KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

# BIG DATA IN WETENSCHAPPELIJK ONDERZOEK MET GEGEVENS OVER PERSONEN

ADVIES



Koninklijke Nederlandse Akademie van Wetenschappen
Postbus 19121, 1000 GC Amsterdam
Telefoon + 31 20 551 0700
knaw@knaw.nl
www.knaw.nl

pdf beschikbaar op www.knaw.nl

Basisvormgeving: Edenspiekermann, Amsterdam
Engelse vertaling samenvatting: Balance, Maastricht
Opmaak: Ellen Bouma, Alkmaar

Illustratie omslag: Carolyn Ridsdale

ISBN 978-90-6984-724-5

Deze publicatie kan als volgt worden aangehaald: KNAW (2018). *Big data in wetenschappelijk onderzoek met gegevens over personen*, Amsterdam, KNAW.

# BIG DATA IN WETENSCHAPPELIJK ONDERZOEK MET GEGEVENS OVER PERSONEN

Koninklijke Nederlandse Akademie van Wetenschappen
April 2018

# VOORWOORD

Big data is het nieuwe goud, wordt vaak beweerd. Er zijn al veel toepassingen van big data in vrijwel alle sectoren van onze samenleving, en vermoedelijk volgen er nog veel meer. Ook voor het wetenschappelijk onderzoek bracht big data een ontwikkeling op gang voor het meten, opslaan en analyseren van data. De tsunami van big data waarmee onderzoekers inmiddels te maken hebben, biedt nieuwe kansen om wetenschappelijk onderzoek in een groot aantal toepassingsgebieden te versterken.

Het voorliggende rapport licht toe waar die kansen liggen. Wetenschapsgebieden die gebruikmaken van big data over personen kunnen hun wetenschappelijke vraagstellingen uitbreiden, verscherpen en verbreden. Het blijkt vaak ook nodig om nieuwe methoden en technieken te ontwikkelen. Anderzijds dagen die gegevensbronnen, methoden en technieken de onderzoekers uit zich te onderwerpen aan kritische interpretatie en kwalitatieve analyse. Het vraagt hen te experimenteren met digitale methoden, te leren werken met digitale bronnen en zich te verdiepen in de 'geheimen' van het algoritmisch en computationeel denken.

Dit alles moet uiteraard gebeuren binnen de toepasselijke (privacy)kaders en richtlijnen. Het rapport verschijnt precies op het goede moment, want op 25 mei 2018 wordt de Algemene verordening gegevensbescherming (AVG) in Nederland van kracht. De komende jaren zullen doorslaggevend zijn voor de manier waarop een aantal aspecten van de AVG nog nader worden uitgewerkt, met name wat betreft de bijzondere waarborgen die gelden voor de verwerking van persoonsgegevens in wetenschappelijk onderzoek. Die uitwerking zal grote invloed hebben op de wijze waarop beheerders van grote datasystemen omgaan met het beschikbaar stellen van gegevens over personen, en welke afweging ze maken bij het waarborgen van de privacy van die personen.

Om onderzoekers in staat te stellen de kansen van big data te benutten, zijn ambitie en investeringen in mensen en middelen nodig. Tegelijkertijd is waakzaamheid geboden voor de risico's van big data. Pas als aan al deze voorwaarden is voldaan, kan het Nederlandse big data-onderzoek Europees en internationaal onderscheidend worden en bijdragen aan het verkrijgen van een vooraanstaande positie in de wetenschap.

Het adviesrapport pleit dan ook voor goede samenwerking bínnen de wetenschappelijke gemeenschap en buiten die gemeenschap met partners uit het publieke en private domein. Daarbij is het essentieel dat een overkoepelende infrastructuur tot stand komt die de lokale infrastructuur voor het verzamelen, verwerken, delen en analyseren van (big) data met betrekking tot personen ondersteunt en aanvult. Hier is een duidelijke rol weggelegd voor het ministerie van OCW dat samen met het Nationaal Platform Open Science een dergelijke infrastructuur kan opzetten. Dit Platform, dat begin 2017 is opgericht en inmiddels wordt aangevoerd door Nationaal Coördinator Karel Luyben, richt zich op zowel *open access* als *open data* om wetenschap 'open' te maken en houden.

De nieuw op te zetten infrastructuur zal ook moeten passen in het kader van het Europese platform waartoe Nederland, Duitsland en Frankrijk het voortouw hebben genomen: het GO FAIR-initiatief. Maar het zijn vooral wetenschappelijk onderzoekers zelf die de kansen van big data de komende jaren zullen moeten benutten. Dit vergt de gezamenlijke inzet van veel denkkracht en daadkracht.

José van Dijck
President KNAW

# INHOUD

# SAMENVATTING

Wetenschappelijk onderzoekers kiezen en construeren big data, net zoals ze dat bij 'gewone' data doen. Big data en 'gewone' data verschillen in hoeveelheid en verscheidenheid. Dit rapport staat stil bij de vragen: Wat gebeurt er met de relatie en wisselwerking tussen data, kennis en toepassingen nu *'gewone' data* zich ontwikkelen naar *big data?* Wat zijn de gevolgen voor de wetenschap, met name voor wetenschapsgebieden die werken met gegevens over personen?

Big data biedt de wetenschappelijke gemeenschap niet alleen kansen, maar stelt hen ook voor nieuwe uitdagingen. Uitdagingen die liggen op het vlak van toegankelijkheid, analyse en opslag van de data maar die ook te maken hebben met juridische en ethische aspecten. Wetenschappelijk onderzoekers hebben in toenemende mate experts nodig om hen in onderdelen van het onderzoek met big data te ondersteunen. Ook levert big data nieuwe vragen op rond replicatie, validiteit, en generaliseerbaarheid van wetenschappelijk onderzoek, en toepasbaarheid van bestaande methoden en statistische technieken.

Het is daarom wenselijk dat onderzoekers die met gegevens over personen werken discipline-overstijgende en discipline-specifieke methoden en technieken ontwikkelen voor de omgang met big data conform de wettelijke kaders en richtlijnen. Dit is noodzakelijk om kansen voor de wetenschap ten volle te benutten en tegelijkertijd zowel de onderzochte personen, als de onderzoekers en hun omgeving te beschermen. Hoewel onderzoek met big data meer de inductieve kant van de wetenschap aanspreekt, is er juist behoefte aan theorievorming. Big data maakt theorievorming misschien nog wel belangrijker dan traditioneel al het geval was. Verder is het raadzaam dat onderzoekers die met gegevens over personen werken teams vormen met dataspecialisten en

# SUMMARY

Researchers select and construct 'big data' in the same way as they do 'regular' data. The difference between big data and regular data comes down to quantity and variety. This report addresses the question of what will happen to the relationship and interaction between data, knowledge and applied research now that 'regular' data are developing into 'big' data? What are the implications for academic research, especially those fields that work with personal data?

Big data offers the research community opportunities, but it also poses new challenges with respect to data accessibility, analysis and storage and raises various legal and ethical issues. Researchers increasingly need to call in experts to assist in using big data. Big data is also generating new questions concerning the reproducibility and validity of research, how far we can generalise from research results, and whether existing methods and statistical techniques remain useful.

Researchers who work with personal data should therefore develop cross-disciplinary and discipline-specific methods and techniques for dealing with big data, in accordance with the statutory frameworks and guidelines. That is what is needed to make full use of big data for the benefit of research while at the same time protecting research subjects, as well as researchers and their networks. Although research involving big data has greater affinity with inductive reasoning, theoretical underpinnings are precisely what is required. Big data may make it even more important to work from a theoretical basis than has traditionally been the case. It is also advisable for researchers who use personal data to work in teams with data specialists and legal and ethical experts, all of whom should be given full credit for their expert support. Another recommendation is that researchers should have easy

collega's met juridische en ethische expertise. Daarbij dient ondersteunende expertise volwaardige erkenning te krijgen. Het is verder wenselijk dat onderzoekers eenvoudig toegang hebben tot deze expertise, bij voorkeur ingebed in een solide en veilige infrastructuur in een onderzoeksinstituut of faculteit.

Huidige en komende generaties onderzoekers die met gegevens over personen werken, moeten worden opgeleid in het gebruik van big data. Voor de individuele universiteiten en umc's, en de VSNU en NFU, is daarbij een belangrijke rol weggelegd. Dit geldt niet alleen voor technische aspecten, maar ook voor zaken als datakwaliteit, ethiek en privacy.

Voor goede aansluiting van lokale databronnen en infrastructuur, en het faciliteren van samenwerking binnen de wetenschappelijke gemeenschap en met partners uit het publieke en private terrein, is het essentieel dat wordt ingezet op een overkoepelende infrastructuur. Bij het inrichten van die nieuwe infrastructuur dient het ministerie van OCW, in overleg met onder meer NWO, SURF en VSNU, met nadruk onderzoek met gegevens over personen te betrekken en aan te sluiten bij initiatieven rond *Open Science*, *Open Data* en *FAIR* data, zoals het internationale GO FAIR-netwerk en het Nationaal Platform Open Science.

Big data biedt veel nieuwe kansen om in Nederland wetenschappelijk onderzoek in een groot aantal toepassingsgebieden te versterken. Als Nederland deze kansen weet te benutten kan het big data-onderzoek internationaal onderscheidend worden en bijdragen aan het verkrijgen van een vooraanstaande positie in de wetenschap.

access to such expertise, preferably embedded in a sound, secure infrastructure within a research institute or faculty.

Current and future generations of researchers who work with personal data should be trained in using big data, a task in which all universities and university hospitals, the Association of Universities in the Netherlands (VSNU), and the Netherlands Federation of University Medical Centres (NFU) have a role to play. Training should address not only technical aspects but also such matters as data quality, ethics and privacy.

To ensure the integration of local data sources and infrastructure, and to facilitate cooperation within the research community and with public- and private-sector partners, an overarching infrastructure is required. When designing this new infrastructure, the Ministry of Education, Culture and Science, acting in consultation with the Netherlands Organisation for Scientific Research (NWO), SURF (collaborative ICT organisation for Dutch education and research) and VSNU, should focus on research that involves personal data and should join Open Science, Open Data and FAIR data initiatives, for example the international GO FAIR network and the Netherlands' Open Science Platform.

Big data offers new opportunities to enhance research in the Netherlands in many different areas of application. If the Netherlands takes advantage of these opportunities, it can boost its international reputation in big data research and attain a prominent position in academic research.

# INTRODUCTIE

## Aanleiding

Wetenschappelijke kennis wordt verkregen door ideeën, logica en waarnemingen met elkaar te combineren en confronteren. De wetenschappelijke gemeenschap verwerkt waarnemingen, ook wel aangeduid als gegevens ofwel data, tot (nieuwe) wetenschappelijke kennis die toepassingen in de maatschappij en in het persoonlijke leven kan vinden. In deze visie zijn data, kennis en toepassingen onlosmakelijk met elkaar verbonden. Wat gebeurt er met de relatie en wisselwerking tussen data, kennis en toepassingen nu *'gewone' data* zich ontwikkelen naar *big data?* Dit rapport gaat met name in op de gevolgen van big data voor wetenschapsgebieden die werken met gegevens over personen.

De opkomst van big data[1] wordt gedreven door innovaties in de informatie- en communicatietechnologie. Nieuwe toepassingen van kennis komen daardoor snel dichterbij, denk aan *personalised medicine*, zelfrijdende auto's en slimme huizen. Ook de manier waarop mensen hun sociale netwerk onderhouden en hun weg vinden in het verkeer is sterk veranderd. Sommige inzichten en toepassingen die tot voor kort *science fiction* leken, zijn door de invloed van big data realiteit geworden.

Deze ingrijpende veranderingen alleen al rechtvaardigen verkenning van de kansen voor de wetenschap. Maar de opkomst van big data heeft ook nog andere gevolgen. Big data over personen wordt gevoed door massaal gegevens over die personen bij elkaar

1 Dit rapport gebruikt de term 'data' als meervoudsvorm, zoals gebruikelijk in het Nederlands. Het rapport gebruikt 'big data' daarentegen als enkelvoud, aangezien dit begrip verwijst naar een verschijnsel en niet per se naar de data zelf.

te brengen en te combineren. Het gebruik van de resulterende datasets en kennis kan belangrijke bedreigingen voor de persoonlijke levenssfeer tot gevolg hebben. Hoe moeten onderzoekers omgaan met deze bedreigingen? Big data vraagt vaak om specifieke data-expertise en is ook minder geschikt voor traditionele statistische gegevensverwerking. Zijn onderzoekers in staat op een juiste manier met big data om te gaan en kunnen zij de potentie ervan benutten?

Het belangrijkste onderscheid tussen 'gewone' data en big data zit in hoeveelheden en grote verscheidenheid aan soorten (hoge *dimensionaliteit*) van gegevens die op talloze terreinen kunnen worden vergaard, opgeslagen, gekoppeld en geanalyseerd. De hoeveelheden en hoge dimensionaliteit blijven explosief toenemen. Het arsenaal aan instrumenten dat continu data produceert groeit, zowel in het wetenschappelijk onderzoek als elders in de maatschappij. Die instrumenten bepalen mede hoe mensen de wereld om zich heen (kunnen) waarnemen. De parallelle ontwikkelingen in hardware en software voor het verwerken en analyseren van data leiden tot zowel nieuwe problemen als nieuwe mogelijkheden. Onderzoekers, bedrijven, overheden, maatschappelijke organisaties en burgers hebben hier onvermijdelijk allemaal mee te maken.

Wereldwijd wordt het innovatie- en economische potentieel van onderzoek met big data onderkend. Dit blijkt onder meer uit grote investerings- en onderzoeksprogramma's in Nederland en daarbuiten, zoals Commit2data (Team ICT, 2015), het big data-thema in de Nederlandse Wetenschapsagenda (2016), de Digitale Samenleving (VSNU, 2016) en diverse EU-programma's (EOSC, 2017; Europese Commissie, 2015, 2017). Big data is de sleutel geworden van het succes van bedrijven als Google en Facebook. Ook in veel wetenschappelijke disciplines, van taalkunde tot en met astronomie, is het gebruik van big data inmiddels een gangbare praktijk. Dit neemt niet weg dat er onder wetenschappelijk onderzoekers ook nog vaak onbekendheid is met dit onderwerp en onzekerheid over de gevolgen van big data voor de onderzoekspraktijk.

Big data biedt veel nieuwe kansen voor wetenschappelijk onderzoek in een groot aantal toepassingsgebieden. Als voorbeeld wordt verwezen naar de snelle ontwikkelingen op het gebied van *personalised health* (zie **Box 1**). Iedereen heeft belang bij goed onderzoek op al die toepassingsgebieden. Het is hierdoor ook in ieders belang dat onderzoek effectief, veilig, en met respect voor de individuele privacy wordt verricht. Tegelijkertijd stelt het fenomeen big data de wetenschappelijke gemeenschap voor problemen van uiteenlopende aard (boyd and Crawford, 2012; Kitchin et al., 2014). Het gaat dan bijvoorbeeld om zaken als toegankelijkheid, analyse en opslag van de data en om juridische en ethische aspecten. Verder is vaak geopperd dat big data het wetenschappelijk proces radicaal zal veranderen. Er zou sprake zijn van een paradigmaverschuiving in de wetenschappelijke methode, en het gebruik van theorie zou overbodig worden (Hey et al., 2009; Bell et al., 2009). Kortom, er is meer dan genoeg aanleiding voor een verkenning van de gevolgen van big data voor de wetenschap.

## BOX 1. *PERSONALISED HEALTH*

**Welke big data is er in het wetenschapsgebied ter beschikking gekomen voor wetenschappelijk onderzoek, die er eerder niet was?**

Technische innovaties bieden steeds meer mogelijkheden om snel en tegen steeds lagere kosten genetische gegevens en data over genexpressie, metabolieten en een breed scala aan *biomarkers* te genereren (*next generation DNA sequencing*, multi-omics). Daarnaast is het verzamelen en analyseren van data over voeding, gedrag, de fysieke en sociale omgeving, gezondheid en ziekte gemakkelijker geworden. Er ontstaat daardoor een enorme 'diepte' van gegevens. Voor een individu zijn data op veel verschillende lagen beschikbaar: omgeving, gedrag, genetisch, expressie en metabolieten. Innovaties in technologie en data-analysetechnieken (bio-informatica) hebben ertoe geleid dat die data ook kunnen worden gekoppeld, geanalyseerd, gevisualiseerd en geïnterpreteerd.

**Welke kansen biedt de beschikbaarheid van deze data voor het stellen en het beantwoorden van onderzoeksvragen in het wetenschapsgebied?**

Door die grote hoeveelheden gedetailleerde data heeft het inductieve, hypothese-genererende onderzoek een vlucht genomen. Dit heeft tot nieuwe onderzoeksvragen geleid en nieuwe perspectieven. Het is beter mogelijk de effecten van de fysieke en sociale omgeving op gezondheid van mensen te onderzoeken, bijvoorbeeld door data over luchtverontreiniging, landgebruik en geluid te koppelen aan gezondheidsuitkomsten en *biomarkers* (Stieb et al., 2017). Ook is het beter mogelijk om de *pathways* te onderzoeken: van genotype en blootstelling naar het ontstaan van ziekte. Hierdoor groeit het inzicht in de verschillen tussen mensen wat betreft het ontstaan en het verloop van ziekten. Daardoor is differentiatie in gezondheidsadviezen en behandeling mogelijk en komt de genezing van bepaalde ziekten dichterbij.

**Welke bedreigingen zijn er voor de toepassing van big data in het wetenschapsgebied?**

Rijke data en hoge datadichtheid per deelnemer maakt anonimisering in dit wetenschapsgebied tegenwoordig vrijwel onmogelijk. Dit kan door privacyoverwegingen het gebruik en delen van data met collega's in de weg staan. Ook vragen data die met hoge snelheid worden gegenereerd om een solide, veilige IT-omgeving, gestandaardiseerde dataverwerkingsprocessen en *data science*-expertise. Er zijn specifieke oplossingen nodig voor het annoteren, visualiseren en interpreteren van data en onderzoeksresultaten. De benodigde faciliteiten, methoden en technieken en *data science*-expertise brengen hoge kosten met zich mee.

**Hoe wordt er omgegaan met deze bedreigingen?**

De biobanken betrekken deelnemers nauw bij het onderzoek. Zij zien deelnemers in toenemende mate als partner (BBMRI-NL, 2014) en koersen op een participatieve dialoog. In diverse onderzoeksgebieden begint een hechte samenwerking te ontstaan tussen inhoudsdeskundigen, dataspecialisten en bio-informatici. Landelijk en internationaal zijn er veel lopende initiatieven voor grootschalige infrastructuur opgezet (onder

andere BBMRI-NL, Data4LifeSciences)[2], die zowel gezamenlijke IT-oplossingen en beleid ontwikkelen als expertise samenbrengen. Elementen hierin zijn *FAIR* data (zie Box 3), de 'Personal Health Train' en de persoonlijke gezondheidsomgeving. Op het vlak van ethische en juridische aspecten voor gebruik van lichaamsmateriaal voor onderzoek is in afwezigheid van wetgeving binnen de biomedische sector de gedragscode Code Goed Gebruik (Federa, 2011) opgesteld.

Vast staat dat de komst van big data onderzoekers uitdaagt nieuwe visies te ontwikkelen op de productie van wetenschappelijke kennis. Bovendien is het noodzakelijk om naast methoden en technieken ook begrippen als privacybescherming en zeggenschap over gegevens te herijken (WRR, 2015). De wetenschap heeft als taak om in brede zin bij te dragen aan een verantwoorde toepassing van big data. Ook voor het behoud van een vooraanstaande positie van Nederland in de wetenschap is het essentieel dat de kansen en bedreigingen voor de toepassing van big data in het wetenschappelijk onderzoek in kaart worden gebracht, inclusief de rol en verantwoordelijkheid van de wetenschappelijke gemeenschap hierin. Deze opgaven zijn leidend geweest voor deze verkenning.

## Taakopdracht commissie 'Big data'

De taak van de commissie, ingesteld door het bestuur van de KNAW, is een brede verkenning van effecten van big data op wetenschappelijk onderzoek en een advies over enkele geselecteerde onderwerpen voor te bereiden. Aangezien big data een zeer diffuus terrein is, maakt de commissie op basis van haar eigen verkenning een beargumenteerde keuze van adviesonderwerpen. De taakopdracht stelt verder dat de commissie haar werkzaamheden dient te verrichten vanuit het perspectief dat de Nederlandse wetenschap internationaal een vooraanstaande en onderscheidende positie in het onderzoek met big data kan verkrijgen. Het accent ligt daarbinnen op wetenschapsgebieden die werken met gegevens over personen. In bijlage 1 is het instellingsbesluit van de commissie 'Big data' opgenomen.

## Doelgroep rapport

De doelgroep van het rapport bestaat primair uit bestuurders en beleidsmakers die te maken hebben met wetenschappelijk onderzoek en secundair uit wetenschappelijk onderzoekers.

2 Verschillende projecten op het gebied van infrastructuur voor *personalised health and medicine*, zoals BBMRI-NL, Data4LifeSciences, DTL en EATRIS, komen samen in het nieuwe initiatief Health Research Infrastructure (Health RI; zie link).

## Gezichtspunt rapport

De commissie heeft gekozen voor het gezichtspunt van de *onderzoeker en de onderzoekomgeving*. Dit is een ander gezichtspunt dan bijvoorbeeld een technisch of economisch gezichtspunt (*for profit*).

## Reikwijdte en beperkingen rapport

Conform de taakopdracht richt de commissie zich op de implicaties van big data voor wetenschappelijk onderzoek. Hierbij gaat het in dit rapport met name om wetenschapsgebieden waar met gegevens over personen kan worden gewerkt. Dit betekent dat bijvoorbeeld onderzoek met data uit de hoge-energiefysica en astronomie of met bedrijfseconomische indicatoren buiten beschouwing zijn gelaten. Het rapport richt zich niet primair op de juridische en ethische aspecten van big data in algemene zin, maar enkel in relatie tot onderzoek met gegevens van personen. Het rapport gaat niet specifiek over technieken die rond analyse van big data in opkomst zijn, zoals *machine-learning*.

## Werkwijze commissie

Voor de totstandkoming van het rapport heeft de commissie in eerste instantie de relevante literatuur grondig bestudeerd. Een belangrijke bevinding is dat er nog geen rijke literatuur bestaat over de effecten van big data op wetenschapsgebieden die werken met gegevens over personen. De commissie heeft daarom een verdiepingsslag gemaakt door tijdens focusgroep-bijeenkomsten te spreken met onderzoekers die regelmatig werken met gegevens over personen. Informatie uit deze bijeenkomsten is gebruikt voor het ontwikkelen van focus en perspectief voor het rapport en is dus niet gebruikt als primaire gegevensbron. De commissie heeft in de periode van maart tot november 2016 drie bijeenkomsten gehouden, elk met een andere focus en samenstelling (bijlage 2 geeft de geconsulteerde personen weer):

1. Focusgroepbijeenkomst over toepassingsgericht onderzoek
   Gericht op vragen als: in welke mate is big data doorgedrongen in wetenschapsgebieden, en wat is de betekenis van big data voor de ontwikkeling van wetenschapsgebieden?
2. Focusgroepbijeenkomst over ict-gerelateerd onderzoek
   Gericht op vragen als: hoe gaat de big data-onderzoeker om met 'verantwoordelijkheid' op punten als transparantie, validiteit en privacy? Ook is gesproken over het inpassen van big data in het onderzoekstraject van data-acquisitie en het hanteerbaar maken van data en interdisciplinaire samenwerking.
3. Klankbordgroepbijeenkomst met onderzoekers uit diverse wetenschapsgebieden.
   Gericht op het bespreken van een eerdere versie van het huidige rapport besproken.

De bevindingen uit de literatuurstudie en de drie bijeenkomsten zijn medebepalend geweest voor de verkenning door de commissie. Op basis van de verkenning heeft de commissie vervolgens een beargumenteerde keuze van adviesonderwerpen gemaakt en op die onderwerpen enkele aanbevelingen geformuleerd.

## Indeling rapport

Dit rapport start met een omschrijving van het begrip big data. Daarna wordt ingegaan op kansen en bedreigingen van big data voor wetenschappelijk onderzoek met gegevens over personen. Vervolgens geeft het rapport aan hoe de kansen van big data kunnen worden benut en hoe de bedreigingen kunnen worden afgewend. In het laatste hoofdstuk staan concrete aanbevelingen.

# WAT IS BIG DATA?

## Verschil tussen 'gewone' data en big data

Technologische innovaties zorgen ervoor dat het verzamelen, verwerken en opslaan van grote hoeveelheden data mogelijk is en dat nieuwe vormen van data ontstaan. Deze ontwikkeling is nog in volle gang. Die grote hoeveelheden, complexe en voortdurend in beweging zijnde data worden kortweg aangeduid als big data. Er zijn diverse omschrijvingen van het begrip big data in omloop. Die hebben met elkaar gemeen dat big data *niet* uit één soort data bestaat, zich *niet* op slechts één locatie hoeft te bevinden en *niet* uit één bron afkomstig is.

In vrijwel elke omschrijving van big data komen 'de vijf V's' terug: 'volume', 'velocity', 'variety', 'veracity' en 'value'. Bij big data gaat het om data van grote omvang ('high in volume'), geproduceerd of bewerkt met grote snelheid ('high in velocity' – real-time observaties die worden gecodeerd op het moment dat ze plaatsvinden), met een grote verscheidenheid in type en vorm ('high in variety' – zowel gestructureerde als relatief ongestructureerde data), waarvan de herkomst en waarheidsgehalte niet altijd goed is in te schatten ('low in veracity') en mogelijk van hoge economische of andersoortige waarde is ('high in value') (Gartner Group, 2011).

Big data is als begrip moeilijk ondubbelzinnig vast te leggen. Alhoewel big data in vrijwel alle wetenschapsgebieden wordt gebruikt, is 'big' relatief en afhankelijk van de discipline. In bijvoorbeeld de natuurkunde geldt vaak een andere opvatting over 'veel' dan in de sociologie. In een aantal wetenschappelijke disciplines was het gebruik van grote hoeveelheden data al langer gangbaar. Dit rapport staat daarom niet langer stil bij een uitputtende begripsbepaling van big data, maar besteedt vooral aandacht

aan de gevolgen voor de wetenschap van data die voldoen aan de kenmerken van big data, waarbij het accent ligt op wetenschapsgebieden die werken met gegevens over personen.

## Geconstrueerd, dus allesbehalve gegeven

Het rapport ziet big data als een variant van data. Data zijn op vele manieren te omschrijven; een gebruikelijke omschrijving stelt data gelijk aan gecodeerde observaties. Die omschrijving maakt meteen duidelijk dat data paradoxaal genoeg allesbehalve *gegeven* zijn. Data zijn altijd *geconstrueerd*. Dit constructieproces kent twee fasen waarin de onderzoeker keuzen maakt. Ten eerste de keuze wat te observeren, en ten tweede hoe observaties te coderen en zo geschikt te maken voor nadere analyse. Deze werkwijze heeft gevolgen voor de plaats van big data in het wetenschappelijk onderzoek. Net als het geval is bij andere data, is big data altijd een product van expliciete of stilzwijgende keuzen. Big data overkomt wetenschappelijk onderzoekers niet: zij kiezen big data om hun eigen onderzoeksvragen te beantwoorden. Voor een goed begrip van de rol van big data in onderzoek is deze constatering van centraal belang.

## Herleidbaar tot personen

Veelal ziet men big data alleen als 'veel data', maar als in het bijzonder wordt gekeken naar onderzoek met gegevens over personen, zitten de uitdagingen meestal niet in het volume (Jin et al., 2015; Jagadish, 2015). Een kenmerk van big data is dat het gaat om data van grote dichtheid in tijd en ruimte, en van hoge precisie, bijvoorbeeld *next generation DNA sequencing* en geografische coördinaten. Dit zijn data waarmee onderzochte personen kunnen worden geïdentificeerd. Anonimiseren of pseudonimiseren[3] van data is in de praktijk meestal niet mogelijk. Dit type gegevens leidt daardoor in de geowetenschappen niet alleen tot vernieuwing maar ook tot nieuwe problemen rond privacy (zie **Box 2**). Bovendien lijkt het door de mogelijkheden tot combinatie van gegevens ook niet langer mogelijk om op voorhand een onderscheid te maken tussen reguliere persoonsgegevens en bijzondere categorieën van persoonsgegevens[4], zoals medische gegevens en gegevens over ras. Dit is problematisch gegeven de

3 Geanonimiseerde gegevens zijn gegevens die niet tot een levend individu (bijvoorbeeld een deelnemer aan wetenschappelijk onderzoek) herleidbaar zijn. Het zijn dus geen persoonsgegevens (zie volgende voetnoot). Pseudonimiseren of coderen is een vorm van verwerken van gegevens waarbij de verbinding tussen een set identificerende gegevens en het individu wordt verwijderd, en een nieuwe verbinding wordt gemaakt tussen een bepaalde set van karakteristieken die verwijzen naar het individu en een of meerdere pseudoniemen. Gepseudonimiseerde gegevens blijven persoonsgegevens.

4 Persoonsgegevens zijn alle gegevens die direct, indirect of via een sleutel herleidbaar zijn tot een levend individu. Anonieme gegevens of gegevens van overleden personen of organisaties zijn geen persoonsgegevens.

verschillende wettelijke regimes voor beide soorten gegevens (Moerel en Prins, 2016). De onderzoeker vindt hiervoor kaders en handvatten in wet- en regelgeving, zoals nieuwe Europese regels inzake de omgang met persoonsgegevens die vanaf mei 2018 gelden.

### BOX 2. GEOPRIVACY

**Welke big data is er in het wetenschapsgebied ter beschikking gekomen voor wetenschappelijk onderzoek, die er eerder niet was?**
Als gevolg van het gebruik van nieuwe technieken, zoals GPS-tracking, sensoren en *beacons*, neemt de beschikbaarheid van zeer nauwkeurige tijd-ruimtelijke en continue data snel toe. Nieuwe geavanceerde technieken in Geografische Informatie Systemen (GIS) maken het mogelijk die data te exploreren, te analyseren en te visualiseren.

**Welke kansen biedt de beschikbaarheid van deze data voor het stellen en het beantwoorden van onderzoeksvragen in het wetenschapsgebied?**
Gedetailleerde geodata in combinatie met andere (gezondheids)kenmerken bieden nieuwe mogelijkheden voor onderzoekers om tot nieuwe discipline-overstijgende inzichten te komen. Een goed voorbeeld zijn allerlei initiatieven om gezondheidsproblemen in ruimtelijke omgevingen te onderzoeken en effectieve interventies in leefstijl en omgeving te ontwikkelen en te beoordelen. In dit kader zijn bijvoorbeeld de concepten 'healthy cities' (WHO, 2017) en 'healthy urban living' (UU, 2017) ontwikkeld. De onderzoeksresultaten worden geleidelijk beschikbaar gesteld aan professionals in de zorg en ruimtelijke ordening.

**Welke bedreigingen zijn er voor de toepassing van big data in het wetenschapsgebied?**
Combinatie van geografische data en analyses maakt het erg makkelijk om personen te identificeren. Woon- en werkplekken, dagelijkse activiteiten en routes kunnen probleemloos worden ontcijferd. Vooral bij gezondheids- en sociale vraagstukken kan dit privacyproblemen veroorzaken.

**Hoe wordt er omgegaan met deze bedreigingen?**
Een gevolg van de ontwikkelingen van methoden en technieken op het gebied van GIS-data dat vanuit ethisch en juridisch oogpunt de noodzaak toeneemt om de geoprivacy van personen te beschermen (Richardson et al., 2015). Dit is een bedreiging die helaas tot nu toe te weinig aandacht heeft gekregen (Bader et al., 2016). Er zijn verschillende manieren om de geoprivacy te beschermen. Een daarvan betreft regulering via wettelijke procedures en ethische richtlijnen. Gebruikers van GIS-data zouden die moeten onderschrijven. Een goed voorbeeld betreft de 'GIS Code of Ethics' van de Urban and Regional Information Systems Association in de VS (URISA, 2003). Daarnaast zijn er technieken beschikbaar om de oorspronkelijke locaties van personen in tijd en ruimte te wijzigen of

te verbergen. Het effect op de analyseresultaten van dergelijke technieken is nog grotendeels onbekend (Richardson et al., 2015). In aanvulling hierop wordt door de National Science Foundation (NSF) in de Verenigde Staten aanbevolen een Geo Virtuele Data Enclave te ontwikkelen (GVDE, zie link). Belangrijke kenmerken van een GVDE betreffen de opslag, combinatie en analyse van data, methoden en technieken op een beveiligde server. Vanaf een lokale desktop kan daarop ingelogd en 'encrypted' mee worden gecommuniceerd zonder dat door de gebruiker beschikbare data kunnen worden gedownload of verzonden. De analyseresultaten worden voor verzending naar de gebruiker gecontroleerd op geoprivacy. Eenzelfde type systeem wordt gebruikt door het Centraal Bureau voor de Statistiek (CBS) ('remote access') en in het domein van de levenswetenschappen door LifeLines ('virtual research workspace') en diverse umc's ('virtual research environment').

## Produceren en hergebruiken van data

Tot enkele jaren geleden waren wetenschappelijk onderzoekers bij uitstek degenen die beschikten over grote hoeveelheden data. Vaak stonden ze aan de bron doordat ze die data zelf verzamelden. Met de komst van big data is ook het secundair datagebruik door onderzoekers sterk toegenomen, bijvoorbeeld van data afkomstig uit sociale media (Twitter of Facebook) of de detailhandel (klantenkaart en online gedrag). Wetenschappelijk onderzoekers zijn samen met vele anderen producenten van data geworden en niet langer een van weinigen met toegang tot data.

Onderzoekers gebruiken dus niet alleen door hen zelf verzamelde primaire data, ze maken in toenemende mate gebruik van door anderen verzamelde secundaire data. Deze ontwikkeling wordt versterkt door initiatieven die het wetenschapsproces transparanter willen maken (Munafò et al., 2017) en het hergebruik van bestaande wetenschappelijke en publieke data stimuleren (Pasquetto et al., 2017; Wilkinson et al., 2016), zoals de beweging richting *Open Science*, *Open Data* en *FAIR* data (*FAIR*: *findable, accessible, interoperable and reusable*; zie **Box 3**). Ze stimuleren wetenschappelijk onderzoekers niet alleen om data te hergebruiken, maar ook om data te delen en toegankelijk te maken voor anderen.

In de wetenschap ontstaat big data veelal binnen grote consortia, waar data worden gedeeld, gekoppeld en geharmoniseerd om gecombineerde data-analyses mogelijk te maken. Door het federatieve karakter van deze initiatieven is er vaak niet langer sprake van één verantwoordelijke. Verantwoordelijkheden en zeggenschap over de gecombineerde data zijn soms complex, bijvoorbeeld rond de verantwoordelijkheid voor de naleving van privacyregelgeving. In plaats van aan de bron te staan van de data zijn wetenschappelijk onderzoekers steeds vaker onderdeel van een keten.

### BOX 3. DE BEWEGING RICHTING *OPEN SCIENCE*, *OPEN DATA* EN *FAIR* DATA

*Open Science* gaat over het beoefenen van wetenschap waarbij iedereen de gelegenheid heeft bij te dragen aan en gebruik te maken van de resultaten van het wetenschappelijk proces (Nationaal Plan Open Science, 2017). Er is een principieel standpunt, namelijk dat onderzoek dat met publieke middelen is gefinancierd ten goede komt aan de gehele samenleving. Daarnaast biedt *Open Science* wetenschappelijk onderzoekers nieuwe kansen voor het uitwisselen van resultaten van hun onderzoek en van methoden, technieken en wetenschapspraktijken. Dit kan de kwaliteit van hun werk ten goede komen. *Open Science* stimuleert data-intensief onderzoek doordat via het koppelen van data, nieuwe wetenschappelijke vragen kunnen worden gesteld (Pasquetto et al., 2017). Om koppeling en hergebruik daadwerkelijk mogelijk te maken, is het essentieel dat data vindbaar en toegankelijk zijn, niet alleen door mensen, maar ook door computers (*machine-actionable*). Hiervoor wordt het *FAIR*-acroniem gebruikt (*findable, accessible, interoperable and reuseable*) (Wilkinson et al., 2016). Dit betekent niet dat alle data altijd *open* (*Open Data*) en voor iedereen toegankelijk moeten zijn. Er is momenteel een internationale beweging om wetenschappelijke data zoveel mogelijk *FAIR* te maken (EOSC, 2017).

## Samenvatting

In dit hoofdstuk is een omschrijving gegeven van big data, en is ingegaan op een aantal aspecten daarvan, met een accent op gegevens over personen. Hieruit volgen kansen en bedreigingen voor de wetenschappelijke gemeenschap bij de toepassing van big data. Big data 'overkomt' onderzoekers niet; alle data, dus ook big data, worden gekozen en geconstrueerd door onderzoekers. Het volgende hoofdstuk werkt verschillende consequenties voor de praktijk van het onderzoek met gegevens over personen verder uit.

# CONSEQUENTIES VAN BIG DATA VOOR DE PRAKTIJK VAN HET ONDERZOEK

De bijzondere eigenschappen van big data ten opzichte van 'gewone' data geven specifieke uitdagingen voor de praktijk van het wetenschappelijk onderzoek met gegevens over personen. In dit hoofdstuk passeert een aantal daarvan de revue.

## Onderzoeksmethoden en -technieken

Technologische innovaties hebben nieuwe mogelijkheden geschapen voor metingen en nieuwe dataverzamelingen die niet eerder mogelijk waren. Computers zijn niet langer alleen een *computing instrument*, maar worden ook gebruikt voor dataverzameling en -verwerking, zoals *information retrieval device* en als *data gathering & analysis system*. In toenemende mate worden data passief verzameld, dus zonder voortdurende aandacht van de onderzoeker en met minder inspanning van de onderzochte personen. Voorbeelden zijn het dragen van *tracers*, uitlezen van sensoren in huizen en *wifi-checking.* In die metingen is de kans op sociaal wenselijke antwoorden kleiner dan in vraaggesprekken en surveyonderzoek. Positief is dat daarmee de geldigheid van de data kan toenemen. Positief is ook dat door instrumenten zoals *tracers* en sensoren data kunnen worden verzameld van personen bij wie dat vroeger lastig was door fysieke of mentale beperkingen, taalbarrières, of problemen met bereikbaarheid voor surveyonderzoekers. Dit zien wetenschappers als kansen voor de toepassing van big data in het onderzoek met gegevens over personen.

Als gevolg van de toegenomen hoeveelheid data, de complexiteit en het *streaming*-karakter zijn er nieuwe methoden en technieken gekomen voor het verwerken en analyseren van de data zoals *data-mining*, *machine-learning* en *deep-learning*. Als voorbeeld kan worden gedacht aan ontwikkelingen in online marketing (zie **Box 4**). Deze technologische veranderingen vragen om aanpassingen van de onderzoeksmethoden

en -technieken. Hierbij gaat het onder meer om het ontwerp en de opzet van het onderzoek dat leidt tot het verkrijgen van de big data. De conventionele statistische toetsen verliezen bijvoorbeeld hun betekenis bij een zeer grote steekproefgrootte. Het gevaar van *p-hacking* blijft aanwezig. Dit wil zeggen dat de wetenschappelijk onderzoeker net zo lang analyses uitvoert tot de 'kans' dat een resultaat op statistisch toeval berust onder de grens van, zeg, 5 procent is gezakt. Het risico op *HARKing* (*Hypothesizing After the Results are Known*) wordt groter doordat die praktijk bij big data meer kans op succes biedt (Munafò et al., 2017).

Ook is er het gevaar dat de beschikbare data de onderzoeksvraag sturen, in plaats van andersom, waardoor er tunnelvisie optreedt. Zoals eerder gesteld, is de eerste en belangrijkste remedie tegen deze bedreigingen het expliciteren van de theoretische verwachtingen van de onderzoekers. Hiernaast moeten, gezien het tekortschieten van conventionele statistische toetsen, nieuwe onderzoeksmethoden en -technieken voor het omgaan met big data worden ontwikkeld. De ontwikkeling van nieuwe onderzoeksmethoden en -technieken beschouwen wetenschappelijke onderzoekers als een kans voor de toepassing van big data in het onderzoek met gegevens over personen.

### BOX 4. CONSUMENTENGEDRAG EN DE EFFECTEN VAN ONLINE MARKETING

**Welke big data is er in het wetenschapsgebied ter beschikking gekomen voor wetenschappelijk onderzoek, die er eerder niet was?**
Tot een jaar of tien geleden waren gegevens over de wijze waarop consumenten reageren op informatie over een product alleen geaggregeerd beschikbaar. Sindsdien zijn dergelijke gegevens massaal online beschikbaar gekomen op microniveau. Hierdoor heeft het onderzoek naar consumentengedrag een grote vlucht genomen, met name in relatie met online marketing. Niet alleen is het nu mogelijk individuele mensen te volgen in hun zoekgedrag naar een product, het is ook mogelijk te experimenteren met de wijze waarop zij reageren op informatie over dat product. In theorie is het mogelijk met miljoenen mensen en producten experimenten te doen.

**Welke kansen biedt de beschikbaarheid van deze data voor het stellen en het beantwoorden van onderzoeksvragen in het wetenschapsgebied?**
Onderzoek naar de invloed van marketing op consumentengedrag was traditioneel gericht op het ontwikkelen van theoretische kennis op het niveau van groepen, die vervolgens werd gevalideerd in experimenten. Een voorbeeld hiervan is het principe van *social proof*, wat wil zeggen dat mensen geneigd zijn te doen wat anderen doen. Dit principe is veelvuldig in experimenten bevestigd: mensen zijn inderdaad meer geneigd bestsellers te kopen. Met gebruikmaking van online marketing kan echter worden gewerkt met veel meer producten en veel meer mensen tegelijk, terwijl van ieder van deze personen het individuele keuzegedrag kan worden geobserveerd. Deze

mogelijkheden hebben ertoe geleid dat in de online marketing inmiddels veel aandacht uitgaat naar de ontwikkeling van algoritmes die de *match* tussen product (en *product pitch*) en consument op individueel niveau zo efficiënt mogelijk maken. De aandacht in het marketing- en consumentengedragsonderzoek is hierdoor voor een deel verschoven van theoretische concepten, zoals *social proof*, naar het ontwikkelen en toepassen van nieuwe methoden en technieken.

**Welke bedreigingen zijn er voor de toepassing van big data in het wetenschapsgebied?**
Door het toegenomen aantal observaties en door technieken vanuit de *machine-learning* kunnen patronen worden ontdekt in data over consumentengedrag. Maar onduidelijk blijft vaak wat die patronen precies zeggen en hoe consumenten daadwerkelijk tot een keuze komen. De uitspraak "hoe meer data, hoe beter" is dus niet altijd waar. Meer data leidt wellicht tot betere beslissingen, maar niet altijd tot een beter begrip van de onderliggende processen. Hier blijft telkens de vraag interessant wat de causale verbanden zijn: het wetenschapsgebied observeert dan wel veel meer keuzegedrag, maar het bepalen van de oorzaken van dat gedrag op individueel niveau is een grote uitdaging.

**Hoe wordt er omgegaan met deze bedreigingen?**
Het streven is er zeker van te zijn dat de gevonden verbanden causaal zijn en niet alleen correlaties beschrijven. De online marketing ontwikkelt zich sterk door het toepassen van methoden en technieken die dit streven ondersteunen. Dit gebeurt onder meer door het inzetten van nieuwe statistische methoden (bijvoorbeeld *doubly robust estimation* van causale effecten). Tevens wordt heterogeniteit steeds vaker expliciet gemodelleerd door het gebruik van hiërarchische modellen.

## Kennis en vaardigheden voor uitvoering van wetenschappelijk onderzoek

Het gebruik van big data voor onderzoek vraagt in toenemende mate om specifieke kennis en vaardigheden die veel wetenschappelijke onderzoekers, specialisten in hun eigen discipline, niet hebben (Munevar, 2017). In het onderzoeksproces waarin big data wordt gebruikt, zijn meer bewerkingen nodig voordat de data kunnen worden geanalyseerd. De bewerkingen worden veelal geautomatiseerd uitgevoerd (via *pipelines*, automatische *data-processing*), en er worden federatieve en/of gedistribueerde systeemanalyses uitgevoerd.

De onderzoeker kan hierdoor behoefte hebben aan kennis en vaardigheden op het vlak van databeheer (metadatering en dataharmonisatie) en IT-oplossingen als een High Performance Computing (HPC)-faciliteit, gedistribueerde systeemanalyses, datavisualisatie en analyse-*pipelines*. Los van inhoudelijke kennis gaat het dus ook om diepgaande kennis van data, methoden en IT-oplossingen. De onderzoeker die de

onderzoeksvraag stelt is inhoudsdeskundig, maar heeft niet altijd de gewenste data-deskundigheid.

In toenemende mate wordt een deel van het onderzoeksproces uitgevoerd door anderen dan de onderzoeker zelf. Opeenvolgende stappen in het onderzoeksproces worden gezet door verschillende personen met specifieke kennis en vaardigheden op het gebied van big data. Het creëren van onafhankelijke rollen in het onderzoeksproces die 'blind' zijn voor de gewenste uitkomst van het wetenschappelijk onderzoek kan bijdragen aan transparantie (Munafò et al., 2017) en de bescherming van de privacy van de onderzochte personen. Het vereist echter wel goede afstemming en vraagt om documentatie en controle op datakwaliteit en validiteit. Naast de specifieke kennis en vaardigheden benodigd voor het goede verloop van dit proces, is werken met big data ook nog eens arbeidsintensief. Als de wetenschappelijke gemeenschap niet adequaat op deze wijze de transparantie van het onderzoeksproces en de privacy van de onderzochte personen zou borgen, kan dit negatieve gevolgen hebben voor onderzoek.

## Privacy en ethiek

Het grote volume en vaak fijnmazigere karakter van data plus het gemak waarmee data worden verzameld, bewerkt, gekoppeld en gedeeld, kunnen tot problemen leiden op het vlak van privacy, eigendom, aansprakelijkheid, jurisdictie en ethiek (Bader et al., 2016; Rathenau Instituut, 2017). Keer op keer blijkt dat big data slecht te anonimiseren is en dus herleidbaar blijft tot personen (Metcalf et al., 2016; Mostert et al., 2016). Enkele bekende voorbeelden staan in **Box 5**. De beweging richting *Open Science*, *Open Data* en *FAIR* data (zie Box 3) en subsidiegevers die transparantie en hergebruik stimuleren, versterken de praktijk van het delen en combineren van data. Juist dit herhaaldelijk delen en combineren van big data, ook wel *journeys of data* (Leonelli et al., 2016) genoemd, vergroot de uitdagingen rond privacy, jurisdictie en ethiek van datagebruik. Wie is verantwoordelijk voor het beschermen van de privacy van onderzochte personen in een keten van datagebruik? Moet er toestemming worden gegeven voor (her)gebruik van data? Hoe en wanneer bepaalt de onderzoeker of geanonimiseerde data afkomstig uit meer bronnen, na koppeling niet toch herleidbaar zijn tot personen? Hoe gaan wetenschappelijk onderzoekers om met de wettelijke voorwaarde van doelbinding die de privacyregelgeving stelt? Is er een waarschuwingsplicht als er potentieel relevante informatie wordt gevonden over onderzochte personen (een discussie die internationaal tevens is gevoerd naar aanleiding van *next generation DNA sequencing*)? Een zorgvuldige beoordeling van de ruimte voor hergebruik van data is vorm te geven door invulling van het criterium van een gerechtvaardigd belang bij hergebruik (Moerel en Prins, 2016).

### BOX 5. ONGEWILDE IDENTIFICATIE

De herleidbaarheid van big data tot individuele personen kan tot curieuze, onbedoelde kennis leiden. Een klassiek voorbeeld is het relaas van de Amerikaanse Thelma Arnold, geabonneerd op America Online. Deze Amerikaanse provider stelde in 2006 drie maanden de geanonimiseerde zoekgeschiedenis van gebruikers beschikbaar voor de wetenschap (Barbaro et al., 2006). Aan de hand van haar zoekopdrachten werd Thelma Arnold al snel geïdentificeerd als gebruiker #4417749 en kon, onder veel meer, zelfs worden vastgesteld dat haar hond zindelijkheidsproblemen had. Een ander voorbeeld, uit 2013, betreft het openbaar maken van gegevens over 173 miljoen taxiritten, wat leidde tot identificatie van personen inclusief hun tijdsbesteding, waaronder bezoek aan nachtclubs (Tockar, 2015). In 2008 werd een studie gepubliceerd die liet zien dat het mogelijk is om personen te identificeren met minieme hoeveelheden DNA in een mix van DNA. Hoewel dat op zichzelf nog geen probleem hoeft te geven, kunnen deze gegevens worden gecombineerd met andere data (zoals locatie) of kunnen kenmerken in het DNA leiden tot gevoeligheden (Homer, 2008).

Wanneer onderzochte personen zich onvoldoende beschermd voelen en het gevoel hebben grip op en zeggenschap over hun data te verliezen, bestaat de kans dat zij minder informatie delen. Daardoor kan er sterke zelfselectie optreden bij het deelnemen aan onderzoek waardoor gegevens minder representatief worden. Het omgaan met gevoelige, mogelijk herleidbare data vereist daarom specifieke kennis en vaardigheden bij allen die een rol spelen in het onderzoeksproces: de wetenschappelijk onderzoeker, de ondersteunende dataspecialisten[5], de ethische commissie, etc. (KNAW, 2016). Tevens is een ethisch verantwoorde *governance* van die data en een veilige infrastructuur onontbeerlijk om de big data te verzamelen, te bewerken, op te slaan en te delen. Onder de infrastructuur vallen diensten zoals pseudonimisatie, Trusted Third Party (TTP, het Centraal Bureau voor de Statistiek en ZorgTTP fungeren bijvoorbeeld als TTP) en *Privacy Impact Assessment* (PIA).

De standaard infrastructuur binnen onderzoeksinstellingen is hier momenteel nog niet op toegerust. Ook sluit het kennisniveau bij wetenschappelijke onderzoekers op het gebied van de juridische en ethische aspecten niet aan bij de gebruikte data of het type onderzoek (Leonelli et al., 2016; Metcalf et al., 2016). Enerzijds kan dit leiden tot onwenselijke situaties waarin de privacy van onderzochte personen onvoldoende wordt gewaarborgd. Anderzijds kan onbekendheid met mogelijkheden en methoden in omgang met data leiden tot het ten onrechte blokkeren van innovatief onderzoek. Dit beschouwt de wetenschap als een bedreiging voor de toepassing van big data in

5 Dataspecialisten zijn bijvoorbeeld de datasteward (die onder andere verantwoordelijk is voor het beheer van de onderzoeksgegevens), de dataconsultant (die de wetenschappelijk onderzoeker de weg wijst naar de juiste data en IT-oplossingen) en de data scientist (die de wetenschappelijk onderzoeker begeleidt in het verzamelen, bewerken en analyseren van data).

het onderzoek met gegevens over personen, indien hiervoor geen aandacht is en geen extra budget en personeel ter beschikking worden gesteld.

De Algemene verordening gegevensbescherming (AVG), die vanaf 25 mei 2018 in Nederland van kracht is, geeft richtlijnen voor het verzamelen en hanteren van data die direct dan wel indirect tot personen zijn te herleiden. Zo is in de AVG de verplichting tot een *Privacy Impact Assessment* (PIA) opgenomen. Individuen krijgen meer rechten en bepaalde organisaties worden verplicht tot het aanstellen van een onafhankelijke privacyfunctionaris en het aanleggen van een register van verwerking van persoonsgegevens.

Zorgvuldige omgang met onderzoekdata in wetenschappelijk onderzoek dient uiteraard met het in werking treden van de AVG op juridisch correcte wijze te worden gecontinueerd. Een aantal aspecten van de AVG moet nog nader worden uitgewerkt, met name inzake de bijzondere waarborgen die gelden voor de verwerking van persoonsgegevens. De uitwerkingen kunnen ruimte geven voor discipline-specifieke ethische normen en zelfregulering.

De juridische werkgroep van het LCRDM (Landelijk Coördinatiepunt Research Data Management, 2016), ingesteld door de VSNU, zal hiervoor aanvullend beleid ontwikkelen. Deze werkgroep heeft tot taak de universitaire instellingen en het bredere onderwijsveld in Nederland beleidsmatig te ondersteunen bij het borgen van zeggenschap, intellectueel eigendom, gegevensbescherming en privacy. Eén uitwerking van de AVG vindt plaats in de *Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek* (VSNU, 2005), waar wetenschappelijke onderzoekers in Nederland zich aan dienen te houden. Die gedragscode wordt nu (stand van zaken april 2018) aangepast aan de nieuwe situatie in 2018. Belangrijk punt onder de nieuwe databeschermingsregels is dat wanneer deelnemers hun toestemming aan het onderzoek intrekken, zij een beroep kunnen doen op het recht ‘om vergeten te worden’. Concreet zal dit betekenen dat de data die tot de betreffende persoon te herleiden zijn, moeten worden gewist dan wel geanonimiseerd. Belangrijk voor big data-onderzoek is ook dat onder de AVG specifieke regels zullen gelden voor data die betrekking hebben op kinderen en jongeren. Overmatige bureaucratisering en protocollering kan hierbij mogelijk een bedreiging gaan vormen doordat het de wetenschappelijke creativiteit afknelt en de onderzoeker van het werk houdt.

### BOX 6. DIGITALE GEESTESWETENSCHAPPEN

**Welke big data is er in het wetenschapsgebied ter beschikking gekomen voor wetenschappelijk onderzoek, die er eerder niet was?**
Er zijn veel data van *social media platforms* en websites in het publieke domein beschikbaar voor mediastudies. Onderzoekers verzamelen met behulp van *scrapers* data van

zo'n platform binnen de beperkingen van de regels van de betreffende aanbieder, en van web-forums, websites. Ook zijn er grote datasets beschikbaar gekomen van bedrijven en organisaties die met onderzoekers samenwerken.

**Welke kansen biedt de beschikbaarheid van deze data voor het stellen en het beantwoorden van onderzoeksvragen in het wetenschapsgebied?**
De geesteswetenschappen omvatten diverse onderzoekstradities, en een beperkt deel daarvan werkt met gegevens over personen (het onderwerp van deze verkenning). Er doen zich volledig nieuwe onderzoeksmogelijkheden voor, bijvoorbeeld naar gebruikersinteractie, disseminatie van content, nieuwe correlaties van mobiel mediagebruik en tv-consumptie. Daarnaast leidt het gebruik van data-analytische tools tot onderzoek naar de epistemische impact daarvan, dus naar vragen zoals: "Hoe beïnvloedt het algoritme van mijn tool de uiteindelijke netwerkanalyse of de visualisatie van data?".

**Welke bedreigingen zijn er voor de toepassing van big data in het wetenschapsgebied?**
Er zijn diverse uitdagingen voor de digitale geesteswetenschappen. Rieder en Röhle (2012) noemen: (a) *objectivity*: data en datasets worden snel als objectief beschouwd en er wordt onkritisch mee omgegaan; (b) *visual evidence*: de visualisatie wordt als bewijs gehanteerd zonder kritisch te onderzoeken welke factoren tot het resultaat hebben geleid; (c) *black boxing*: het zicht is beperkt op wat de analysetools doen en hoe ze dat doen; (d) *institutions:* de rol van instituties in het inzetten en reguleren van datapraktijken, en de zeggenschap van bedrijven over toegang tot en gebruik van data en analysetools, en (e) *total science:* de valkuil om één methode als passend te zien voor een brede reeks van onderzoeken, vragen en verklaringen.
In de praktijk is het in dit wetenschapsgebied, net als in veel andere wetenschapsgebieden, mogelijk om in zeer korte tijd met betrekkelijk weinig technische kennis veel data over personen te downloaden, wat kan leiden tot inbreuk op de privacy. Hier schieten de bestaande ethische en juridische kaders tekort. De grote hoeveelheid data en de beschikbare data-analytische tools nopen onderzoekers tot het continu blijven doordenken van die kaders. Het gebrek aan beschikbare expertise kan de toepassing van nieuwe datapraktijken en -analyse in dit wetenschapsgebied bemoeilijken.

**Hoe wordt er omgegaan met deze bedreigingen?**
Bewustwording van risico's bij gebruik en toepassing van data en het ontwikkelen van nieuwe methoden en technieken heeft ook binnen de mediastudies veel aandacht. Belangrijke opgaven liggen op het vlak van ethische en juridische kennis en bij het opleiden van studenten binnen de mediastudies. Onderzoekers participeren in discussies over ethiek, data, datamanagement, en over onderwijsontwikkeling. Bijvoorbeeld bij het ontwikkelen van data-analytische tools is permanente discussie over de epistemische impact ervan, en voor de ethische review van onderzoeksprojecten is een specifieke tool ontwikkeld (De Ethische Data Assistent / DEDA).

## Validiteit en generaliseerbaarheid

Grote hoeveelheden data, soms van de gehele populatie, kunnen bijdragen aan betere generaliseerbaarheid van de uitkomsten van onderzoek. Om de vertaling te maken van onderzoeksresultaten naar doelpopulatie (generaliseren), of om de validiteit van resultaten te beoordelen, heeft de onderzoeker inzicht nodig in de herkomst van de data. Bij secundair gebruik gaat het om data die voor andere doeleinden zijn verzameld, ook wel *function creep* genoemd (WRR, 2016). De methoden en technieken voor het verzamelen en verwerken van die data zijn specifiek toegesneden op een bepaalde vraag of behoefte. De data van Twitter zijn bijvoorbeeld niet primair bedoeld voor onderzoek en zijn niet per se representatief voor de populatie die de onderzoeker wil beschrijven (zie **Box 6**). De wetenschappelijk onderzoeker kan dit type data gebruiken voor het beantwoorden van onderzoeksvragen rond gedrag, communicatie en opinies, maar loopt daardoor mogelijk tegen vragen aan rond interne validiteit en generaliseerbaarheid (boyd et al., 2012). Te stellen vragen zijn bijvoorbeeld: "Zeggen de data echt iets over wat ik wil meten?" en "Weet ik genoeg over de context waarbinnen de data zijn verzameld om een schatting te kunnen maken van mogelijke selectieproblemen en de geldigheid van de data?". Door onvoldoende zicht op de geldigheid van de data en op de context waarbinnen de data zijn verzameld, kan de externe geldigheid van de uitkomsten van het onderzoek in het geding zijn. Ook is er een grotere kans op fouten bij data uit meerdere bronnen, doordat deze gemakkelijker – bewust of onbewust – zijn bewerkt of vervuild. Wetenschappelijke onderzoekers zien dit als bedreigingen voor de toepassing van big data in het onderzoek met gegevens over personen.

Een ander risico voor de validiteit van het onderzoek is dat de wetenschappelijk onderzoeker bij de analyse van big data technieken als *machine-learning* en *deep-learning* kan gebruiken. Die technieken zijn gebaseerd op correlaties en associaties in de data. De wetenschappelijk onderzoeker dient zich er bewust van te zijn dat uitkomsten van die analyses geen harde bewijzen voor causaliteit zijn (zie Box 4). Bovendien zijn de methoden en technieken die daarbij worden gehanteerd allerminst hypothesevrij, bijvoorbeeld door de keuze van waarnemingen en van coderingsprocessen. Hierbij speelt het *closure principle*, dat luidt dat de onderzoeker bij het verzamelen van informatie ergens moet beginnen en het startpunt reeds informatie geeft over het te vinden antwoord. De wijze waarop de wetenschappelijk onderzoeker technieken toepast, bepaalt welke data worden gebruikt en hoe die data worden benaderd. Dit zijn kortom keuzen die de validiteit en generaliseerbaarheid van de uitkomsten van onderzoek met big data kunnen beïnvloeden.

## Verificatie en replicatie

Wetenschappelijke kennis staat of valt bij de (mogelijkheden voor) verificatie en replicatie van de gevonden uitkomsten. Verificatie en replicatie zijn bij het gebruik van grote hoeveelheden data des te belangrijker vanwege de grotere kans op toevallige

bevindingen (doordat meer data gemakkelijker beschikbaar zijn). Aan de ene kant geeft big data mogelijkheden voor verificatie- en replicatieonderzoek. Er zijn immers meer data beschikbaar en data kunnen makkelijker worden gekoppeld en vergeleken. Aan de andere kant lijken de huidige eisen voor het langdurig opslaan van data voor verificatie- en replicatiedoeleinden niet meer haalbaar, juist vanwege de grote omvang en het beweeglijke karakter van big data (*realtime streaming* data).

Verificatie en replicatie kan lastiger zijn doordat de zeggenschap over big data complexer is dan bij data uit primaire bronnen. Zeggenschap over data is soms niet toe te wijzen of te achterhalen als gevolg van het combineren van data uit meer bronnen. Vanwege de complexiteit van databewerkingen en -analyses zijn er meer mensen die data kunnen muteren, waardoor het voor de wetenschappelijk onderzoeker soms onmogelijk is goed te achterhalen wat de data-kwaliteit is. Ook kan het voor de wetenschappelijk onderzoeker problematisch zijn te waarborgen dat de data intact, langdurig bewaard en beschikbaar blijven voor verificatie- en replicatiedoeleinden. De wetenschappelijke gemeenschap beschouwt dit als bedreigingen voor de toepassing van big data in het onderzoek met gegevens over personen.

## De rol van publieksparticipatie

Eén van de grootste kostenposten bij het verzamelen van grote hoeveelheden gegevens over personen is het verzamelen zelf. Om selectie-effecten tegen te gaan is het streven om zeer grote en zo volledig mogelijke groepen te bereiken, en het verhogen van de participatiemotivatie via publieksparticipatie is daarbij een belangrijk middel. Ook kan bij een echt hoge deelname zelfrapportage of beschikbaar stellen van eigen data een wezenlijk onderdeel worden. Bijvoorbeeld in geval van het grootschalig verzamelen van gezondheidsgegevens en -materialen in biobanken is de éénrichtingsopstelling vanuit de onderzoeker als waarnemer tegenover zijn onderzoeksobject minder vruchtbaar gebleken. Men ziet de deelnemers aan de biobanken in toenemende mate als partner (BBMRI-NL, 2014) en koerst voor het wetenschappelijk onderzoek aan op een participatieve dialoog, ook met het oog op de vertrouwenswinst die dit oplevert.

## Multidisciplinair onderzoek

Onderzoekers worden gedwongen samen te werken door (her)gebruik en koppelen van data uit andere disciplines. In wetenschappelijk onderzoek met big data wordt een multidisciplinaire aanpak daardoor steeds gangbaarder. Dit geldt niet alleen voor samenwerking met dataspecialisten, maar ook voor samenwerking met onderzoekers uit andere disciplines. Dit heeft als positief effect dat nieuwe perspectieven worden gecreëerd ('innovatie door buitenstaanders'): het bevordert een open blik en kan tunnelvisie voorkómen. Als de wetenschappelijk onderzoekers echter onvoldoende inhoudelijke kennis hebben of een tekort aan inzicht hebben in de condities waarbinnen de data zijn verzameld, kunnen zij de data niet op een juiste manier interpreteren.

## Samenvatting

Dit hoofdstuk schetst kansen en bedreigingen die volgen uit gebruik van big data bij onderzoek met gegevens over personen. Deze kansen en bedreigingen zijn te vinden bij de onderzoeksmethoden en statistische technieken, kennis en vaardigheden in teamverband, privacy en ethiek, validiteit en generaliseerbaarheid, verificatie en replicatie, publieksparticipatie, en multidisciplinariteit. Het volgende hoofdstuk geeft aan wat er nodig is om de kansen te benutten en de bedreigingen af te wenden.

# WAT IS ER NODIG OM DE KANSEN TE BENUTTEN EN DE BEDREIGINGEN AF TE WENDEN?

In het vorige hoofdstuk is een aantal kansen en bedreigingen geïdentificeerd voor onderzoekers die met big data over personen werken. Dit hoofdstuk geeft aan wat er in ieder geval nodig is om deze kansen te benutten en de bedreigingen af te wenden.

## Nieuwe methoden en technieken, procedures en werkwijzen inzake big data in onderzoek

De formulering van wetenschappelijke theorieën is door het fenomeen big data met gegevens over personen allerminst een achterhaald wetenschappelijk gebruik geworden. Integendeel: het belang van theorieën is juist toegenomen doordat er nieuwe wegen worden geopend voor onderzoek (Levallois, 2013). Big data vraagt om gedegen onderzoeksvragen, methoden en technieken, en in het geval van secundaire data staat de vraag voorop of de onderzoeksvraag met de beschikbare data wel te beantwoorden is. Tevens is er een grote behoefte aan replicatie- en validatiestudies (KNAW, 2018). In dit opzicht is er door de komst van big data dus niets veranderd. Doordat veel methoden en technieken uit een tijdperk stammen van vóór big data, is het niet vanzelfsprekend dat een onderzoeker die met gegevens over personen werkt, beschikt over kennis hoe met big data om te gaan. Hierdoor kan de onderzoeker niet altijd terugvallen op een gevalideerde methode of techniek. Onderzoekers zullen gezamenlijk discipline-overstijgende, en anders discipline-specifieke nieuwe methoden en werkwijzen moeten ontwikkelen.

## BOX 7. DAGBEVOLKING

**Welke big data is er in het wetenschapsgebied ter beschikking gekomen voor wetenschappelijk onderzoek die er eerder niet was?**
Mensen laten tegenwoordig talloze digitale sporen na. Ze gebruiken hun mobiele telefoon en OV-chipkaart, rijden over verkeerslussen en worden gezien door camera's in de openbare ruimte. Dat biedt vergaande mogelijkheden om statistische informatie samen te stellen over de aanwezigheid en mobiliteit op iedere plaats en ieder moment van de dag. De kennis over de zogeheten dagbevolking, verplaatsingspatronen, aantal personen bij evenementen en demonstraties, woon-werkverkeer, toerisme, vrijetijdsbesteding en migratie wordt hierdoor veel breder en dieper dan inzichten uit traditionele bronnen als enquêtes en overheidsregistraties.

**Welke kansen biedt de beschikbaarheid van deze data voor het stellen en het beantwoorden van onderzoeksvragen in het wetenschapsgebied?**
Het massale karakter en lage kosten van big-databronnen maken veel gedetailleerder onderzoek mogelijk op alle terreinen rond aanwezigheid en verplaatsingen van mensen. Het gaat hierbij zowel om beschrijvend als verklarend onderzoek. Waarom neemt de mobiliteit toe? Waarom worden sommige evenementen massaal bezocht en andere niet? Wat zijn de achtergronden van de toenemende migratiestromen tussen landen? Naast wetenschappelijk inzicht leveren de nieuwe bronnen ook baten op voor overheidsbeleid. Gemeenten worden op dit moment bijvoorbeeld gefinancierd op basis van het aantal ingeschreven inwoners (de nachtbevolking). Overdag zijn er vaak veel meer of veel minder mensen aanwezig, waar in de uitkering nu geen rekening mee wordt gehouden.

**Welke bedreigingen zijn er voor de toepassing van big data in het wetenschapsgebied?**
Er zijn twee bedreigingen. Gegevens over de aanwezigheid van mensen zijn zeer privacygevoelig. Er zijn daarom juridische en technische maatregelen nodig om te garanderen dat de gegevens alleen worden gebruikt zonder kans op onthulling van individuele gegevens en zonder dat niet meer mensen toegang hebben dan strikt noodzakelijk is. De andere bedreiging is dat de eigenaren van de data om meerdere redenen vaak erg terughoudend zijn om hun data ter beschikking te stellen. Ze zijn bang ongunstig in de publiciteit te komen, vrezen hoge kosten zonder er zelf baat bij te hebben, en zijn bij strategische data bang dat hun concurrenten er toegang toe krijgen.

**Hoe wordt er omgegaan met deze bedreigingen?**
Onderzoekers dienen zich zeer bewust te zijn van de privacy-gevoeligheid van hun data en daar consciëntieus mee om te gaan. Dit kan bijvoorbeeld via opleidingen, het afleggen van een eed en beoordeling door een ethische commissie. Verder dienen rechten- en toegangsbeheer en *logging* goed te zijn geregeld en dienen gegevens waar mogelijk geanonimiseerd of gepseudonimiseerd te zijn, zoals in het Stelsel van Sociaal-statistische

Bestanden van het CBS. In de privacywetgeving (de Wet bescherming persoonsgegevens en de Algemene verordening gegevensbescherming) moet het gebruik van persoonsgegevens voor onderzoeksdoelen expliciet worden geregeld. Voor de tweede bedreiging is het vooral van belang dat zowel voor de (vaak private) eigenaren van big-databronnen als voor onderzoekers de toegankelijkheid en het gebruik van de gegevens uit die bronnen goed geregeld is. Dat schept duidelijkheid voor alle partijen.

## Solide en veilige infrastructuur

Onderzoekers die met gegevens over personen werken, hebben voor het doen van onderzoek in toenemende mate behoefte aan lokale toegang tot een solide en veilige infrastructuur (Nationale Wetenschapsagenda, 2016; NWO, 2017). Als voorbeeld kan worden gedacht aan het onderzoek naar 'dagbevolking' door het Centraal Bureau voor de Statistiek (zie **Box 7**). Omdat bij onderzoek vaak meerdere partijen zijn betrokken en de benodigde infrastructuur te complex en te kostbaar is om lokaal op te brengen, is een landelijke en internationale infrastructuur die ondersteunt en aanvult van belang. Onder infrastructuur worden zowel de IT- en softwarefaciliteiten verstaan als de experts die de wetenschappelijk onderzoeker ondersteunen bij het gebruik van big data.

Naast de nodige opslagmogelijkheden voor grote databestanden, veilige manieren om data te koppelen en delen, en High Performance Computing (HPC)-faciliteiten, zullen ook onderdelen als *FAIR*-datacatalogi (voor hergebruik en replicatie- en validatiestudies) en systemen voor dataverzameling en –bewerking tot de reguliere infrastructuur gaan behoren.

Ook systemen voor publieksparticipatie dienen zowel nationaal als lokaal te worden gefaciliteerd, bijvoorbeeld voor inzicht van onderzochte personen in de eigen data en *opt-out*-registratie voor personen die niet langer willen meewerken. Er zijn belangrijke Europese initiatieven op het gebied van data-infrastructuur (EOSC, 2017). Het is essentieel dat Nederland en, op lokaal niveau, instituten en faculteiten hier effectief op aansluiten om internationaal onderzoek mogelijk te blijven maken.

## Ondersteuning door dataspecialisten en expertise vanuit juridische en ethische hoek

In toenemende mate zullen anderen dan de wetenschappelijke onderzoeker een deel van het onderzoeksproces uitvoeren. Verschillende personen met elk een specifieke expertise zijn dan betrokken bij de opeenvolgende stappen in dit proces. Ondersteuning door dataspecialisten binnen elk onderzoeksteam zal noodzakelijk zijn.

Onderzoek met big data vereist dus multidisciplinariteit in het onderzoeksteam en een nauwe, volwaardige samenwerking tussen de wetenschappelijk onderzoeker en dataspecialisten.

Het wetenschapsgebied van de *recommender systems* (zie **Box 8**) kan als voorbeeld worden genoemd. Deze volwaardigheid houdt onder meer in dat de onontbeerlijke bijdrage van dataspecialisten aan het onderzoek voldoende wetenschappelijke erkenning vindt. Daarnaast is toegankelijke expertise vanuit juridische en ethische hoek essentieel (Metcalf et al., 2016; Mostert et al., 2016). Deze expertise zal in een vroeg stadium van het onderzoek moeten worden betrokken om te voorkomen dat in een latere fase bijstelling van het onderzoek nodig is danwel onzekerheid ontstaat over eigendom en bewaren van data. Ethische commissies moeten de mogelijke risico's voor de grondrechten van onderzochte personen en de mogelijke maatschappelijke voordelen van het onderzoek tegen elkaar afwegen. Tevens dient op beleidsniveau te worden gereflecteerd op manieren om big data-onderzoek te ondersteunen, bijvoorbeeld door het ontwikkelen van gedragscodes.

### BOX 8. *RECOMMENDER SYSTEMS*

**Welke big data is er in het wetenschapsgebied ter beschikking gekomen voor wetenschappelijk onderzoek, die er eerder niet was?**
Het relatief nieuwe wetenschapsgebied van de *recommender systems* werkt met data uit websystemen (interacties, reviews, plaatjes, etc.) en *social media*; die data kunnen een *streaming* karakter hebben. Zo laat bijvoorbeeld het onderzoek met *user-interaction* data zien waar mensen online op hebben geklikt, wat zij hebben bekeken of gekocht, en hoe zij reageren op de suggesties van het *recommender system*.

**Welke kansen biedt de beschikbaarheid van deze data voor het stellen en het beantwoorden van onderzoeksvragen in het wetenschapsgebied?**
In het verleden werden in het wetenschappelijk onderzoek hypothesen over het gedrag van mensen getoetst. Nu kan men, met onderzoek aan *recommender systems,* trends ontdekken door het gedrag van enkele mensen te analyseren, met die analyseresultaten terug te gaan naar de hypothesen, en vervolgens meer specifieke hypothesen te toetsen op een veel grotere populatie. Door deze onderzoeksopzet te combineren met de correcte onderzoeksmethoden en -technieken, die bijvoorbeeld voorkómen dat het verschijnsel *spurious correlations* optreedt, is er sprake van een andere manier van werken. Zodoende kan het wetenschapsgebied bijdragen aan wetenschappelijke inzichten over het gedrag van mensen en de wereld om ons heen.

**Welke bedreigingen zijn er voor de toepassing van big data in het wetenschapsgebied?**
Voor het maken van voorspellingen in het rechts- en het schoolsysteem, en ook in andere systemen met professionals, worden steeds vaker algoritmen gebruikt. Bijvoorbeeld

rechters in de Verenigde Staten kunnen op basis van algoritmen voorspellen wat de recidivekans van overtreders is en op basis daarvan straffen opleggen. Maar algoritmen kunnen het menselijke oordeel in systemen met professionals niet geheel vervangen, hooguit ondersteunen ze dit. Als onderzoekers in het wetenschapsgebied van de *recommender systems* big data en algoritmen te gemakkelijk zouden gebruiken, kan dit niet alleen de privacy in gevaar brengen van de betrokkenen in het onderzoek, maar ook de kwaliteit van de onderzoeksresultaten ondergraven. Wetenschappelijk onderzoekers mogen big data en algoritmen dus niet argeloos gebruiken. Anderzijds ervaren zij een toegenomen druk om met het bedrijfsleven samen te werken of toegepaste onderzoeksprojecten te doen, bijvoorbeeld ten behoeve van valorisatie, waardoor zij minder tijd hebben om de data echt goed te bekijken en over wetenschappelijke vraagstellingen na te denken. Het is voor hen lastig geworden de 'rust' te vinden (*slow science*) die nodig is voor goed wetenschappelijk onderzoek.

**Hoe wordt er omgegaan met deze bedreigingen?**
Het is de verantwoordelijkheid van de onderzoeker om de onderzoeksvraag – en niet de data of de technieken – centraal te houden in het onderzoek. In het wetenschapsgebied van de *recommender systems* wordt veelal in interdisciplinaire teams samengewerkt, waar de onderzoeksvraag centraal staat. Aan studenten in dit wetenschapsgebied, ook aan studenten met een 'engineering'-achtergrond, wordt geleerd in teams samen te werken.

## Opleiden van wetenschappelijk onderzoekers in het gebruik van big data

Teneinde wetenschappelijk onderzoekers die met gegevens over personen werken meer vaardig te maken in het omgaan met data, zou dit binnen het wetenschappelijke onderwijs een standaard onderdeel moeten zijn van het curriculum. Bij dit onderdeel dienen behalve basiskennis over (big) data ook de ethische en juridische kanten en de mogelijke maatschappelijke impact worden belicht (Leonelli, 2016). Deze onderdelen verdienen aandacht in alle fasen van het hoger onderwijs, maar in het bijzonder in de PhD-opleidingen.

## Samenvatting

Dit hoofdstuk geeft aan wat nodig is om de kansen van big data te benutten en de bedreigingen af te wenden. Hierbij zijn vier resultaatgebieden benoemd. Ten eerste de ontwikkeling van methoden en technieken die bij uitstek geschikt zijn voor de analyse van big data. Ten tweede de bouw en het onderhoud van een solide onderzoekinfrastructuur voor het werken met big data. Ten derde de verdere ontwikkeling van de ondersteuning door dataspecialisten en expertise vanuit ethische en juridische hoek. En ten slotte de ontwikkeling en implementatie van onderwijsonderdelen over data,

big data, en datamanagement in alle fasen van het hoger onderwijs. Het volgende hoofdstuk formuleert in het verlengde hiervan in de aanbevelingen welke actor ermee aan de slag moet.

# AANBEVELINGEN

Big data biedt tal van kansen om het wetenschappelijk onderzoek te versterken. Als Nederland deze weet te benutten, kan het onderzoek dat gebruik maakt van big data internationaal onderscheidend worden en bijdragen aan het verkrijgen van een internationaal vooraanstaande positie. Behalve kansen levert big data ook een aantal bedreigingen op wat betreft gegevens en privacy van personen. Om de kansen voor de toepassing van big data in het wetenschappelijk onderzoek ten volle te kunnen benutten en tegelijk de onderzochte personen, de onderzoekers en hun onderzoekersomgeving te beschermen, wordt aanbevolen:

### VOOR DE WETENSCHAPPELIJKE ONDERZOEKER DIE MET GEGEVENS OVER PERSONEN WERKT:

- Houd voor ogen dat big data theorievorming misschien nog wel belangrijker maakt dan traditioneel al het geval was. De beschikbaarheid van grote hoeveelheden gegevens mag niet leiden tot een tunnelvisie in wetenschappelijk onderzoek.
- Ontwikkel waar mogelijk nieuwe discipline-overstijgende, en anders discipline-specifieke methoden en technieken voor het gebruik van big data voor wetenschappelijk onderzoek. Deze methoden en technieken dienen in overeenstemming te zijn met een zorgvuldige omgang met big data en wettelijke kaders en richtlijnen.
- Vorm onderzoeksteams met dataspecialisten en met juridische en ethische expertise, waarbij deze ondersteunende experts volwaardige erkenning krijgen voor hun bijdrage.

VOOR HET ONDERZOEKSINSTITUUT OF DE FACULTEIT:

- Zorg dat onderzoekers die met gegevens over personen werken lokaal toegang hebben tot een solide en veilige infrastructuur, vergelijkbaar met een lab dat zowel onderzoekers als onderzochte personen een veilige omgeving biedt. Toegang kan betrekking hebben op lokale en nationale faciliteiten (SURF, DANS en Netherlands eScience Center, en op het niveau van wetenschapsgebieden bijvoorbeeld Health RI) of op Europees niveau de European Open Science Cloud (EOSC). Nationale faciliteiten dienen veilig lokaal verbonden te zijn en te passen in het onderzoeksproces.
- Big data stelt hoge eisen aan databeheer, -verwerking, -(her)gebruik en -analyse. Onderzoekers hebben daarbij in toenemende mate behoefte aan ondersteuning door dataspecialisten en juridische en ethische expertise. Die ondersteuning dient duurzaam en dichtbij het onderzoeksproces te zijn ingebed, bij voorkeur in een onderzoeksinstituut of faculteit. Het is belangrijk dat dataspecialisten en juridische en ethische experts niet ad hoc worden bijeengeroepen, maar permanente ondersteuning bieden. Experts met kennis van zaken als het gaat om het gebruik van big data voor onderzoek zijn momenteel nog schaars. Dergelijke ondersteuning is nog niet standaard beschikbaar en zal dus vaak moeten worden ontwikkeld. Er zal door instituten en faculteiten in moeten worden geïnvesteerd.

VOOR DE UNIVERSITEITEN/UMC'S EN DE VSNU/NFU:

- Leid de huidige en komende generaties onderzoekers die met gegevens over personen werken op in het gebruik van big data, zowel in de technische aspecten – zodat ze in staat zijn samen te werken met dataspecialisten – als in zaken zoals verificatie, datakwaliteit, ethiek en privacy.
- Zorg dat bestaande gremia, zoals ethische commissies en juridische afdelingen, goed zijn toegerust door dataspecialisten te laten deelnemen aan het beoordelingsproces. Hierbij kan worden gedacht aan een aanpak die is beschreven in het KNAW-rapport over informaticaonderzoek (KNAW, 2016). Dit is een taak die op het niveau van universiteiten of op boven-universitair niveau kan worden vormgegeven, bijvoorbeeld door regelmatig scholing aan te bieden, bijeenkomsten te organiseren en best practices te belichten. Om te onderstrepen dat het niet om vrijblijvende activiteiten gaat, zou hierbij op instellingsniveau met enige regelmaat een check op kwaliteitsproces en kwaliteitsborging kunnen worden doorlopen.
- Bespreek met de toezichthouder (Autoriteit Persoonsgegevens) complexe casussen en kwetsbare toepassingen bij het (her)gebruik van data, waarbij diverse partijen zijn betrokken (waaronder samenwerking in internationale consortia). Stimuleer, in navolging van de 'Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek' (VSNU, 2005, verwachte update in 2018), een uitwerking van de AVG voor wetenschappelijk onderzoek. Ook kan het opstellen van een discipline-specifieke gedragscode zoals de Code Goed Gebruik (Federa, 2011) uitkomst bieden.

VOOR HET MINISTERIE VAN OCW:

- Neem het initiatief om een overkoepelende infrastructuur tot stand te brengen die de lokale infrastructuur ondersteunt en aanvult bij het verzamelen, verwerken, delen en analyseren van (big) data met betrekking tot personen. Het gaat daarbij om faciliteiten en ondersteuning door dataspecialisten en juridische en ethische expertise die de reikwijdte van individuele instituten of de faculteiten overstijgen. Deze infrastructuur is bedoeld om samenwerking binnen de wetenschappelijke gemeenschap en met partners uit het publieke en private terrein tot stand te brengen. De infrastructuur kan alleen tot stand komen met structurele extra investeringen (NWO, 2017; KNAW, 2016), en kan worden gezien als een nutsvoorziening, die voortbouwt op ervaringen van onder meer het CBS, DANS, BBMRI en Data4LifeSciences en recente nationale initiatieven die zijn opgenomen in de Nationale Roadmap Grootschalige Wetenschappelijke Infrastructuur (NWO, 2016). Het ministerie dient bij het inrichten van die nieuwe infrastructuur in overleg met onder meer NWO, SURF en VSNU met nadruk onderzoek met gegevens over personen te betrekken en aan te sluiten bij initiatieven rond *Open Science*, *Open Data* en *FAIR* data, zoals het internationale GO FAIR-netwerk en het Nationaal Platform Open Science.

# BIJLAGEN

## Bijlage 1. Instellingsbesluit commissie 'Big data'

Het bestuur van de KNAW besluit, gelet op mogelijk verstrekkende implicaties van 'big data' voor wetenschappelijk onderzoek en op kansen en bedreigingen die zich daarbij kunnen voordoen, en tevens gelet op artikel 8 van het *Reglement van de KNAW*, tot het instellen van de commissie 'Big data', hierna te noemen de commissie.

### Artikel 1. Taakopdracht

De commissie verricht haar werkzaamheden vanuit het perspectief dat de Nederlandse wetenschap op het gebied van onderzoek met 'big data' een vooraanstaande positie in het internationale veld kan verkrijgen en het Nederlandse 'big data'-onderzoek internationaal onderscheidend kan zijn. Daarbinnen ligt het accent op wetenschapsgebieden die werken met gegevens over personen.

'Big data' worden omschreven als data die een of meer van de volgende kenmerken hebben:
- Een zeer grote omvang (in terabytes of petabytes) – *volume.*
- Een hoge creatiesnelheid (in, of bijna in real-time) – *velocity.*
- Een grote variatie (waaronder zowel gestructureerde als ongestructureerde data) – *variety*.

Het betreft niet alleen data die voor het verrichten van wetenschappelijk onderzoek worden verzameld, maar ook de – snel groeiende stroom – data die voor andere doelen worden verzameld (denk aan 'twitterdata' en 'consumentendata').

De commissie heeft twee taken:
1. Uitvoeren van een brede verkenning naar effecten van 'big data' op wetenschappelijk onderzoek.
2. Voorbereiden van een KNAW-advies over enkele geselecteerde onderwerpen.

*1.1. Verkenning*

De eerste taak is het uitvoeren van een brede verkenning naar de effecten van 'big data' op wetenschappelijk onderzoek. De volgende vragen kunnen daarbij leidend zijn:
1. Op welke wijze en in welke mate verandert wetenschappelijk onderzoek als gevolg van 'big data'?
2. Wat zijn kansen en bedreigingen voor de toepassing van 'big data' in het Nederlands wetenschappelijk onderzoek?
3. Hoe kunnen kansen worden benut en de bedreigingen het hoofd worden geboden, en onder welke voorwaarden kan dergelijk onderzoek worden gefaciliteerd?
4. Waar liggen de grenzen van de toepassing van 'big data' voor wetenschappelijk onderzoek en onder welke voorwaarden is dergelijk onderzoek (wettelijk) gelegitimeerd?
5. Is de kwaliteit van 'big data' toereikend voor wetenschappelijk onderzoek en welke kwesties behoeven aandacht om de kwaliteit van 'big data' te garanderen?

*1.2. Advies*

De tweede taak is het voorbereiden van een KNAW-advies over enkele geselecteerde onderwerpen. Aangezien 'big data' een zeer diffuus terrein betreffen, maakt de commissie op basis van haar eigen verkenning een goed beargumenteerde keuze van adviesonderwerpen – vooral gelet op kansen en bedreigingen.

*1.3. Werkwijze*

Voor de eerste taak kan de commissie zich mede baseren op een aantal te organiseren oriënterende bijeenkomsten. De eerste bijeenkomst kan reeds bestaande nationale en internationale 'big data'-initiatieven vanuit de wetenschap en het bedrijfsleven in beeld brengen met een focus op de technische aspecten van 'big data'. Deze bijeenkomst legt de basis voor enkele wetenschapsinhoudelijke bijeenkomsten waaraan kan worden bijgedragen vanuit wetenschapsgebieden die werken met gegevens over personen en instellingen op het gebied van 'big data'.

## Artikel 2. Samenstelling en instellingsduur

Tot leden van de commissie worden op persoonlijke titel benoemd:
- Voorzitter: prof. dr. C.W.A.M. (Kees) Aarts (*Political science*, Universiteit Twente[6])

6 Prof. dr. Kees Aarts werkt sinds begin 2016 als hoogleraar Politieke instituties en gedrag en als decaan van de Faculteit Gedrags- en Maatschappijwetenschappen bij de Rijksuniversiteit Groningen.

Overige leden:

- Prof. dr. M.J. (Martin) Dijst, *Spatial mobility and urban development*, Universiteit Utrecht
- Prof. dr. J.S.H. (Johan) van Leeuwaarden, Wiskunde, Technische Universiteit Eindhoven)
- Prof. dr. G.J.B. (Gert-Jan) van Ommen, Humane genetica, LUMC
- Prof. mr. J.E.J. (Corien) Prins, Recht en informatisering, Tilburg University
- Prof. dr. M. (Maarten) de Rijke, *Information processing and internet*, Universiteit van Amsterdam
- Prof. dr. A. (Arjen) van Witteloostuijn, Economie en management, Tilburg University[7])
- Prof. dr. S. (Sally) Wyatt, *Digital Cultures*, Universiteit Maastricht

De commissie wordt ingesteld tot de aanbieding van het rapport aan het bestuur van de KNAW, binnen een jaar na de eerste commissievergadering[8].

Namens het bestuur is prof. dr. J.F.T.M. (José) van Dijck, president KNAW, agenda-lid van de commissie[9].

De commissie wordt ondersteund vanuit het bureau van de KNAW door een secretariaat[10]. Als ambtelijk secretaris van de commissie wordt benoemd: dr. A. (Ans) Vollering (senior beleidsmedewerker).

## Artikel 3. Kwaliteitsbeheer

De leden van de commissie hebben voordat zij zijn benoemd, kennis genomen van de code ter voorkoming van oneigenlijke beïnvloeding door belangenverstrengeling en de verklaring daarvan ingevuld en geretourneerd, voorafgaand aan de eerste vergadering van de commissie.

De leden van de commissie hebben kennis genomen van de *Handleiding adviezen en verkenningen* van de KNAW die op 21 mei 2013 is vastgesteld door het bestuur van de KNAW. Het beleid omtrent *review* is beschreven in bijlage A van deze handleiding. Van dit beleid wordt niet afgeweken.

7 Prof. dr. Arjen van Witteloostuyn is sinds begin 2018 hoogleraar Business and economics aan de School of Business and Economics (SBE) van de Vrije Universiteit Amsterdam. Hij is ook decaan van SBE.

8 De instellingstermijn van de commissie is verlengd tot februari 2018.

9 In 2016 is prof. dr. José van Dijck als agenda-lid van de commissie opgevolgd door prof. dr. mr. Marc Groenhuijsen, bestuurslid KNAW.

10 Het schrijfteam, dat bestaat uit prof. dr. Kees Aarts en dr. Ans Vollering, is in 2017 uitgebreid met dr. Salome Scholtens, programma manager, Dept. of Human Genetics, UMCG.

### Artikel 4. Nazorg en communicatie

De commissie besteedt aandacht aan de nazorg en communicatie rondom haar bevindingen.

### Artikel 5. Kosten en vergoedingen

De leden ontvangen op basis van art. 18 lid 2 van het *Reglement van de KNAW* een vergoeding voor de gemaakte reiskosten.

### Artikel 6. Geheimhouding

De leden nemen geheimhouding in acht ten aanzien van alle informatie die in het kader van de uitvoering van haar taak als vertrouwelijk is aan te merken.

Aldus vastgesteld door het bestuur van de Koninklijke Nederlandse Akademie van Wetenschappen op 14 september 2015 te Amsterdam.
Namens het bestuur van de KNAW,
Mr. Mieke Zaanen
Algemeen directeur

## Bijlage 2. Geconsulteerde personen (stand van zaken januari 2018)

Deelnemers aan de focusgroepbijeenkomst met toepassingsgerichte onderzoeksgroepen op 16 maart 2016:

- Prof. dr. Jaap Abbring, Hoogleraar Econometrics, TiU
- Prof. dr. Annelien Bredenoord, Hoogleraar Ethiek van Biomedische Innovatie, UMCU
- Prof. dr. Wilco Hazeleger, Hoogleraar Climate Dynamics, WUR
- Prof. dr. Natali Helberger, Hoogleraar Informatierecht, UvA
- Prof. dr.ir. Dick den Hertog, Hoogleraar Econometrics and Operations Research, TiU
- Prof. dr. Barend Mons, Hoogleraar Biosemantics, LUMC
- Prof. dr. Brenda Penninx, Hoogleraar Psychiatric Epidemiology, VUMC

Deelnemers aan de focusgroepbijeenkomst met ict-gerelateerde onderzoeksgroepen op 15 juni 2016:

- Prof. dr. Geert-Jan Houben, Hoogleraar Web Information Systems, TUD
- Dr. René Penning de Vries, Boegbeeld ICT
- Prof. dr. Jan-Willem Romeijn, Hoogleraar Philosophy of Science, RUG
- Dr. Marco Spruit, Onderzoeker organisatie en informatie, UU
- Prof. dr. Maarten van Steen, Hoogleraar Grootschalige gedistribueerde computersystemen, UT

- Mr. Wilbert Tomesen, Vicevoorzitter Autoriteit Persoonsgegevens

Deelnemers aan de klankbordgroep met onderzoekers uit diverse wetenschapsgebieden op 4 november 2016:
- Dr. Sanne Abeln, UHD, Faculty of Science, Bioinformatics, VU
- Jasper Bovenberg, JD, LLM, PhD, Advocaat en jurist
- Dr. Hannes Datta MSc, UD, TS of Economic and management, TiU
- Dr. Claudia Hauff, UD, Web information systems group, TUD
- Dr. Willem Herder, Co-founder Pacmed
- Dr. Peter-Bram 't Hoen, UHD, Department of Human genetics, LUMC
- Prof. dr. Maurits Kaptein PDEng, bijzonder hoogleraar data science & health, en UD, TS Social and behavioral sciences, TiU
- Prof. dr. Martha Larson, Hoogleraar Multimedia Information Technology, RU
- Dr. Annika Richterich, UD, Faculty of Arts & Social Sciences, UM
- Dr. Jeroen de Ridder, UHD, Section Genetics, UMCU
- Dr. Mirko Schäfer, UD, Faculteit Geesteswetenschappen, UU
- Dr. Katrijn Van Deun, UHD, TS Social and behavioral sciences, TiU
- Dr. Janneke van der Zwaan, eScience Research Engineer, Netherlands eScience Center

## Bijlage 3. Reviewprocedure

Een conceptversie van dit rapport is conform de *Handleiding Adviezen KNAW* beoordeeld door:
- Prof. dr. Annelien Bredenoord, Hoogleraar Ethiek van Biomedische Innovatie, UMCU
- Prof. dr. Wilco Hazeleger, Hoogleraar Climate Dynamics, WUR
- Prof. dr. Jan Willem Klop, Emeritus hoogleraar Toegepaste logica, VU
- Dr. Huib van de Stadt, Hoofddirecteur Sociaaleconomische en Ruimtelijke Statistieken CBS

Daarnaast is die versie becommentarieerd vanuit de Raad voor Geesteswetenschappen, de Raad voor Medische Wetenschappen, de Raad voor Natuur- en Technische Wetenschappen, en de Sociaal-Wetenschappelijke Raad.

De KNAW is de beoordelaars en de leden van de adviesraden veel dank verschuldigd. Zij dragen geen verantwoordelijkheid voor de inhoud van het rapport.

# SELECTIE GERAADPLEEGDE LITERATUUR


Bader, M.D.M., Mooney, S.J., Rundle, A.G., Protecting Personally Identifiable Information When Using Online Geographic Tools for Public Health Research, *American journal of public health*,106(2):206-208, 2016 (https://www.nature.com/articles/nbt.3918?WT.ec_id=NBT-201707&spMailingID=54477982&spUserID=ODkwMTM2NjI1NQS2&spJobID=1201876024&spReportId=MTIwMTg3NjAyNAS2)

Barbaro, M., Zeller, T., jr, A Face Is Exposed for AOL Searcher No. 4417749, *The New York Times*, 2006 (http://www.nytimes.com/2006/08/09/technology/09aol.html)

BBMRI-NL, *De donor als partner. Hoe patiënt en publiek te betrekken bij besluitvorming over biobanken en registraties*, 2014 (https://www.bbmri.nl/wp-content/uploads/2015/10/richtsnoerned_def.pdf)

Bell, G., Hey, T., Szalay, A., Beyond the data deluge, *Science*, 06;323(5919):1297-1298, 2009 (http://www.uvm.edu/pdodds/files/papers/others/2009/bell2009a.pdf)

boyd, d., Crawford, K., Critical questions for big data, *Information, Communication & Society*,15(5):662-679, 2012 (https://people.cs.kuleuven.be/~bettina.berendt/teaching/ViennaDH15/boyd_crawford_2012.pdf)

EOSC, Science Business, Governing the European Open Science Cloud, 2017 (https://sciencebusiness.net/system/files/reports/Governing%20the%20European%20Open%20Science%20Cloud%20v5.pdf)

Europese Commissie, Digital Single Market. Making Big Data work for Europe, 2015 (https://ec.europa.eu/digital-single-market/en/making-big-data-work-europe)

Europese Commissie, Sustainable research infrastructure. A call for action, Commission staff working document, 2017 (https://www.google.nl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0ahUKEwiK9Nvik_rXAhULEVAKHbJyCV0QFgg5MAI&url=https%3A%2F%2Fec.europa.eu%2Fresearch%2Finfrastructures%2Fpdf%2Fri_policy_swd-infrastructures_2017.pdf&usg=AOvVaw22JrU2Ov9KkMdscf3HfMRS)

Federa, Code Goed Gebruik. De Gedragscode Verantwoord omgaan met lichaamsmateriaal ten behoeve van wetenschappelijk onderzoek, 2011 (https://www.federa.org/code-goed-gebruik)

Gartner Group, Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data, *Gartner Group press release*, 2011 (https://www.gartner.com/newsroom/id/1731916)
Hey, T., Tansley, S., Tolle, K., *The Fourth Paradigm. Data-Intensive Scientific Discovery*, 2009 (https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/M091000H.pdf)
Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W., Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays, PLOS genetics, 2008 (https://doi.org/10.1371/journal.pgen.1000167)
Jagadish, H.V., Big Data and Science: Myths and Reality, *Big Data Research*, 2(2):49-52, 2015 (https://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_152501.pdf)
Jin, X., Wah, B.W., Cheng, X., Wang, Y., Significance and Challenges of Big Data Research, *Big Data Research*, 2(2):59-64, 2015 (https://dl.acm.org/citation.cfm?id=2991439)
Kitchin, R., Big Data, new epistemologies and paradigm shifts, *Big Data & Society*,1(1):205395171452848, 2014 (https://people.cs.kuleuven.be/~bettina.berendt/teaching/ViennaDH15/kitchin_2014.pdf)
KNAW, Ethische en juridische aspecten van informatica-onderzoek, 2016 (https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/20160919-advies-ethische-en-juridische-aspecten-van-informaticaonderzoek-web)
KNAW, Replication studies. Improving reproducibility in the empirical sciences, 2018 (https://knaw.nl/nl/actueel/publicaties/resolveuid/3ba19549eab047fcbd8daf8d43748c05)
Landelijk Coördinatiepunt Research Data Management, Vuistregel gegevensbescherming en privacy, 7 juni 2016 (https://www.lcrdm.nl/binaries/content/assets/subsites-diensten/lcrdm/juridisch/vuistregel-gegevensbescherming-en-privacy-201600607_juridische-werkgroep-landelijk-coordinatiepunt-research-data-management.pdf)
Leonelli, S., Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems, *Philos Trans A Math Eng Sci*, 2016 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124067/)
Levallois, C., Steinmetz, S., Wouters, P., Sloppy Data Floods or Precise Social Science Methodologies? Dilemmas in the Transition to Data-Intensive Research in Sociology and Economics, p. 151-182, in: Wouters, P. et al (eds), Virtual Knowledge: experimenting in the humanities and the social sciences, MIT Press, 2013 (http://thebook.virtualknowledgestudio.nl/table-of-contents/chapter-6)
Metcalf, J., Crawford, K., Where are human subjects in Big Data research? The emerging ethics divide, *Big Data & Society,* 1-14, 2016 (http://journals.sagepub.com/doi/pdf/10.1177/2053951716650211)
Ministerie van OCW, Kamerbrief over big data in onderwijs en wetenschap. Inventarisatie en essays, 5 juni 2015 (https://www.rijksoverheid.nl/documenten/kamerstukken/2015/06/05/kamerbrief-over-big-data-in-onderwijs-en-wetenschap-inventarisatie-en-essays)
Ministerie van OCW, Kamerbrief over big data in onderwijs, cultuur en wetenschap, 28 juni 2016 (https://www.rijksoverheid.nl/documenten/kamerstukken/2016/06/28/kamerbrief-over-big-data-in-onderwijs-cultuur-en-wetenschap)
Moerel L., Prins, J.E.J., Privacy for the Homo Digitalis: Proposal for a New Regulatory Framework for Data Protection in the Light of Big Data and the Internet of Things, in: *Homo Digitalis, Preadviezen Nederlandse Juristen-Vereniging*, Kluwer 2016, pp. 11-12 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784123)

Mostert, M., Bredenoord, A.L., Biesaart, M.C., Van Delden, J.J., Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach, *European journal of human genetics*, 24(7):1096, 2016 (https://www.ncbi.nlm.nih.gov/pubmed/26554881)

Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie Du Sert, N., et al., A manifesto for reproducible science, *Nature Human Behaviour*, 2017 (https://www.psy.ox.ac.uk/publications/682604)

Munevar, S., Unlocking Big Data for better health, *Nature Biotechnology*, 35(7):684, 2017 (https://www.nature.com/articles/nbt.3918?WT.ec_id=NBT-201707&spMailingID=54477982&spUserID=ODkwMTM2NjI1NQS2&spJobID=1201876024&spReportId=MTIwMTg3NjAyNAS2)

Nationaal Plan Open Science, WWW-document, 2017 (https://www.openscience.nl/binaries/content/assets/subsites-evenementen/open-science/nationaalplanopenscience_nl.pdf)

*Nederlandse Wetenschapsagenda, Toegankelijke en verantwoorde waardecreatie uit big data,* 2016 (https://wetenschapsagenda.nl/wp-content/uploads/2016/06/09-Big-Data-verantwoord-gebruiken-Game-Changer-31052016.pdf)

NSF (National Science Foundation), WWW-document, 2015 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1244691&HistoricalAwards=false)

NWO, Nationale Roadmap Grootschalige Wetenschappelijke Infrastructuur, 2016 (https://www.nwo.nl/documents/nwo/permanente-commissie/roadmap-grote-onderzoeksfaciliteiten)

NWO, Topwetenschap vereist topinfrastructuur. Adviesrapport nationale digitale infrastructuur Permanente Commissie voor Grootschalige Wetenschappelijke Infrastructuur. ICT-subcommissie, 2017 (https://www.nwo.nl/documents/ew/adviesrapport-nationale-digitale-infrastructuur-juni-2017)

Pasquetto, I.V., Randles, B.M., Borgman, C.L., On the Reuse of Scientific Data, *Data Science Journa*l, 2017 (https://works.bepress.com/borgman/409/)

Rathenau Instituut, Opwaarderen: borgen van publieke waarden in de digitale samenleving, 2017 (https://www.rathenau.nl/nl/publicatie/opwaarderen-borgen-van-publieke-waarden-de-digitale-samenleving)

Richardson, D.B., Kwan, M-P, Alter, G., McKendry, J.E., Replication of scientific research: addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research, *Annals of GIS*, 21, 101-110, 2015 (http://www.tandfonline.com/doi/abs/10.1080/19475683.2015.1027792)

Rieder, B., Röhle, T., Digital Humanities: Five Challenges, in: D. M. Berry (Ed.), *Understanding digital humanities* (pp. 67–84), Houndmills: Palgrave Macmillan, 2012 (http://www.palgrave.com/us/book/9780230292642)

Stieb, D.M., Boot, C.R., Turner, M.C., Promise and pitfalls in the application of big data to occupational and environmental health, *BMC Public Health*, 17:372, 2017 (https://bmcpublichealth.biomedcentral.com/track/pdf/10.1186/s12889-017-4286-8?site=bmcpublichealth.biomedcentral.com)

Team ICT, COMMIT2DATA, *White Paper. Proposal for a National Public-Private Research and Innovation Program on Data Science, Stewardship and Technology Across Top Sectors,* 2015 (http://www.bdvc.nl/images/Rapporten/COMMIT2DATA.pdf)

Tockar, A., *Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset*, Neustar Research, 2015 (https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/)

UU (Universiteit Utrecht), WWW-document, 2017 (https://www.uu.nl/en/research/sustainability-healthy-urban-living)

URISA (Urban and Regional Information Systems Association), A GIS Code of Ethics. WWW-document, 2003 (http://www.urisa.org/about-us/gis-code-of-ethics/)

VSNU, *Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek,* 2005 (http://www.vsnu.nl/files/documenten/Domeinen/Accountability/Codes/Gedragscode%20persoonsgegevens.pdf)

VSNU, *De Digitale Samenleving. Nederland en zijn universiteiten: internationale pioniers in mensgerichte informatietechnologie*, 2016 (http://www.vsnu.nl/files/documenten/Publicaties/VSNU_De_Digitale_Samenleving.pdf)

WHO (World Health Organization), WWW-document, 2017 (http://www.euro.who.int/en/health-topics/environment-and-health/urban-health/activities/healthy-cities)

Wilkinson, M.D., Dumontier, M., Aalbersberg, IJ.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J-W., Silva, Santos L.B. da, Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.C. 't, Hooft, R., Kuhn, T., Kok, R., Kok, J.N., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., Schaik, R. van, Sansone, S-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Lei, J. van der, Mulligen., E. van, Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K.J., Zhao, J., The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data*, 2016 (https://www.nature.com/articles/sdata201618)

WRR, *The public core of the Internet. An international agenda for Internet governance* (rapport nr. 94) 2015 (https://english.wrr.nl/publications/reports/2015/10/01/the-public-core-of-the-internet)

WRR, *Big Data in een vrije en veilige samenleving* (rapport nr. 95) 2016 (https://www.wrr.nl/publicaties/rapporten/2016/04/28/big-data-in-een-vrije-en-veilige-samenleving)