbreken. In elke stap van dataverzameling, tot aan datavisualisatie, worden keuzes en interpretaties gemaakt die het eind-

19 Bijeenkomst Data Science Center, 08-04-2014
20 http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/

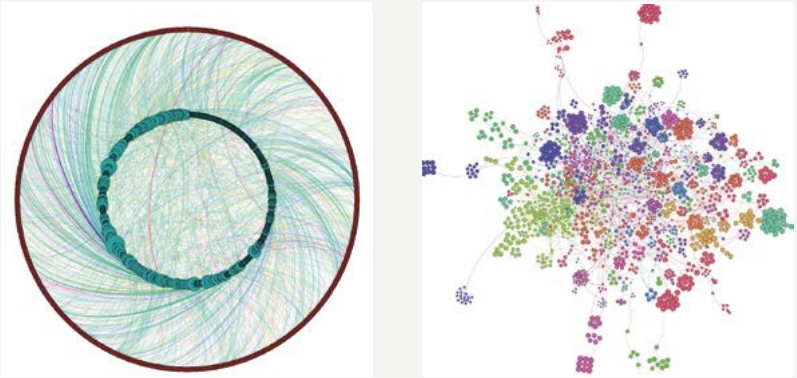
resultaat beïnvloeden – net als ieder statistisch onderzoek. Maar de gemaakte keuzes zijn aan het eind van het dataproces niet altijd meer zichtbaar, zeker voor de manager of beleidsmaker die zijn beslissing op de data moet baseren. In plaats van een mythisch vertrouwen toe te kennen aan dergelijke analyses, blijft het dus van essentieel belang voor, enerzijds data-analisten om zich bewust te zijn van die keuzes, en daarover openheid te blijven geven, en anderzijds voor managers en beleidsmakers, om vragen over die te blijven stellen over de gemaakte keuzes.

Kader 5.1 De schone schijn van data-onderzoek

In opdracht van het Rathenau Instituut legden onderzoekers van de Utrecht Data School de keuzes in een typisch data-analyse bloot. Ze deden dit aan de aan hand van een netwerkanalyse van de connecties van 'bestuurlijk Nederland'. Het oorspronkelijk doel was om Twitter voor de analyse te gebruiken. Maar Nederlandse CEO's bleken amper actief op Twitter. Als alternatief zijn bronnen van de Volkskrant, Management Scope, IEX.nl en Quote gebruikt. Elk van deze bronnen gebruiken andere indicatoren en andere wegingen om de 'machtigste' of meest 'invloedrijke' personen te bepalen. Sommigen leggen meer nadruk op verband met de politiek (zoals de Volkskrant), anderen meer op de financiële middelen van 'machtige' personen (zoals Management Scope). De IEX toont bijvoorbeeld de grootste bedrijven in termen van koerswaarde, maar wat voor invloed deze personen elders uitoefenen is niet zichtbaar.

De onderzoekers voegden bovengenoemde bronnen samen in een centrale dataset, waarbij de dubbelingen (veroorzaakt door verschillende benamingen van personen, bedrijven en functies in iedere dataset) zijn verwijderd. De onderzoekers gebruikten verschillende software-algoritmes om de connecties van bestuurlijk Nederland weer te geven. Ze ontdekten dat sommige algoritmes meer rekening houden met hoe de visualisatie eruit ziet (de esthetische waarde), dan de eigenschappen van de dataset zelf. Personen werden bijvoorbeeld netjes gepositioneerd in een cirkel, maar die plaats correspondeerde niet altijd precies met de waarde in de dataset.

Wie de verschillende mogelijkheden benut die visualisatietools bieden, kan met dezelfde dataset tot zeer verschillende presentaties van dezelfde data komen. De onderzoekers ontworpen aan de hand van de casus twee uiteenlopende visualisaties. De een benadrukt dat de betrokken personen en organisaties nauw met elkaar verweven zijn (de 'inner circle'); bij de ander ligt de nadruk op de verschillende clusters in Nederland, en duidt op meer spreiding van de macht ('ons kent ons').



Bron:
Bouwman, X., D. van Geenen en M. van der Goes (2013). *De schone schijn van data-onderzoek. Het proces van netwerkanalyse onder de loep genomen.* Utrecht Data School #2.

Voor de Amerikaanse stad Boston werd bijvoorbeeld de app StreetBump ontwikkeld, die op basis van gegevens uit sensoren in smartphones gaten en slechte plekken in het wegdek detecteert en doorgeeft aan de wegbeheerder die het onderhoud verzorgt. De app is een mooi voorbeeld van hoe data op een slimme manier ingezet kunnen worden om dienstverlening in de publieke sector te verbeteren, maar hij is ook tekenend voor de problemen rondom representativiteit. Omdat de data afkomstig zijn van smartphones (specifieker iPhones), komen er alleen gegevens binnen van een bepaalde subgroep van de populatie. Ouderen en mensen met lagere inkomens die geen dure smartphone bezitten of de StreetBump app niet installeren, vallen buiten het plaatje, met als gevolg dat armere of vergrijsde buurten systematisch minder onderhoud aan het wegdek zouden kunnen krijgen. Doordat de ontwikkelaars van de app zich van deze bias bewust werden, is uiteindelijk besloten de app in te zetten bij weginspecteurs die elk deel van het wegennet evenveel bezoeken (Podesta et al. 2014).

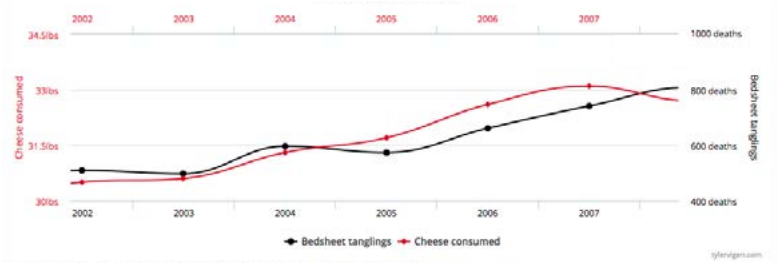
Een ander voorbeeld komt naar voren op een bijeenkomst van het Data Science Center van de TU Eindhoven.[21] Een instelling als het CBS besteedt jaarlijks veel geld aan onderzoek naar consumentenvertrouwen. Gebruikmaken van bestaande (big) data zou een snelle, kostenbesparende manier kunnen bieden om het consumentenvertrouwen te meten. De OESO (2013b, p.17) onderschrijft de potentie van big data voor statistische bureaus. Twitterberichten kunnen bijvoorbeeld worden ingezet als indicator voor consumentenvertrouwen. Voor een bureau als het CBS ligt hier echter ook een probleem, omdat het onduide-

21 DSC/e, 08-04-2014

lijk is in hoeverre deze big data bijvoorbeeld uit twitterberichten daadwerkelijk een betrouwbare afspiegeling zijn van het consumentenvertrouwen. Zolang het verband tussen de twee variabelen niet duidelijk is, kan er niet op Twitter worden vertrouwd als graadmeter voor consumentenvertrouwen. De verkeerde voorspelling van Google Flu Trends illustreert dit punt: het algoritme dat het aantal griepgevallen voorspelde aan de hand van zoektermen in de zoekmachine, deed een aantal jaren achtereen een goede voorspelling, maar kwam in 2013 met een grove overschatting (Butler 2013).[22] Een instelling die betrouwbare statistiek moet produceren als grond voor beleid en beslissingen, kan zich een dergelijke onnauwkeurigheid niet veroorloven. Voor 'vertrouwd' statistisch onderzoek met steekproeven hebben we een theoretisch kader om de foutmarge in te schatten, maar is dat ook mogelijk voor big data? Hoe zou dat eruit moeten komen te zien?

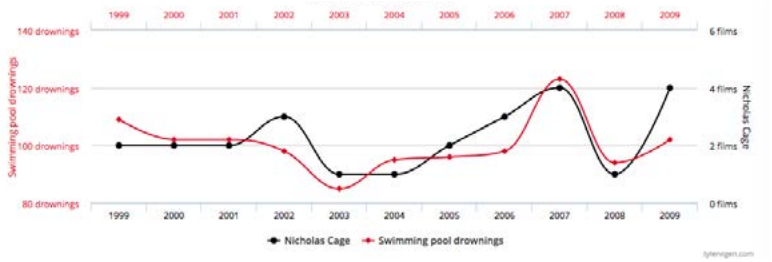
Statistici uiten tot slot ook zorgen over het feit dat in steeds grotere bergen gegevens de kans ook groter wordt dat niet-bestaande verbanden gevonden worden. *'Big data may mean more information, but it also means more false information.'* (Taleb 2013). De verbanden die met data worden gevonden, zeggen bovendien niets over de oorzaak van een fenomeen, zoals Tyler Vigen op zijn website Spurious correlations probeert duidelijk te maken met grafieken. Figuur 1 toont een ogenschijnlijk verband tussen hoeveel kaas mensen eten en hoeveel mensen overlijden doordat ze komen vast te zitten in hun dekbed. Figuur 2 toont een ogenschijnlijk verband tussen films waarin Nicolas Cage speelt en het aantal mensen dat verdrinkt in een zwembad.

Figuur 1 Verband tussen kaasconsumptie en vastzitten in dekbed



22 De verklaring is dat zoektermen een relatief zwak signaal zijn voor griepgevallen. Een gebruiker kan op griep of gerelateerde termen zoeken om een hoop verschillende redenen. Wanneer een andere externe factor (bv. media-aandacht voor een onderwerp) ervoor zorgt dat er meer op griep gerelateerde termen wordt gezocht, klopt de voorspelling niet meer.

Figuur 2 Verband tussen film met Nicolas Cage in de hoofdrol en kans op overlijden door val in zwembad



5.2 Complexiteit, transparantie, toezicht en controle

Inzicht in de manier waarop gegevens worden verzameld en geanalyseerd, is een belangrijke voorwaarde om een inschatting te kunnen maken van de waarde van een analyse. *'Perhaps the most dangerous is the technologist who never understands the limitations of data, never understands what data isn't telling you.' (Loukides 2013).* Dit inzicht wordt echter bemoeilijkt door de steeds complexere aard van dataverzamelingen en analyses. Eerder werd al beschreven dat het moeilijk is om zicht te houden op datastromen en op hoe data worden uitgewisseld en gebruikt (WEF 2014). Ook de manieren waarop data wordt geanalyseerd met softwaretools en algoritmen nemen toe in complexiteit. Tal Zarsky (2013), een Israëlische datamining-expert, wijst op het gebrek aan transparantie van voorspellende systemen:

> *Data mining might point to individuals and events, indicating elevated risk, without telling us why they were selected. Here, the software makes its selection decisions based upon multiple variables (even thousands) […] It would be difficult for the government to provide a detailed response when asked why an individual was singled out to receive differentiated treatment by an automated recommendation system. The most the government could say is that this is what the algorithm found based on previous cases.*

Volgens hem kan transparantie worden ingebouwd, bijvoorbeeld via auditlogs[23], maar moet hierbij een afweging gemaakt worden tussen transparantie en efficiëntie, omdat auditlogs het systeem kunnen vertragen (Zarsky 2015).[24]

23 Een audit log is een chronologische vastlegging van alle gebeurtenissen en handelingen, en tijdstip daarvan, waarmee acties van gebruikers, en de software, kan worden nagelezen.

24 Interview Tal Zarksy, 2014

Het controleren van algoritmes kan ook moeilijk zijn, omdat de manier waarop bedrijven gegevens verwerken onderdeel van een beschermde bedrijfsstrategie zijn. Tarleton Gillespie (2013) beschrijft hoe Twitter gebruikmaakt van een algoritme om het Trending Topic te bepalen. Het algoritme is niet inzichtelijk voor gebruikers en externe partijen, waardoor het ondoorzichtig is waarom iets als 'populair' onderwerp is geselecteerd. De website kan echter moeilijk inzicht geven in haar werkwijze, omdat ze daarmee misbruik in de hand zou werken en competitief voordeel zou verliezen.

> *Revealing the "secret sauce" of their algorithm in greater detail risks helping those who would game the system. Everyone from spammers to marketers to activists to 4chan tricksters to narcissists might want to optimize their tweets and hashtags so as to Trend.* (Gillespie 2013)

Het punt over hoe controle te houden op algoritmes wordt belangrijker, naarmate algoritmes meer automatische beslissingen gaan nemen, bijvoorbeeld bij de uitvoering van overheidsbeleid (ook wel '*algoritmic regulation*' genoemd; regulering door algoritmes). Dit gebeurt bijvoorbeeld al op de beurs, waarin algoritmes zelf aandelen kopen en verkopen. Dit heeft een paar keer tot grote dalingen op de beurs geleid, omdat algoritmes op elkaar reageerden en binnen fracties van seconden aandelen bleven verkopen (zie kader 5.2.).

Ook voor het bepalen van kredietscores en het krijgen van leningen is het belangrijk om te kunnen controleren hoe het systeem de score bepaalt en of dat rechtvaardig is (zie kader 5.3.), of dat het een juiste beslissing is. Zulke systemen kunnen grote effecten hebben op het leven van mensen. Software mist soms belangrijke contextinformatie om mee te wegen bij een beslissing. Vorig jaar berichtte de *New York Times* bijvoorbeeld over een man die net van baan gewisseld was en zijn hypotheek niet opnieuw kon financieren. Zijn nieuwe baan met onzekere inkomsten vormde een te groot risico. Dat klinkt logisch, maar de man was Ben Bernanke, voormalig voorzitter van de Amerikaanse centrale bank, die net een miljoenencontract had afgesloten voor een boek en zeer gevraagd werd om lezingen te geven (Irwin 2014).

Een ander voorbeeld is het reguleren van onlinecontent waarbij software automatisch bepaalt welke content wel en welke content niet geplaatst kan worden (bijvoorbeeld discriminerende teksten of hatelijke teksten). De Russische internetcriticus Yevgeni Morozov stelde in 2014 dat de software van Silicon Valley daarmee de samenleving een nieuw soort conservativiteit oplegt, waarbij algoritmes bepalen wat cultureel wel of niet is toegestaan (Morozov 2014). Hij vindt daarom dat er gebruik moet worden gemaakt van externe '*algorithmic auditors*' om dit te controleren, maar hij gaat niet in op wie die inspectie zou moeten doen, of hoe dat georganiseerd moet worden.

Hoewel het belang van transparantie vanuit verschillende hoeken wordt onderschreven (zie Zarksy 2013, p. 1506) als voorwaarde om controle en

toezicht te kunnen houden op data-analyses en voorspellende software – zowel door het publiek als door overheden of bedrijven – blijkt het geven van betekenisvolle transparantie nog een grote uitdaging in de ontwikkeling van big data.

Kader 5.2 Automatische beurshandel

Flash Crash

Op 6 mei 2010 ontstond er een crash op de Amerikaanse beurzen, waarbij in half uur zeer veel geld verloren ging. Gedurende de dag herstelden de beurzen weer. Verschillende instanties deden onderzoek naar de oorzaak. De crash lijkt te zijn veroorzaakt door een combinatie van factoren, waaronder de geautomatiseerde handelssystemen (ook *high frequency trading* genoemd) die in enkele seconden grote hoeveelheden aandelen aan elkaar verkochten. Op een gegeven moment werden er meer dan 27.000 contracten in slechts veertien seconden verhandeld in de vicieuze cirkel die de systemen hadden gecreëerd. In vier minuten daalde de koers met 3 procent. Er was ook een wachtrij van orders in een cruciale server ontstaan, waardoor de ICT-systemen niet meer de werkelijke koersen konden weergeven. In 2012 ontstond er een flash crash op de Indiase beurzen door een stroom foutieve orders van een beurshandelaar en het niet op tijd leggen van de handel.

Om dergelijke verliezen in de toekomst te voorkomen, zijn in veel beurzen limieten ingesteld bij zeer plotselinge koersdalingen en –stijgingen en kan de beurs dan worden stilgelegd. Ook zijn er regels en procedures voor 'software-bugs' gekomen. De software faalde bijvoorbeeld in 2012, toen het bedrijf 'Knight Capital Group' 440 miljoen dollar verloor in 30 minuten (4 keer het netto inkomen van het bedrijf). De handelssoftware van de het bedrijf kochten aandelen hoog, en verkocht ze voor een lage proces – het tegenovergestelde dus van goede handelsstrategie. Ook tijdens de beursgang van Facebook in 2012 waren er problemen. Nasdaq werd bekritiseerd voor slecht ontworpen software, die leidde tot een overbelast systeem, en het verkeerd verwerken van handelsorders.

Bronnen:

Schoemaker, R. (2010) Concurrerende algoritmes oorzaak beurscrash. Webwereld 4 oktober 2010 http://webwereld.nl/netwerken/45293-concurrerende-algoritmes-oorzaak-beurscrash

Bloomberg (2012) Nasdaq Chief Blames Software for Delayed Facebook Debut, 22 mei 2012. http://www.bloomberg.com/news/articles/2012-05-20/nasdaq-ceo-says-poor-design-in-ipo-software-delayed-facebook

Houtman, J. (2012) Beurs India beleeft 'flash crash'. Financieel Dagblad, 5 oktober 2012. http://fd.nl/frontpage/beleggen/847082/beurs-india-beleeft-flash-crash

Kader 5.3 Kredietscore als black box

Citron en Pasquale (2014) waarschuwen voor de gevaren van datasystemen als black boxes aan de hand van kredietbeoordelingssystemen in de VS. Bij het afsluiten van een lening maken Amerikaanse kredietverstrekkers vaak gebruik van een bureau dat de kredietwaardigheid bepaalt op basis van gegevens over eerder financieel gedrag. Hoe deze scores tot stand komen, is onduidelijk, omdat het om complexe systemen en berekeningen gaat die bovendien meestal bedrijfsgeheim zijn. Het systeem is dus een black box. Voor klanten betekent dat dat ze niet weten welke variabelen invloed hebben op hun kredietwaardigheid en waarom hun in een bepaald geval precies een lening geweigerd wordt. Het is ook niet inzichtelijk als het systeem een 'denkfout' maakt of zich op verkeerde data berust, en het is dus moeilijk om daartegen bezwaar te maken. Hoewel de kredietscore een subjectieve inschatting is van iemands kredietwaardigheid, wordt er een onwrikbare waarde aan gehecht stellen.

Bron:
Timmer, J. et al. (2015) Berekende risico's: Verzekeren in de data samenleving.

5.3 Discussiepunten

Big data zijn een nieuw instrument om inzichten en kennis te produceren. De mogelijkheden en de beperkingen van dit instrument moeten we – deels – nog leren kennen. Hoe kan worden voorkomen dat door onjuiste interpretatie van gegevens en correlaties verkeerde beslissingen over mensen worden genomen? Welke check en balances zijn nodig? Bij wie liggen verantwoordelijkheden? En welke (nieuwe) vaardigheden zijn hiervoor nodig? Hoe kunnen managers en beleidsmakers zicht blijven houden op de gemaakte keuzes in de big data-analyse? Hoe zorgen we voor transparantie om controle en toezicht op data-analyses mogelijk te maken? En hoe kan hierin rekening worden gehouden met bescherming van bedrijfsgevoelige informatie?

6 Competenties en vaardigheden

6.1 Datavaardig

Het verantwoord omgaan met big data vraagt om expertise bij degenen die met big data werken of op basis van data-analyses hun beslissingen moeten maken. Een goede datascientist moet zich bewust zijn van de beperkingen van zijn gegevens en analyses. Dat vergt dat hij meer moet zijn dan alleen een goede computerwetenschapper of mathematicus, maar ook oog moet hebben voor de sociale context van een vraagstuk of bepaalde dataset (Eric Meyer, Oxford Internet Institute, 03-10-2013). Het Data Science Center van de Technische Universiteit Eindhoven profileert de datascientist dan ook als iemand met een brede kennisbasis; iemand met kennis op het gebied van computerscience en wiskunde, maar ook op het gebied van sociale wetenschappen, privacy en ethiek (DSC/e, 08-04-2014).

Bedrijven signaleren een tekort aan mensen met deze brede expertise. McKinsey (2011) schat dat er in de VS in 2018 zo'n 140.000 tot 190.000 specialistische datascientists nodig zullen zijn. Het beroep datascientist wordt uitgeroepen tot '*sexiest job of the 21st century*' door Harvard Business Review. Ook de OESO (2013b) onderschrijft het belang van een goede kennisbasis voor de ontwikkeling van de data-economie. Het tekort aan kennis en goed opgeleide mensen wordt gezien als belemmering voor de innovatie.

Er is niet alleen specialistische kennis nodig. Naarmate data-driven processen en werkwijze een grotere rol gaan spelen, komen ook meer managers en beleidsmakers hier direct mee in aanraking. Dat vraag ook van hen bepaalde datavaardigheden om een verantwoorde inschatting te kunnen maken over hoe data-analyses worden ingezet, en welke beslissingen hierop kunnen worden genomen. Naast de 140.000 tot 190.000 datascientists zullen er volgens McKinsey in de VS in 2018 ook zo'n 1,5 miljoen managers nodig zijn die de vaardigheden bezitten om big data goed in te zetten. In een strategische innovatie-agenda van het Europese Big Data Value Partnership wordt het zorgen voor voldoende competenties een centrale uitdaging genoemd. Er moeten nieuwe curricula worden opgezet aan universiteiten, maar ook gedeelde onderzoeksprojecten en *innovation spaces* om de samenwerking tussen industrie en universiteiten te bevorderen (Big Data Value cPPP 2014, p. 26)

Een andere benadering is om de tools om big data mee te verwerken, toegankelijker te maken. Door aansluiting te zoeken met bekende tools als Excel probeert Microsoft bijvoorbeeld het voor eindgebruikers gemakkelijker te maken om met big data aan de slag te gaan. Er wordt ingezet op 'democratisering' van big data tools (Microsoft 2011).

6.2 Discussiepunten

De roep om expertise op het gebied van big data is sterk. Wat moet er gebeuren om deze behoefte te voorzien? Hoe is de kennisbasis in Nederland? Waar liggen de kansen voor de Nederlandse kenniseconomie? Hoe kan ervoor gezorgd worden dat niet alleen specialisten maar ook managers en beleidsmakers die te maken hebben met big data de juiste vaardigheden bezitten?

7 Veranderende machtsverhoudingen

Met de opkomst van een datagedreven economie ontstaan nieuwe verdienmodellen, waarin het verwerken van gegevens centraal staan. Deze modellen zijn een bron van innovatie en economische groei (OESO 2013b). Denk bijvoorbeeld aan de snelle opkomst van online platformen als Uber en Airbnb, die optimaal gebruik maken van data en slimme software. Ismail et al. (2014) noemen deze platformen 'exponentiele organisaties'. De platformen kunnen zeer disruptief voor bestaande organisaties. Zo zet Uber de bestaande taxibranche op zijn kop met zijn op data en slimme software gebaseerde bedrijfsmodel. Het (mobiele) internet zorgt voor een eenvoudige en goed toegankelijke verbinding tussen vraag en aanbod van taxi's, en software zorgt voor dynamische prijzen die zich aanpassen aan drukte of luwte.[25] Ondertussen lijkt er in allerlei sectoren een 'Uber' mogelijk, denk bijvoorbeeld aan Helpling voor schoonmakers, of Carenzorgt.nl voor mensen die zorg nodig hebben, of kunnen bieden. Ook in de verzekeringssector is te zien hoe in een extreem scenario andere partijen dan verzekeraars – namelijk internetbedrijven die over data beschikken – een rol kunnen gaan spelen in het aanbieden van verzekeringen (Timmer et al. 2015).

De grote waarde van data – en van software om die waarde uit data te halen – zorgt er ook voor dat verhoudingen tussen consumenten, bedrijven en overheden veranderen. De vraag wie de data bezit en ervan kan profiteren, staat centraal in hoe verhoudingen in de opkomende data-economie en datasamenleving vorm krijgen. Op het niveau van de individuele consument of burger is er discussie over wie de eigenaar is, of zou moeten zijn, van de gegevens die een individu produceert, en over hoe het individu zich moet verhouden tot bedrijven en overheden die steeds meer over hem weten (zie paragraaf over profilering en digitale autonomie). Voor bedrijven en overheden speelt de vraag welke partijen de data bezitten en wie daar in economisch opzicht het meeste voordeel uit haalt. In deze paragraaf kijken we naar veranderende machtsverhoudingen die ontstaan door big data en datagedreven verdienmodellen. Hoe beïnvloeden zij de economische kansen voor Europa?

7.1 Positie van bedrijven en Europa in het big-data-ecosysteem

Het data-ecosysteem wordt op dit moment gedomineerd door Amerikaanse partijen, zoals Amazon, IBM, Google, Oracle en Microsoft. Daarmee bestaat de vrees dat Europa niet volledig zal kunnen profiteren van de economische groei als gevolg van de 'big-datarevolutie' (European Big Data Value Partnership

25 Zie bijvoorbeeld Surowiecki, J. (2014) In Praise of Efficient Price Gouging, MIT Technology Review, 19 augustus 2014, http://www.technologyreview.com/review/529961/in-praise-of-efficient-price-gouging/

2014). Europa wil daarom meer inzetten op 'leiderschap in platformen voor de digitale industrie'.[26]

> *"Europe is belatedly discovering its failure to develop many of the platforms underpinning the online economy. Much of the world's digital territory has in effect been ceded to America without a fight."* (The Economist 2015)

> *A great challenge is also Europe's position in the development of the next digital platforms that will gradually replace the current Internet and mobile platforms. We have so far missed many opportunities in this field and our online businesses are today dependent on a few non-EU players world-wide: this must not be the case again in the future.* Mr Oettinger [Commissaris Europese Commissie voor digitale economie en samenleving] (EC 2015).

Volgens de Europese Commissie zijn de barrières voor de EU divers: het ontbreekt aan vergelijkbare industriële capaciteiten als de VS, de financiering en innovatie op het gebied van data is subkritisch en vaak ongecoördineerd. Er zijn te weinig data-experts die ondernemingskansen benutten. Verder is de Europese markt gefragmenteerd door verschillende talen en juridische kaders. Ook heeft het midden- en kleinbedrijf moeilijkheden om de markt te betreden (EC 2014). Daarom investeert de Europese Commissie in het bundelen van bestaande publiek-private initiatieven in onderzoek naar nieuwe technologieën en diensten, wordt een Europees netwerk van kenniscentra opgezet om het aantal dataspecialisten te vergroten, en wordt open data en open standaarden gestimuleerd evenals het beschikbaar stellen van infrastructuur voor onderzoek (ibid). Ook de verdere eenwording van de Euroepse markt (denk aan de wetgeving over intellectueel eigendom en de Dienstenrichtlijn) vermindert fragmentatie en kan een stimulerende rol spelen voor het opzetten van nieuwe *data driven* bedrijven.[27] Daarvoor is ook het aantrekken van kapitaal en goed personeel, en een aantrekkelijke ondernemingscultuur van belang.

Er bestaan ook zorgen over de mogelijke gevaren van monopolisering in het data-ecosysteem. De dominante positie van enkele spelers zoals Google en Amazon kan leiden tot onwenselijke centralisatie van macht (Mayer Schonberger & Cukier 2013b). In zijn boek *Who Owns the Future* beschrijft internetgoeroe Jaron Lanier dat enkele belangrijke grote IT-bedrijven een centrale positie in het data-ecosysteem veroveren, waarmee ze controle en overzicht hebben op een groot aantal van de datastromen en diensten die over het netwerk verlopen

---

26 Oettingers speech at Hannover Messe, Europe's future is digital, 14 april 2015, https://ec.europa.eu/commission/2014-2019/oettinger/announcements/speech-hannover-messe-europes-future-digital_en

27 Interview Bart van Ark in Est, van R. & L. Kool (2015)

(Lanier 2013; Kreijveld, Deuten & Van Est 2014). De dominantie van enkele grote spelers kan leiden tot een concentratie van gegevens en waarde bij deze partijen, die het moeilijker maakt voor nieuwe partijen om tot de markt toe te treden. Het tegengaan van monopolisering zou bovendien moeilijker zijn op het gebied van data, omdat het hier in tegenstelling tot sectoren zoals software en producten moeilijker is om de continue veranderende grootte van de markt te schatten en de positie van een bepaalde speler daarin te bepalen (Mayer-Schonberger & Cukier 2013b). Europa klaagde in april 2015 Google aan voor machtsmisbruik.[28] De Europese Commissie onderzoekt of Google in zijn zoekmachine eigen diensten voortrekt en daarmee concurrenten benadeelt.

Ondanks de dominante positie van Amerikaanse partijen in het big data-speelveld, worden er ook kansen gesignaleerd voor Europese partijen. Bijvoorbeeld op het gebied van aanbieden van veilige cloud-diensten onder de Europese Dataprotectie regulering ('privacy als kans'). Bedrijven zoals het Finse F-secure, Deutsche Telekom en Orange bieden clouddiensten aan vanaf Europese bodem, met privacy en databescherming als 'selling point'. In Nederland heeft KPN CloudNL gelanceerd als clouddienst van eigen bodem. Met technische en juridische maatregelen probeert het bedrijf klanten zekerheid te bieden, bijvoorbeeld dat gegevens niet op verzoek aan de Amerikaanse overheid zullen worden geleverd. Toch blijven garanties daarvoor moeilijk te geven (Koenis 2014).

Vanuit de EU wordt er door middel van Europese Cloud Strategie als onderdeel van de Digitale Agenda stimulans gegeven aan ontwikkelingen rondom cloud computing. Oprichter Robert Knapp van het bedrijf CyberGhost (VPN-diensten) ziet een duidelijk voordeel voor Europa op dit gebied, al geeft hij wel aan dat er betere randvoorwaarden moeten komen:

> *'The European Union is overall a good place for privacy-related projects, companies (…) For the first time in history we have an advantage towards American companies… Cyber security and the United States of American is simply a no-go. It doesn't fit together. It's not working. So we are able to build maybe some unicorns here in Europe. But for that we need an ecosystem.'*

Andere sterke kanten van Europa liggen op het vlak van data-analyse en het bouwen van toepassingen op de bestaande laag van diensten:

> *Large US IT and Internet companies currently have an unquestionable lead on Big Data infrastructure & storage techniques [...] The field of Big Analytics & Data Visualization (predictive and decision support systems) is much more open. The EU has an undeniable competitive advantage*

28 http://europa.eu/rapid/press-release_IP-15-4780_en.htm

> *here, thanks to the very high mathematical and computer literacy level of EU engineers and research scientists as well as the solid base of industries which own most of the underlying data assets, unlike the end consumer data sets dominated by consumer facing web companies in the US.* (European Big Data Value cPPP - Strategic Research and Innovation Agenda - April 2014, p. 14)

Volgens Minister Kamp van Economische Zaken is Nederland goed toegerust om te kunnen profiteren van de potentie van big data (EZ 2014): "De telecominfrastructuur in Nederland – een van de basisvoorwaarden voor het transport van data – is van wereldklasse. Nederland kent goedlopende bedrijven die actief zijn op gebied van big data en er vindt veel en hoogstaand ICT- en wetenschappelijk onderzoek plaats."

7.2 Discussiepunten

Vooral Amerikaanse spelers zijn sterk vertegenwoordigd in de datagedreven economie. De uitdaging voor Europa, en Nederland, is om de nieuwe mogelijkheden van datagedreven bedrijfsmodellen te ontdekken en in te zetten. Daarbij verdient de verdere eenwording van de Europese (digitale) markt aandacht, als ook het aantrekken van kapitaal en personeel en het creëren van een aantrekkelijke ondernemingscultuur de aandacht. Ook het nieuwe Europese kader voor gegevensbescherming kan een innovatiekans voor Nederland en Europa betekenen voor de ontwikkeling van privacyvriendelijke diensten, bijvoorbeeld gericht op empowerment van de consument om met zijn eigen data aan de slag te gaan.

8 Conclusie: maatschappelijke uitdagingen

In deze achtergrondstudie is een beeld geschetst van een aantal centrale discussies rondom big data, gestart vanuit de vraag: Wat is er nodig om de ontwikkeling van big data te kunnen benutten, en hoe kan ervoor worden gezorgd dat dit op een maatschappelijk verantwoorde manier gebeurt? De studie laat zien dat hierbij verschillende spanningen ontstaan. We vatten de belangrijkste uitdagingen hier samen.

8.1 Naar een realistische kijk op big data

Een deel van de waarde van big data bestaat uit het kunnen combineren van verschillende databronnen, binnen organisaties en tussen organisaties. Maar in de praktijk blijkt het daadwerkelijk delen van data om verschillende redenen heel lastig: ICT-systemen zijn op verschillende manieren geoptimaliseerd in silo's of in 'oude' ICT-systemen, garanties over de kwaliteit en beveiliging van data zijn lang niet altijd aanwezig, en ook de incentives om data te delen ontbreken vaak. Onze studie over big data in de verzekeringspraktijk laat zien dat de verwachtingen over wat er mogelijk is ook vooruitlopen op wat er daadwerkelijk gebeurt (Timmer et al. 2015). Ondanks de torenhoge verwachtingen is het toepassen big data in de praktijk dus niet vanzelfsprekend.

De ontwikkeling van data stewardship, datamarkten en licentiemodellen kan daarom helpen om data makkelijker te delen. Ze zorgen namelijk niet alleen voor een makkelijke toegang, ze maken ook vergelijkbaarheid en kwaliteitsgaranties mogelijk, en ze kunnen er daarnaast voor zorgen dat de data worden opgeschoond en in de juiste formats worden aangeleverd. Partijen die data bezitten, kunnen die tegen vergoeding op datamarkten aanbieden In sommige sectoren wordt al gewerkt met dergelijke licentiemodellen. De Nationale Databank Wegverkeersgegevens bijvoorbeeld stelt data zowel 'open' en gratis beschikbaar, als onder licentie en tegen betaling, met een onderscheid in serviceniveau. Andere partijen en sectoren kunnen daarvan leren. Partijen, op nationaal en internationaal niveau, zullen daar samen afspraken over moeten maken, en ruimte moeten creëren om daarmee in de praktijk te experimenteren.

8.2 Creëren van datavaardigheid

De verwachtingen of percepties over big data omgeven zijn dus omgeven door een zekere mythevorming. Er bestaat een groot geloof in big data als nieuwe, objectieve kennisbron. Maar big data zijn niet neutraal. Ook big data blijven een statistische analyse, die zijn eigen beperkingen kent. Elke dataverzameling, -analyse en visualisatie gaat gepaard met keuzes over welke databronnen wel of niet mee worden genomen, hoe de dataset 'gereed' wordt gemaakt voor

analyse, en welk algoritme wordt gebruikt voor de analyse of visualisatie. Dat is niet altijd zichtbaar bij het eindresultaat.

Het is daarom belangrijk dat iedereen, inclusief beslissers en beleidsmakers, kritisch blijven kijken naar wat een big-data-analyse vertelt. Wanneer zijn correlaties voldoende om beleid op te baseren, en wanneer niet? Over welke sociale context vertelt de analyse iets? Steeds vaker blijken big data eerder een manier om nieuwe vragen te stellen en te onderzoeken, dan dat ze het definitieve antwoord brengen. Big data brengen de hypothese, er is vervolgonderzoek nodig om die hypothese te valideren.

De verantwoorde inzet van big data vraagt daarom om datavaardigheid bij managers, beleidsmakers én burgers om de mogelijkheden en beperkingen van data-analyes en datagedreven omgevingen te leren kennen. Welke keuzes zijn gemaakt? Hoe beïnvloeden die keuzes het eindresultaat en de waarde die er aan toe te kennen valt?

8.3 Heldere afspraken over gegevensgebruik

Er bestaat veel discussie over hoe de 'logica' van big data (alles verzamelen, later kijken of en hoe het gebruikt kan worden) zich verhoudt tot het dataprotectierecht. Met name de wettelijke eisen aangaande doelbinding en toestemming vragen voor het verzamelen en gebruiken van gegevens lijken op gespannen voet te staan met big data. Een van de interessante kenmerken van big data is dat het vooraf vaak niet duidelijk is welke nieuwe inzichten worden gevonden. Dat maakt het moeilijk om vooraf een doel te formuleren of vooraf betekenisvol toestemmingen te geven of te vragen voor het gebruik van data. Op dit moment wordt het Europese juridische kader voor persoonsgegevens aangepast, mede om nieuwe technologische ontwikkelingen als big data mogelijk te maken. Dat betekent dat doelbinding, dataminimalisatie, transparantie, toestemming en accountability ook de komende jaren de juridische uitgangspunten zijn uitgangspunten zijn waaraan elke partij die big data toepast of wil toepassen, zich ten minste zal moeten houden.

Daarnaast blijkt dat wat past binnen de juridische kaders, niet altijd aansluit bij de verwachtingen en percepties van burgers en consumenten. Denk aan de ophef over de voorstellen van ING en Equens over het vermarkten van gegevens, of TomTom over het geaggregeerd delen van rijgedrag met de politie. Voor het vertrouwen van burgers en consumenten in 'data-gebruikers' is transparantie van deze organisaties over gegevensgebruik, én daadwerkelijk een zekere mate van controle over dat datagebruik, zeer belangrijk. Dat vraagt van organisaties niet alleen dat ze nadenken over hoe zij nieuwe waarde kunnen halen uit gegevens van klanten of burgers, maar ook hoe diezelfde klanten of burgers toegang hebben tot hun data. Sommige verzekeraars experimenteren bijvoorbeeld met een 'cafetariamodel', waarbij klanten zelf aangeven waarvoor

data gebruikt mogen worden en waarvan ze hun toestemming ook kunnen intrekken (Timmer et al. 2015).

8.4 Het belang van autonomie en gelijke behandeling

Tegelijk is duidelijk dat de discussie over big data en privacy verdergaat dan het beschermen van persoonsgegevens. Het gebruik van big data en profilering gaat ook over het waarborgen van autonomie, vrije keuze en gelijke behandeling. Op basis van profielen, al dan niet op geaggregeerd niveau, worden besluiten genomen over de behandeling van specifieke individuen. Dat kan gaan om prijs of toegang tot een dienst of product. Als er essentiële verschillen in behandeling ontstaan, kan dat discriminerend zijn of leiden tot ongerechtvaardigde uitsluiting. Omdat het voor individuen vaak niet duidelijk is dat gegevens over hen verzameld worden, en hoe die bij hen 'terug'komen, is het voor hen moeilijk om zich hiertegen te beschermen. Met het steeds slimmer worden van omgevingen, en de opkomst van een datagedreven samenleving waarin beslissingen steeds meer leunen op data-analyses, wordt het geven van transparantie over meer dan alleen de sec verzamelde gegevens steeds belangrijker. Individuen moeten informatie krijgen over hoe een profiel wordt gebruikt, en hoe dat beslissingen over hen beïnvloedt. Op dit moment schieten de mogelijkheden van consumenten en burgers tekort om zich te verweren tegen foutieve beslissingen en de consequenties daarvan (zie bijvoorbeeld Munnichs et al 2009; 2014). Er dient dus meer aandacht te komen voor het versterken van de positie van burgers en consumenten in de praktijk. Soms kan dat heel praktisch: geen schriftelijk bezwaarverzoek indienen, maar direct online.[29]

8.5 Grip houden op automatische (software)beslissingen

Data-analyses worden steeds complexer, en gaan een steeds grotere rol spelen bij het nemen van beslissingen, bijvoorbeeld op de beurs, bij het verstrekken van krediet of bij het opsporen van fraude. Dat gaat niet altijd goed, en het is lastig te achterhalen waar of waarom het fout ging, ook voor toezichthouders. Soms is de inzet van software erg complex, waardoor er weinig grip meer is op wat de systemen doen. Dat laat het voorbeeld van de *flash crash* van mei 2010 zien. De beurs noteerde in korte tijd een enorm verlies en herstelde in de loop van de dag weer. Daarop volgde langdurig onderzoek naar wat er precies gebeurd was en wat de rol was van algoritmen die in miniseconden in aandelen handelen. Transparantie kan in de softwaresystemen worden ingebouwd, bijvoorbeeld via auditlogs, maar vraagt om een afweging tussen transparantie, efficiëntie en snelheid. Soms is het controleren van algoritmes moeilijk, doordat ze onderdeel zijn van een beschermde bedrijfsstrategie.

29 Hierbij dient natuurlijk rekening gehouden te worden met minder digivaardige mensen; een offline mogelijkheid moet ook mogelijk blijven.

Er moet meer aandacht komen voor hoe controle en toezicht op complexe systemen geregeld kan worden om systeemdwang te voorkomen. Dat gaat ook over de bovengenoemde positie van burgers en consumenten. De ervaring met meer 'traditionele' ICT-systemen leert dat ICT-systemen in de praktijk nooit foutloos zijn en dat eenmaal gemaakte fouten zeer hardnekkig zijn, en moeilijk te corrigeren zijn. De gevolgen voor burgers of consumenten kunnen zeer schrijnend zijn (Kamerstukken II 2014-2015; Munnichs et al. 2009; Ombudsman 2013). De inzet van slimme software voor grote data-analyses maakt het dus noodzakelijk om werkbare correctieprocedures te organiseren.

8.6 Experimenteren met data-driven modellen

Datagedreven bedrijfsmodellen worden steeds belangrijker. Met het vergaren van data en die effectief kunnen inzetten in bedrijven ontstaat een nieuwe ondernemingscultuur, waarbij data het middelpunt vormen. Dat kan zeer disruptief zijn voor bestaande bedrijven of sectoren. Zo is in de verzekeringssector te zien hoe data nieuwe verzekeringsmodellen voor bestaande partijen mogelijk maken, maar ook hoe in een extreem (disruptief) scenario hele andere partijen dan verzekeraars – internetbedrijven die over data beschikken – een rol kunnen gaan spelen in de verzekeringswereld (Timmer et al. 2015). Het ontdekken van de nieuwe mogelijkheden van data vraagt om ruimte voor experiment, mét ruimte voor reflectie en evaluatie en inachtneming van publieke waarden als privacy, vrije keuze of solidariteit. Het nieuwe Europese kader voor gegevensbescherming kan ook een innovatiekans voor Nederland en Europa betekenen voor het ontwikkelen van diensten, gericht op empowerment van de consument om met zijn eigen data aan de slag te gaan.

8.7 Slotbeschouwing

De datagedreven samenleving is volop in ontwikkeling: data worden een steeds belangrijker onderdeel van bedrijfsmodellen en organisatieprocessen, zowel in het bedrijfsleven als in de publieke sector. Diverse voorbeelden laten zien hoe krachtig het gebruik van data kan zijn. Het is dus niet verwonderlijk dat organisaties zoeken naar hoe ze de slag naar 'data-driven' kunnen maken. Deze studie laat zien dat ze daarbij tegen verschillende hindernissen aan lopen: hoe kunnen data gedeeld worden, welke competenties en vaardigheden zijn daarvoor nodig, hoe borgen we privacy en digitale autonomie, hoe houden we grip op automatische beslissingen en hoe bereiden we ons voor op mogelijke disruptieve 'data-driven' organisatiemodellen. Ook wordt duidelijk dat in elk domein of sector verschillende dynamieken spelen. Gegevensuitwisseling in de zorg kent geheel andere technische en organisatorische uitdagingen, juridische kaders en ethische risico's dan bijvoorbeeld gegevensuitwisseling in de mobiliteitssector. Dat vraagt, naast de algemene uitdagingen die deze studie benoemt, ook om nader inzicht in de specifieke uitdagingen per domein.

Literatuur

Adam, D. I. Kramer, J. Guillory & J. Hancock (2014). 'Experimental evidence of massive-scale emotional contagion through social networks' In: *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 111, No 24. http://www.pnas.org/content/111/24/8788.full

AFM (2010) *High frequency trading: De toepassing van geavanceerde handelstechnologie op de Europese markt*. 18 november 2010. toepassing-hft-op-europese-markten-afm.pdf

Boyd, D. & K. Crawford (2011). '*Six Provocations for Big Data. A Decade in Internet Time'*, Symposium on the Dynamics of the Internet and Society, September 2011.

Brooks, D. (2013). '*The Philosophy of Data*'. New York Times' http://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html?_r=0

Butler, D. (2013). 'When Google got flu wrong' In: Nature 494 (7436), pp. 155-156. http://www.nature.com/news/when-google-got-flu-wrong-1.12413

Cate, F., P. Cullen & V. Mayer-Schönberger (2014). '*Data Protection Principles for the 21st century: Revising the 1980 OESO guidelines*'.

Cavoukian, A., A. Dix, K. Emam (2014). *The unintended consequences of privacy paternalism*. Privacy and Information Commissioner Canada: Toronto https://www.privacybydesign.ca/index.php/paper/unintended-consequences-privacy-paternalism/

CPB (2014). *Kiezen voor privacy. Hoe de markt voor persoonsgegevens beter kan*. Den Haag: Centraal Planbureau.

Citron, D.K. & F. Pasquale (2014). 'The Scored Society: Due Process for Automated Predictions'. *In: Washington Law Review, 89, no. 1.*

Dumbill, E. (2012). *Planning for big data: A CIO's Handbook to the Changing Data Landscape*. Sebastopol (Canada): O'Reilly Publishing.

Dwork. C. & D. Mulligan (2013). 'It's not privacy, and it's not fair' In: *Stanford Law Review* 35, http://www.stanfordlawreview.org/online/privacy-and-big-data/its-not-privacy-and-its-not-fair

Esmeijer, J., T. Bakker, S. de Munck (2013). *Thriving and surviving in a data-driven society*. TNO Rapport.

European Big Data Value Partnership (2014). '*Strategic Research and Innnovation Agenda*'. http://www.bdva.eu/sites/default/files/europeanbigdata-valuepartnership_sria__v1_0_final.pdf

Europese Commissie (1995). Richtlijn van het Europees Parlement en de Europese Raad. '*Over de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens*' (95/46/EC). Brussel, 24 oktober 1995.

Europese Commissie (2010). '*A comprehensive approach on personal data protection in the European Union COM*' (201) 609 Final.

Europese Commissie (2014). '*Towards a thriving data-driven economy*', Communication of the European Commission, 442 final. Brussels. https://ec.europa.eu/digital-agenda/en/towards-thriving-data-driven-economy

Europese Commissie (2015) '*Europe's future is digital*'. Mr. Oettinger's speech at Hannover Messe, 14 april 2015, https://ec.europa.eu/commission/2014-2019/oettinger/announcements/speech-hannover-messe-europes-future-digital_en

Europese Commissie (2015) *Proposal on new data protection rules to boost EU Digital Single Market* supported by Justice Ministers, 15 juni 2015, http://data.consilium.europa.eu/doc/document/ST-9565-2015-INIT/en/pdf

Europese Gemeenschap (2000). *Handvest van de Grondrechten van de Europese Unie*, (2000/C 364/01) www.europarl.europa.eu/charter/pdf/text_nl.pdf

Europees Parlement (2014) *Security of e-government systems*. Final Report. Report for STOA by ETAG Consortium 2014

Evans, P.C. & M. Annunziata (2012). *Industrial Internet. Pushing the Boundaries of Minds and Machines*. General Electric.

Gartner (2011). '*Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data*', persbericht, www.gartner.com/it/page.jsp?id=1731916

Gillespie, T. (2013). '*The Relevance of Algorithms*.' In: Gillespie, T. et al. (2013) *Media Technologies*. Cambridge, MA: MIT Press.

Gitelman, L. (Red.) (2013). *Raw Data is an Oxymoron*. MIT Press: Cambridge MA

Gillespie, T. (2012) Can An Algorithm be Wrong. *LIMN: Crowds and Clouds*. http://limn.it/can-an-algorithm-be-wrong/

Greenwald, G. (2013). 'Revealed: how US and UK spy agencies defeat internet privacy and security' In: *The Guardian*, 06-09-2013.

Gutwirth, S. and R. Gellert (2011) Privacy en dataprotectie: sterk verweven maar toch verschillend. pp 47-68-In: Frissen, V., Kool, L, Van Lieshout (2011). *Jaarboek ICT en Samenleving 2011: De transparante samenleving*. Media Update: Gorredijk.

Ham, J. van der (2014). '*Ethisch witwassen´*, Blog Datadenkers. Den Haag: Rathenau Instituut. https://datadenkers.wordpress.com/2014/11/27/ethisch-witwassen/#more-229

Hey, T., S. Tansley, K. Tolle (2009). *The Fourth Paradigm: Data-intensive scientific discovery*. Microsoft Research.

Hildebrandt, M. (2011) *Privacy na de 'computationele wending'?* In: Frissen, V. et al. (2011). Jaarboek ICT en Samenleving 2011: De transparante samenleving. Media Update: Gorredijk.

Hildebrandt, M. (2015). *Smart technologies and the (ends of) Law*. Edward Elgar Publishing: Cheltenham.

IBM (2012a). '*Demystifying Big Data: A Practical Guide To Transforming The Business of Government*'. Washington: TechAmerica Foundation..

IBM (2012b). '*Big Data Comes of Age'*. EMA inc. and 9sight Consulting Research Report.

IBM (2013). '*The Four V's of Big Data*' http://www.ibmbigdatahub.com/infographic/four-vs-big-data

IDC (2014). '*Digital Universe of opportunities: Rich data and the increasing value of the Internet of Things*'. IDC Digital Universe study http://www.emc.com/leadership/digital-universe/2014iview/digital-universe-of-opportunities-vernon-turner.htm

IDC (2013) *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east – United States*. IDC Digital Universe study

Irwin, N. (2014). 'Why Ben Bernanke Can't Refinance His Mortgage' In: *New York Times*, 2 oktober 2014. http://www.nytimes.com/2014/10/03/upshot/why-ben-bernanke-cant-refinance-his-mortgage.html?_r=0&abt=0002&abg=1

Kamerstukken II 2012-13, 32761, nr. 49. Verwerking en bescherming persoonsgegevens. Kabinet's visie e-privacy. Brief van de Minister van Economische Zaken, 24 mei 2013. https://zoek.officielebekendmakingen.nl/kst-32761-49.html

Kamerstukken II 2013-14, 26 643, nr. 298. Informatie- en communicatietechnologie. Notitie Vrijheid en veiligheid in de digitale samenleving. Een agenda voor de toekomst. Brief van de minister en staatssecretaris van Veiligheid en Justie en van de minister van Binnenlandse Zaken en Koninkrijksrelaties

Kamerstukken II 2014-2015, 33 326, nr 5. Parlementair onderzoek naar ICT-projecten bij de overheid. Eindrapport.

Koenis, C. (2014) *De pijnpunten van KPN's volledig Nederlandse cloud*. In: Computerworld 23 mei 2014

Kosinski, M., Stillwell D., & Graepel. T (2012) Private traits and attributes are predictable from digital records of human behavior In: *Proceedings of the National Academy of Sciences of the United States of America* vol. 110 no. 15

Kreijveld, M., J. Deuten, R. van Est (2014). *De kracht van platformen*. Den Haag: Rathenau Instituut.

Kroes, N. (2013). *'Big Data for Europe. Speech - Big data for Europe'* European Commission - SPEECH/13/893 07/11/2013 http://europa.eu/rapid/press-release_SPEECH-13-893_en.htm

Lanier, J. (2013). *Who owns the future*. Simon & Schuster: New York.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). *Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season*. SSRN, 13 maart 2014. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2408560

Loukides, M. (2013). '*Big data is dead, long live big data: Thoughts heading to Strata*' http://radar.oreilly.com/2013/02/big-data-hype-and-longevity.html

Mayer-Schonberger, V. & K. Cukier (2013a). *Big Data: A Revolution that Will Transform How we Live, Work and Think*. Houghton Mifflin Harcourt.

Mayer-Schonberger, V. & K. Cukier (2013b). *The Rise of Big Data*. Foreign Affairs. May/June 2013.

McKinsey (2011). *Big Data: The Next Frontier for Innovation, Competition and Productivity*.

Microsoft (2011). '*Microsoft Big Data Solution Sheet*' Accessed online: http://download.microsoft.com/download/1/8/B/18BE3550-D04C-4B3F-9310-F8BC1B62D397/MicrosoftBigDataSolutionSheet.pdf

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2013). '*Actieplan Open overheid*' September 2013. http://www.rijksoverheid.nl/onderwerpen/digitale-overheid/documenten-en-publicaties/rapporten/2013/09/01/actieplan-open-overheid.html

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2015). '*Resultaten inventarisatie open data*'. 10 juli 2015.

Ministerie van Economische Zaken (2014). Big data en privacy. 19 november 2014.

Morozov, Y. (2014). 'The rise of data and the death of politics' In: *The Guardian*, 20 juli 2014, http://www.theguardian.com/technology/2014/jul/20/rise-of-data-death-of-politics-evgeny-morozov-algorithmic-regulation

Munnichs, G., Besters M. en Schuijff (2009). *Databases. Over ICT-beloftes, informatiehonger en digitale autonomie*. Rathenau Instituut: Den Haag.

Munnichs, G. & Kool, L. (2014). *De autonome burger in de informatiesamenleving. Hand-out inwerkprogramma tijdelijke commissie ICT-projecten bij de overheid*. Rathenau Instituut: Den Haag.

NESSI (2012). *Big Data: A New World of Opportunities*. White Paper.

OESO (2008). *OESO Recommendation for Enhanced Access and More Effective Use of Public Sector Information.*

OESO (2013a). 'OESO Privacy Framework' http://www.OESO.org/sti/ieconomy/privacy.htm

OESO (2013b). 'Exploring Data Driven Innovation as a New Source of Growth: Mapping the Issues Raised by "Big Data".'

Ombudsman (2013). Mijn onbegrijpelijke overheid. Jaarverslag over 2012.

Orange (2014). '*The Future of Digital Trust*'.

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

Reding, V. (2012). 'The European Data Protection Framework for the Twenty-First Century' In: *International Data Privacy Law* (2) , http://idpl.oxfordjournals.org/content/early/2012/06/25/idpl.ips015.full.pdf+html

Roosendaal, A. & T. van den Broek & A.F van Veenstra (2014). 'Vertrouwen in big data toepassingen: accountability en eigenaarschap als waarborgen voor privacy' In: *Privacy en Informatie*, vol. 2014, nr. 3.

Roosendaal, A. & Kool, L. (2012). 'Perspectieven op de waarde van persoonsgegevens' In: C. Prins, A. Vedder & F. Van der Zee (2012). *De transformatieve kracht van ICT. Jaarboek ICT en samenleving 2012*. Gorredijk: Media Update.

Rubinstein, I. (2012). 'Big Data: End of privacy of new beginning?' In: *International Data Privacy Law,* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2157659

Taleb N. (2013). 'Beware of the Big Errors' In: *Big Data WIRED*. http://www.wired.com/2013/02/big-data-means-big-errors-people/

Tene, O. & J. Polonetksy (2013) *A Theory of Creepy: Technology, Privacy and Shifting Social Norms*. Yale Journal of Law & Technology.

The Economist (2010). 'Data, data everywhere' In: *The Economist*, 25 februari 2010. http://www.economist.com/node/15557443

The Economist (2015). 'Nothing to stand on' In: *The Economist*, 18 april 2015, http://www.economist.com/news/business-and-finance/21648606-google?frsc=dg|a

Timmer J., I. Elias, L. Kool & R. van Est (2015). *Berekende risico's: verzekeren in de data-samenleving*. Den Haag: Rathenau Instituut.

Van Est, R. & L. Kool (2015) *Werken aan de robotsamenleving. Visies en inzichten uit de wetenschap over de relatie technologie en werkgelegenheid*. Den Haag: Rathenau Instituut.

Vedder (1998). 'Het einde van de individualiteit? Datamining, groepsprofilering en de vermeerdering van brute pech en dom geluk' In: *Privacy en Informatie* 1(3), pp. 115-120.

WEF (2012). 'Personal data: emergence of a new asset class'. World Economic Forum.

WEF (2014). 'Rethinking Personal Data: A New Lens for Strengthening Trust'. World Economic Forum.

Zarksy (2013). *'Transparent Predictions'*, SSRN. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2324240

Bijlage 1: Stakeholders

In de periode 2014 en 2015 heeft het Rathenau Instituut bijeenkomsten over big data bijgewoond van Platform voor de informatiesamenleving (ECP), Data Science Center TU/E, Big data wetenschapsjournalisten, Nationale DenkTank Big Data, Science Café Big Data TU/e, ministerie van Economische Zaken).

In deze periode heeft het projectteam gesprekken gevoerd met vertegenwoordigers van het Ministerie van Economische Zaken, Microsoft, Bits of Freedom, TU Eindhoven, Oxford Internet Institute, Cloud66, Alex Pentland, Tal Zarsky, Jennifer Lynch en het MIT Media Lab.

Op 9 april 2014 heeft het Rathenau Instituut bijgedragen aan een technische briefing voor de Tweede Kamer over big data.

Voorafgaand aan deze publicatie heeft het Rathenau Instituut de Utrecht Data School een studie laten uitvoeren naar de manier waarop gekozen algoritmes en gemaakte keuzes bij het verzamelen, analyseren en visualiseren van gegevens bepalen welk verhaal een dataset vertelt. Zie ook: https://datadenkers.wordpress.com/2014/02/24/de-schone-schijn-van-datavisualisaties/

Via een blog rondom het thema big data heeft het Rathenau Instituut ingespeeld op actuele ontwikkelingen en discussies over big-datatoepassingen. Verschillende externe partijen hebben ook aan het blog bijgedragen, zoals TNO, Consortium for Science, Policy & Outcomes CSPO, de Correspondent, Studyflow, Ministerie van Binnenlandse Zaken, Bits of Freedom, Universiteit Siegen, Universiteit Utrecht en de Utrecht Data School.

Wie was Rathenau?
Het Rathenau Instituut is genoemd naar professor dr. G.W. Rathenau (1911-1989). Rathenau was achtereenvolgens hoogleraar experimentele natuurkunde in Amsterdam, directeur van het natuurkundig laboratorium van Philips in Eindhoven en lid van de Wetenschappelijke Raad voor het Regeringsbeleid. Hij kreeg landelijke bekendheid als voorzitter van de commissie die in 1978 de maatschappelijke gevolgen van de opkomst van micro-elektronica moest onderzoeken. Een van de aanbevelingen in het rapport was de wens te komen tot een systematische besturdering van de maatschappelijke betekenis van technologie. De activiteiten van Rathenau hebben ertoe bijgedragen dat in 1986 de Nederlandse Organisatie voor Technologisch Aspectenonderzoek (NOTA) werd opgericht. NOTA is op 2 juni 1994 omgedoopt in Rathenau Instituut.

De datagedreven samenleving is volop in ontwikkeling: data en slimme software worden een steeds belangrijker onderdeel van bedrijfsmodellen en organisatieprocessen. Diverse voorbeelden laten zien hoe krachtig het gebruik van data kan zijn. Denk aan dijken die zelf aangeven wanneer onderhoud nodig is met behulp van sensoren, of aan de inzet van slimme camera's om de veiligheid te verbeteren. De verwachtingen van big data om maatschappelijke problemen op te lossen zijn hoog. Tegelijkertijd lopen organisaties in de praktijk tegen diverse hindernissen aan, zoals bij het delen of combineren van data, garanties bieden over kwaliteit en beveiliging, het borgen van privacy en grip houden op automatische beslissingen.

Deze achtergrondstudie is bedoeld voor beleidsmakers en beslissers die meer willen weten over wat big data nu eigenlijk zijn, en welke maatschappelijke en economische kwesties samenhangen met het gebruik van big data. De studie brengt de stand van zaken van het debat over de verantwoorde inzet van big data in kaart.

Met de *Datagedreven samenleving* wil het Rathenau Instituut een bijdrage leveren aan de verdere gedachtevorming over verantwoord datagebruik. Dat krijgt echter pas echt vorm in de praktijk. De studie is een startpunt voor een gezamenlijke dialoog tussen overheden, bedrijven en het maatschappelijk middenveld om nader invulling te geven aan de voorwaarden voor een verantwoorde datasamenleving.

ISBN 978-90-77364-71-0