



Big data en (breedte)sport

Kansen voor sportonderzoek

Remko van den Dool

Paul Hover

Reinout Moes

Big data en (breedte)sport

Kansen voor sportonderzoek

Remko van den Dool
Paul Hover
Reinout Moes

Mulier Instituut
sociaal-wetenschappelijk sportonderzoek

*Postbus 85445 | 3508 AK Utrecht
Herculesplein 269 | 3584 AA Utrecht
+31 (0)30 721 02 20 | www.mulierinstituut.nl
info@mulierinstituut.nl | [@mulierinstituut](https://twitter.com/mulierinstituut)*

Inhoud

Samenvatting	7
1. Inleiding	9
2. Big data: een overzicht	11
2.1 Groot in belofte	11
2.2 Omschrijving big data	12
2.3 Facetten big data	14
2.4 Initiatieven overheden en onderwijsinstellingen	16
2.5 Kansen en beperkingen	16
3. Sportonderzoek en big data	21
3.1 Persoonlijke meters en apps	22
3.2 Uitingen via media	25
3.3 Dataregisters voor sportonderzoek	30
3.4 Openbare observaties	32
4. Combinatie bronnen big data	35
5. Reflectie: betekenis big data voor sport	39
5.1 Methodische implicaties big data voor sportonderzoek	39
5.2 Kansen big data voor sportonderzoek	41
5.3 Big data ambities Mulier Instituut	43
Bijlage I. Hulpmiddelen verzamelen big data	45
Bijlage II. Voorbeeldstudie: voetbal en ongewenst gedrag	47
Bijlage III. Enquêtedata voor sportonderzoek	51
Literatuur	53

Samenvatting

In dit rapport heeft het Mulier Instituut gepoogd om meer duidelijkheid te scheppen over de kansen van big data voor het sportonderzoek. Het Mulier Instituut heeft dat gedaan vanuit de constatering dat klassieke dataverzamelmethode hun beperkingen kennen en dat big data vaak in het nieuws zijn, zonder dat altijd helder is hoe deze nieuwe wijze van data verzamelen zich verhoudt tot meer traditionele wijzen van dataverzameling. Voor het project is de beschikbare literatuur geraadpleegd, zijn interviews gehouden met een beperkt aantal experts en is een casestudie uitgevoerd rond ongewenst gedrag in het (Engelse) voetbal. De focus in het project lag op het onderzoek in de breedtesport.

Samenvattend stellen we dat big data als bijzonder betekenisvol gelden voor de verdere doorontwikkeling van het (breedte-)sportonderzoek. Tegelijk signaleren we ook dat ook big data duidelijke grenzen kennen. Zo zijn er grote vragen te stellen over representativiteit van data, en is de verkrijgbaarheid van data een thema. Er zijn dikwijls meerdere, complexe bewerkingsslagen nodig om tot resultaten te komen. Zonder contextuele en theoretische kennis is ook bij big data geen betekenis aan de uitkomsten van het onderzoek te ontleen. Als Mulier Instituut zien we de kracht van big data vooral waar dergelijke gegevens worden ingezet in combinatie met andere, meer traditionele dataverzamelingen (zoals een enquête).

Het ontbreekt voornamelijk aan een krachtige en eenduidige definitie van het begrip big data. Het Mulier Instituut heeft daarop vier bronnen van big data onderscheiden:

- **Persoonlijke meters:** (*beweeg*)apps en draagbare meters (*wearables*) zijn waardevolle methoden om data over bijvoorbeeld inspanning te verzamelen, zeker in combinatie met aanvullende informatie via een vragenlijst. Of dit ook geldt voor (big) data die buiten een onderzoekssetting zijn verkregen, blijft de vraag. Het gebruik lijkt daarvoor te selectief.
- **Media:** voor *sociale media*-analyses zijn alleen Twitterberichten in voldoende mate beschikbaar. Facebook is veel minder als bron van (ongestructureerde) big data beschikbaar, omdat pagina's worden afgeschermd voor de buitenwereld. De vraag is daarmee of sociale media als bron voor (sport)onderzoek een rol zal spelen. Voor *onderzoek naar traditionele media* lijken de mogelijkheden door de big data revolutie sterk verruimd. Door het internet kost het verzamelen en analyseren van bronnen, niet alleen tekst, ook video en geluid, veel minder tijd. Dat biedt tal van nieuwe kansen voor (geautomatiseerde) technieken voor inhoudsanalyse.
- **Dataregisters:** *Dataregisters* vormen steeds vaker een onmisbare bron voor onderzoekers en bieden ongekennde mogelijkheden voor nadere analyses, vooral door koppeling aan andere bronnen zoals enquêtes een rol spelen. Hoewel veel dataregisters (deels) openbaar zijn, vormt de toegankelijkheid van registers vanwege commerciële belangen en privacybescherming in toenemende mate een thema.
- **Openbare observaties:** Een speciale categorie van big data hebben we *observaties (in de openbare ruimte)* genoemd. Dit kan gaan om het oppikken van signalen (wifi, bluetooth etc.) maar ook de analyse van gedrag met camerabeelden. Een voorbeeld hiervan is het tellen van evenementenbezoekers via smartphone-abonnementen of via camerabeelden.

In de aankomende jaren zal het Mulier Instituut verdere investeringen plegen in het werken met big data. De ambitie van het instituut is om in 2020 voldoende veelsoortige ervaring met big data te hebben opgedaan om op projecten waar dat relevant voor is, big data zinvol te kunnen inzetten.

1. Inleiding

Steeds vaker komen data beschikbaar. Zoals bij het gebruik van sociale media of het kopen van artikelen in een webshop. Deze informatie is bruikbaar voor diverse sectoren van de maatschappij. Er is een sterk geloof dat deze zogenaamde big data een rol kunnen spelen bij vraagstukken zoals het verbeteren van het leefklimaat in steden en dorpen. Qua ontwikkeling bevinden big data zich nu in een ‘goudkoortsfase’ (Kool et al., 2015) waarbij de verwachtingen hooggespannen zijn en met big data volop wordt geëxperimenteerd.

Big data kunnen ook voor sport en bewegen belangrijk worden. In de Nationale Kennisagenda Sport en Bewegen spelen big data een grote rol bij de drie thema’s: ‘Beter presteren’, ‘Een leven lang bewegen’ en ‘De waarde(n) van sport’ (NWO, 2016). Daarnaast heeft het Topteam Sport het initiatief genomen voor een Sport Data Valley (www.sportdatavalley.nl). Hiermee willen zij de uitwisseling van data die relevant zijn voor sport bevorderen.

In voorliggend rapport gaan we in op de betekenis van big data voor het (breedte)sportonderzoek. Het Mulier Instituut schrijft dit rapport om sportonderzoekers te inspireren en zelf een leidraad te hebben voor big data in het sportonderzoek. In het rapport bespreken we enkele concrete mogelijkheden voor het gebruik van big data. Daarbij verliezen we de lange termijn beloften van big data niet uit het oog. We realiseren ons dat gezien de snelle ontwikkelingen van big data een regelmatige herziening van de inhoud van dit rapport gewenst is.

De vragen die we voor dit rapport stellen zijn als volgt:

1. Hoe benut het sportonderzoek in Nederland momenteel big data?
2. Welke kansen zijn er voor big data en sportonderzoek?
3. Welke grenzen zijn er voor big data in sportonderzoek?

Afgeleide vragen zijn of big data de bestaande methoden (deels) kunnen vervangen en wat de validiteit is van informatie afkomstig uit big data.

Onze dank gaat uit naar prof. Steven Vos van Fontys Sporthogeschool en TU Eindhoven en Jelte Timmer (MA) van de WRR voor hun commentaar op ons conceptrapport. Verder danken wij prof. Aarnout Brombacher van TU Eindhoven en dr. Laura Jonker van KNVB voor hun visie op de bruikbaarheid van big data voor sportonderzoek.

Leeswijzer

In het volgende hoofdstuk schetsen wij de contouren van big data in het algemeen. We gaan in op de sectoren waar de invloed in de toekomst het grootst zal zijn. Vervolgens werken we voor vier big data bronnen de stand van zaken voor sportonderzoek uit. Deze bronnen zijn gegevens van (beweeg)meters/apps, (sociale) media, dataregisters van overheid en bedrijven en informatie dat afkomstig is van observaties, onder andere in de openbare ruimte. In hoofdstuk 4 bespreken we de combinatiemogelijkheden van big data bronnen. Tot slot reflecteren we op de kansen en beperkingen voor sportonderzoek van big data en beschrijven we de big data ambities van het Mulier Instituut.

2. Big data: een overzicht

Grote hoeveelheden data worden tegenwoordig onophoudelijk en overal gegenereerd. Deze data bieden perspectief voor kennisontwikkeling en beleid voor publieke en private organisaties. Maar hoe beloftevol is deze ontwikkeling volgens experts, wat zijn big data en wat zijn kansen, beperkingen en randvoorwaarden bij de toepassing van big data? Eerst gaan we in dit hoofdstuk in op de toekomstverwachtingen (2.1). Vervolgens beschrijven we wat met big data worden bedoeld (2.2) en uit welke facetten big data bestaan (2.3). Daarna gaan we in op initiatieven van organisaties op dit gebied (2.4) en passeren de kansen en beperkingen de revue.

2.1 Groot in belofte

De laatste jaren is sprake van een ware data-explosie: de hoeveelheid data is met grote sprongen gegroeid. De drijvende kracht achter deze grote groei is de digitalisering (of dataficatie) van de samenleving, een maatschappij die altijd ‘aan staat’ en waarin voortdurend met grote snelheid gegevens worden verkregen. Door het wijd verbreide gebruik van smartphones, OV-chipkaarten, camera’s in de openbare ruimte, pinbetalingen, websitebezoeken, online aankopen, allerlei apps en sociale media door consumenten, worden bewust of onbewust grote hoeveelheden data digitaal gegenereerd en opgeslagen.

De voortdurend groeiende omvang van data is van een andere orde dan tien jaar geleden; tegenwoordig gaat dat veel sneller. Ook de reikwijdte van de digitalisering is toegenomen: in vrijwel alle sectoren is deze ontwikkeling actueel (van landbouw, industrie tot logistiek en publieke dienstverlening), niet in de laatste plaats in de sport. Het is dan ook niet verwonderlijk dat experts reppen over ‘*the era of the information age*’ (Jin et al., 2015), ‘*the age of big data*’ en over ‘*the datafication of everything*’ (Millington & Millington, in voorbereiding). Een reeks cijfers illustreren de snelheid waarin data worden gegenereerd. Experts schatten dat dagelijks 2,5 miljard gigabytes aan data worden gecreëerd. Daarmee kunnen per minuut 27.000 iPads gevuld worden (VSO, 2014). Als gevolg van deze processen dijt de hoeveelheid big data met de seconde verder uit. Verwacht wordt dat de groei van big data de komende decennia exponentieel zal toenemen. Er is genoeg reden om aan te nemen dat we aan het begin van een nieuw datatijdperk staan.

Big data zijn niet alleen ‘big’ in termen van omvang en aanhoudende groei, maar ook in hun belofte (IBM, 2011). Het gonst van de kansen die big data zouden bieden om kennis te produceren. Experts menen dat big data een toenemende impact zullen hebben op de manier van leven en werken. Big data kunnen in de toekomst uitgroeien tot een belangrijke bron van concurrentievoordeel in het bedrijfsleven en ook overheden kunnen er de vruchten van plukken (Bloem et al., 2012). Volgens Jin et al. (2015) zal de mate waarin big data succesvol worden benut, ook worden beschouwd als een indicator voor de ‘kracht’ van een land waarmee aanzien en een sterke reputatie verworven kan worden, net als indicatoren voor welvaart en economie. Peter Sondergaard van Gartner, een leidend onderzoeks- en adviesbureau in de informatietechnologiesector, verwoordt de waarde als volgt: “*Information is the oil of the 21st century, and analytics is the combustion engine*” (Gartner, 2011).

Naast het geloof in de potentie van big data voor onderzoek en beleid, zijn er ook vragen en uitdagingen. Eén van die uitdagingen is om in die gigantische hoeveelheid data de weg te vinden en (en deel van) het antwoord te destilleren op prangende vragen, bijvoorbeeld voor beleidsvorming. John Naisbitt, gerenommeerd (IT-)futuuroloog, zegt het treffend: “*We have for the first time an economy based on a key resource [information] that is not only renewable, but self-generating. Running out of it*

is not a problem, but drowning in it” (Van den Belt & Nillesen, 2014). In paragraaf 2.5 wordt hier nader op ingegaan.

Omdat we aan het begin van de big data ontwikkeling staan, is er geen scherp zicht op de mogelijkheden die big data bieden en hoe big data te analyseren zijn. Opzienbarend is dat niet, zoals nu big data een relatief nieuw fenomeen zijn, zo was het internet dat begin jaren '90. Destijds was duidelijk dat internet tot grote veranderingen en nieuwe mogelijkheden zou leiden. Maar hoe snel het internet zich zou verbreiden, welke veranderingen aanstaande waren en hoe mogelijkheden benut konden worden, was toen ook nog niet duidelijk. Rond big data bevinden we ons nu in een vergelijkbare situatie, waarbij volop verkend en geëxperimenteerd wordt.

2.2 Omschrijving big data

Wat zijn big data?

De term big data wordt al ruim tien jaar gebezigd en tot op heden is er geen eenduidig en breed gedragen definitie van wat big data zijn (o.a. Manyika et al., 2011). Het is alles wat geen *small data* zijn. Volgens McKinsey Global Institute, een befaamd management- en consultancybedrijf, verwijst big data naar “*datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze*” (Manyika et al., 2011). In het geval van small data zijn deze zaken wel mogelijk. Een meer theoretische definitie van big data is die van Mayer- Schönberger en Cukier die stellen dat big data betrekking hebben op het vermogen van de samenleving om informatie op nieuwe manieren in te zetten voor het verkrijgen van nuttige inzichten of waardevolle goederen en diensten (Bongers et al., 2015).

De WRR (2016) hanteert geen scherp omlijnde definitie van big data, maar ziet big data als een samenspel van ontwikkelingen en niet zozeer als een vastomlijnd en definieerbaar gegeven. Wel onderscheidt de WRR (2016) drie kenmerken van big data:

1. Data: het gaat om grote hoeveelheden gestructureerde en ongestructureerde data uit verschillende bronnen;
2. Analyse: de analyse is ‘data-driven’ en zoekt geautomatiseerd naar correlaties;
3. Gebruik/toepassing: de analyses moeten leiden tot ‘*actionable knowledge*’ (ingrepen in de realiteit op basis van bestandsanalyses).¹

Opvallend aan de verschillende definities is dat gestructureerde data - data met een duidelijke structuur en betekenis (zoals een dataset met rijen en kolommen) - door McKinsey niet tot big data worden gerekend en door de WRR (2016) en Bongers et al. (2015) wel. De definitie van McKinsey kan als ‘eng’ worden beschouwd en die van de andere twee bronnen als relatief ‘ruim’. De indruk leeft dat in de praktijk meestal wordt uitgegaan van de ‘ruime’ definitie en dat ook gestructureerde data tot big data worden gerekend. Voorbeelden van gestructureerde data zijn data in rijen en kolommen, zoals in Excel- en SPSS-bestanden. Tevens is het van belang dat de drie bronnen eensluidend zijn over het feit

¹ Op analyse en gebruik wordt in de volgende paragraaf nader ingegaan.

dat big data niet verwijzen naar data sec, maar ook naar een proces van gegevensverzameling (en opslag), analyse en gebruik.

Aspecten van big data

Big data worden in de literatuur getypeerd aan de hand van de zogenaamde V's. Sommige experts halen drie V's aan, anderen meer (o.a. Jin et al., 2015; Millington & Millington; WRR, 2016). De zeven V's zijn:

1. *Variety*: verscheidenheid van data (beeld, tekst, geluid) en bronnen;
2. *Volume*: hoeveelheid;
3. *Velocity*: snelheid waarmee data veranderen;
4. *Veracity*: mate betrouwbaarheid;
5. *Value*: toegevoegde waarde (t.o.v. beschikbare data);
6. *Viscosity*: tegenstrijdigheid van data;
7. *Virality*: snelheid waarmee data gedeeld kunnen worden.

Bij big data wordt gemakkelijk gedacht aan grote volumes, maar de V's laten zien dat big data niet alleen over volume gaan. De zeven V's bieden de mogelijkheid om verschillende big data te typeren en op de verschillende aspecten te vergelijken en te beoordelen.

Dat big data een belangrijke rol in de toekomst gaan spelen staat buiten kijf. Dit betekent niet dat kwantitatief onderzoek op basis van gegevensverzameling met steekproeven niet meer van waarde is. Dit type dataverzameling blijft van waarde omdat hiermee antwoord gegeven kan worden op een scherp geformuleerde informatiebehoefte, op een specifiek moment en onder een specifieke doelgroep.

Categorisering

Big data kunnen op zes manieren worden gecategoriseerd, zoals tabel 2.1 laat zien. Ten eerste is er het onderscheid gestructureerd-ongestructureerd. Een database met gegevens van klanten die in rijen en kolommen is opgebouwd, is een voorbeeld van een gestructureerde bron, een voorbeeld van een ongestructureerde bron is een brede waaier van zoektermen die in een online zoekmachine zijn ingetypd. Ten tweede zijn er interne en externe bronnen. Interne bronnen verwijzen naar data die binnen de eigen organisatie voorhanden zijn; data die buiten de eigen organisatie zijn te verkrijgen zijn externe data. Ten derde zijn er big data op papier en in digitale vorm. Ofschoon er tegenwoordig veel data digitaal beschikbaar zijn, komt het voor dat (oude) gegevens op papier beschikbaar zijn. Dit betreft bijvoorbeeld informatie in historische archieven. Ten vierde zijn er openbare en niet openbare data. Openbare data zijn voor iedereen eenvoudig toegankelijk (dit wordt ook wel aangeduid met open data) en niet-openbare data zijn data die niet gedeeld worden door een organisatie (o.a. defensie, politie, data van bedrijven met commerciële waarde). Ten vijfde kunnen data bewust en onbewust verzameld zijn. Bewust verzamelde data betreffen registraties die door de betreffende personen zelf worden gedaan (o.a. geboorteregistratie), terwijl onbewust verzamelde data niet (of minder) actief worden verstrekt (o.a. surfgedrag op internet). Tot slot is er het onderscheid tussen analyseerbare data, ook wel *machine-readable* genoemd (o.a. data die in software ingelezen kunnen worden) en niet (goed) analyseerbare data (zoals data die in PDF-bestanden staan en die niet eenvoudig in software geïmporteerd kunnen worden).

Tabel 2.1 Categorisering van big data op basis van structuur en bron

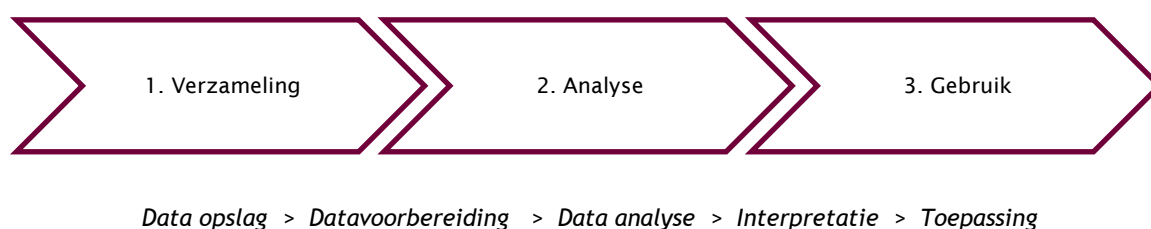
Gestructureerd	Ongestructureerd
Interne bron	Externe bron
Digitaal	Op papier
Openbaar	Niet openbaar
Bewust verzameld	Onbewust verzameld
Analyseerbaar (<i>machine-readable</i>)	Niet analyseerbaar

Deze categorisering laten zien dat big data erg divers zijn en lang niet altijd goed toegankelijk en goed bruikbaar.

2.3 Facetten big data

In de vorige paragraaf werd reeds beschreven dat met big data niet alleen de data zelf bedoeld worden, maar ook de verzameling, analyse en het gebruik ervan. In die zin is big data een ‘werkwoord’. De WRR (2016) onderscheidt bij big data drie hoofdfasen en vijf deelfasen (figuur 2.1).

Figuur 2.1 Fasen van big data



Bron: naar WRR (2016)

De figuur laat zien dat dataverzameling, analyse en gebruik opeenvolgende fasen zijn, maar in de praktijk is dit niet altijd het geval. Na analyse kunnen bijvoorbeeld eerst nieuwe gegevens worden verzameld, die vervolgens ook worden geanalyseerd alvorens te worden gebruikt. Wat houden deze drie fasen in?

Verzameling

Big data worden bewust en onbewust en door menselijke handelingen of door sensoren en apparaten gegenereerd. Menselijke handelingen zijn bijvoorbeeld zoekgedrag op online zoekmachines, een online spel spelen, een online blog schrijven, films op YouTube plaatsen, gebruik van apps en internetbankieren. Er zijn sensoren in tal van apparaten, van auto's, armbanden en horloges tot kleding die allerlei informatie genereren die ergens op een server in de wereld wordt opgeslagen. Sportkleding en accessoires kunnen bijvoorbeeld hartslag, lichaamstemperatuur, ademhalingsfrequentie, locatie (route) en snelheid registreren.

In het kader van privacy is de beveiliging van data van groot belang. Er zijn bekende voorbeelden van cybercrime waarbij hackers aan de hand van ingenieuze software de gegevens van miljoenen personen uit databases wisten te verkrijgen, waaronder paspoortgegevens. Beveiliging is aan de orde bij het

verzamenen, opslaan, uitwisselen en koppelen van data. Hier ligt voor zowel de consumenten zelf, als voor overheden en bedrijven, een opgave.

Analyse

Net als voor de klassieke productiefactoren natuur, arbeid en kapitaal zijn data pas van waarde als deze worden benut: eerst dienen deze te worden geanalyseerd en vervolgens te worden toegepast (DeltIQ Viewpoint, 2013). Kenmerkend voor big data is dat gegevens uit verschillende bronnen in een voorbereidende fase eerst worden geaggregeerd. De data worden als het ware ontschot: dat data uit verschillende sectoren, domeinen of afdelingen gebruikt worden. Big data krijgen pas meerwaarde als ze op een intelligente manier worden gecombineerd en geanalyseerd (NWO, 2016). Dit proces wordt met de term *data science* geduid. Dit impliceert dat big data projecten samenwerking en interactie vereisen - tussen organisaties maar ook binnen (afdelingen van) één organisatie - om ervoor te zorgen dat data niet in afzonderlijke opslagplaatsen bewaard blijven en dat de onderlinge samenhang kan worden bekeken. Doordat grote hoeveelheden data uit uiteenlopende bronnen met elkaar in verband worden gebracht, kunnen in de analysefase nieuwe en onverwachte inzichten ontstaan. Een gebruikte metafoer in deze fase is die van een goudzoeker die in een berg modder op zoek is naar een goudklompje.

De analysemethoden die bij big data projecten worden toegepast zijn niet allemaal nieuw, zoals clusteranalyse en regressieanalyse laten zien. Frequent gebruikte methoden in de analysefase zijn onder meer profiling (patroonherkenning), *machine learning* (het verschijnsel dat computers het vermogen verwerven om iets te doen waarvoor ze niet expliciet zijn geprogrammeerd) en *datamining* (vinden van patronen en correlaties). Er is veel specifieke software die in het kader van big data analyse wordt gebruikt, waaronder Hadoop, Cassandra, MapReduce, Pige, Hive, Hbase en R. Deze software is in staat in snel tempo grote hoeveelheden gegevens te analyseren.

Net als bij andere typen onderzoek, is ook bij big data onderzoek onderscheid tussen beschrijvend onderzoek, verklarend onderzoek en voorspellend onderzoek. Met name het voorspellend onderzoek wordt door experts gezien als de kracht van big data omdat dit effectiviteit en efficiency zou kunnen bevorderen.

Gebruik

Het gebruik of de toepassing van big data heeft implicaties voor beleid en management. Ten eerste kort cyclisch handelen: big data laten zich niet goed rijmen met meerjarige vastomlijnde beleidsplannen (Hiemstra et al., 2014). Dat komt doordat big data organisaties in de gelegenheid stellen om 24 uur per dag ontwikkelingen in de branche en samenleving te monitoren en daarop actie te ondernemen. Ten tweede bevelen experts vaak aan om niet te snel te veel hooi op de vork te nemen en voor de korte termijn geen torenhoge ambities te hebben. De eerste experimentele stappen met big data gedijen in een leeromgeving en bij kortlopende (experimentele) projecten. Niet zelden wordt gestart met het op orde brengen van de interne data, bij veel organisaties is daar nog veel winst te boeken. Ten derde is logischerwijs de vereiste expertise nodig die in de organisatie voldoende draagvlak heeft. Met de opkomst van big data neemt de vraag naar personen met kennis van IT, programmeren en statistiek toe. Er ontstaan nieuwe functies zoals *data scientists*, *data visualisers* en *data virtualisation experts* die ook de beperkingen van big data en de software dienen te kennen. Deze specialistische kennis is de komende jaren schaars (Manyika et al., 2011) en dat zet een rem op innovatie (Kool et al., 2015).

Louter technische kennis volstaat echter niet; voor het formuleren van conclusies en acties is ook contextinformatie nodig.

2.4 Initiatieven overheden en onderwijsinstellingen

Big data staan hoog op de agenda's van publieke en private organisaties. Dat geldt voor Nederland (o.a. WRR, 2016) maar ook voor de Europese Unie (European Commission, 2016) en de Verenigde naties (United Nations, 2014). De EU voert de website *SoBigData* met veel informatie, onder andere van bruikbare databestanden (www.sobigdata.eu).

Ook onderwijsinstellingen investeren in big data. De Universiteit Utrecht heeft de Utrecht Data School opgezet waarbij wordt samengewerkt met overheden en het bedrijfsleven.² In Rotterdam is *The Erasmus Center for Data Science and Business Analytics*³ en in Eindhoven opende de technische Universiteit het *Data Science Center*.⁴ Naast universiteiten zetten ook hogescholen in op big data. Fontys Sporthogeschool kent een lectoraat big data (lector Gerard Schouten)⁵ en de Hogeschool van Amsterdam is partner in het Europese netwerk *Big Data Value Association* (samenwerkingsverband van bedrijven en onderzoeksinstituten).⁶ Daarnaast is er in Rotterdam de kenniswerkplaats *Urban Big Data*.⁷

Voor de sport is het *Dutch Sport Data Center* vermeldenswaardig, waarin de Universiteit Leiden (Leiden *Centre of datascience*), het LUMC, de TU Delft en Amsterdam Institute for Sport Science (AISS) samenwerken. Het centrum doet onderzoek rond de thema's sportprestatie, medische rehabilitatie en gehandicaptensport en sporteconomie.

Tot slot: eind september 2016 is het CBS het *Center for Big Data Statistics* (CBDS) gestart. Binnen het CBDS gaan nationale en internationale partijen uit de overheid, het bedrijfsleven, de wetenschap en het onderwijs samenwerken op het gebied van big data-technologie en big data-methoden voor de productie van officiële statistieken (persbericht CBS, 27 september 2016).

2.5 Kansen en beperkingen

Big data bieden kansen voor publieke en private organisaties om nieuwe kennis te ontwikkelen. Daarnaast zijn er beperkingen: de bruikbaarheid en uitkomsten mogen ook niet worden overschat. De belangrijkste kansen en beperkingen zijn hierna opgesomd en kort toegelicht (zie o.a. WRR, 2016;

² www.dataschool.nl

³ <https://www.erim.eur.nl/centres/data-science-business-analytics/>

⁴ <https://www.tue.nl/en/university/departments/mathematics-and-computer-science/research/research-institutes/data-science-center-eindhoven-dsce/>

⁵ <https://fontys.nl/Over-Fontys/Fontys-Hogeschool-ICT/Onderzoek/Lectoraat-Big-Data.htm>

⁶ <http://www.bdva.eu/>

⁷ <https://www.hogeschoolrotterdam.nl/onderzoek/projecten-en-publicaties/creating-010/smart-and-inclusive-society/kenniswerkplaats-urban-big-data/>

Hiemstra et al., 2014; Manyika et al., 2011; Kool et al., 2015).⁸ Bepaalde kansen en beperkingen gelden ook voor regulier enquête-onderzoek (o.a. correlatie is niet hetzelfde als causaliteit).

Kansen

Big data kunnen een organisatie in staat stellen om efficiënter en effectiever te opereren.⁹ Het succesvol benutten van big data kan bijvoorbeeld bijdragen aan het efficiënt inzetten van schaarse middelen (o.a. politie, ambulances). Zo gebruikt de politie in Los Angeles de data van eerdere inbraakmeldingen om effectievere patrouilleroutes uit te stippelen (Kool et al., 2015). Big data kunnen een organisatie die een gezonde leefstijl promoot bijvoorbeeld inzicht geven in de wijken of gemeenten waar relatief veel personen wonen die er een ongezonde leefstijl op nahouden. Hierdoor kunnen geld en middelen worden gericht op die groepen waar waarschijnlijk de meeste gezondheidswinst te behalen valt.

Big data bieden gelegenheid om terug in de tijd te kijken en de historie te construeren op basis van feiten. Op basis van GPS-signalen, bel- en surfgedrag en camera's in de openbare ruimte kan worden achterhaald waar iemand zich heeft bevonden, met wie die heeft gebeld en hoe het internetgebruik eruit zag. Mede door het inzetten van big data is bijvoorbeeld de dader van de bomaanslagen tijdens de marathon van Boston in 2013 opgepakt.

Daarnaast kan een situatie *real-time* worden gemonitord. Door middel van big data (GPS-signalen, camera's) kunnen bijvoorbeeld het aantal personen bij een sportevenement of uitgaansgebied worden geschat, evenals de verplaatsingen van deze groep. Hiermee kunnen onder andere incidenten worden voorkomen doordat politie tijdig kan ingrijpen (Kool et al., 2015).

Tot slot kunnen big data van voorspellende waarde zijn. Gedrag en oorzaak-gevolg relaties in het verleden zijn geen garantie voor toekomstige ontwikkelingen, maar kunnen wel een houvast of enige vorm van zekerheid bieden.

Beperkingen

Er gloren niet alleen kansen, big data kennen ook beperkingen.

De kans is reëel dat bij big data sprake is van een vertekening (bias). De groep personen uit een dataset hoeft bijvoorbeeld niet per definitie een goede afspiegeling te zijn van de groep waar men in geïnteresseerd is. Bovendien geven data inhoudelijk niet altijd goed weer wat bedoeld of gebeurd is (gebrekkige validiteit).

Er worden verkenningen uitgevoerd in hoeverre big data een substituuut kunnen zijn voor bepaalde onderzoeken onder de Nederlandse bevolking. Zo heeft het CBS gekeken naar de ontwikkeling van het

⁸ Op juridische aspecten wordt hier niet ingegaan, zie daarvoor bijvoorbeeld WRR (2016).

⁹ Efficiëntie is de relatie tussen financiële middelen en tijd enerzijds en de prestaties anderzijds. Met andere woorden: met welk budget en binnen welke termijn is welke prestatie geleverd? Effectiviteit betreft de relatie tussen de prestaties en de (maatschappelijke) effecten.

consumentenvertrouwen op basis van nationaal steekproefonderzoek en de sentimenten op sociale media (Twitter en Facebook). Het CBS concludeert dat er een hoge correlatie bestaat tussen deze sentimenten en het consumentenvertrouwen (VSO, 2014). Zeker op de korte termijn zal het bevolkingsonderzoek echter niet worden vervangen door de sentimentsanalyse op sociale media. De ingetypete woorden op sociale media kunnen met de huidige softwaretools niet allemaal juist worden gekwalificeerd en geïnterpreteerd (denk bijvoorbeeld aan ironie en typefouten). Bovendien wordt de afhankelijkheid van de populariteit van sociale media platforms onder bevolkingsgroepen wellicht te groot (Twittergebruikers vormen bovendien geen goede afspiegeling van de Nederlandse bevolking).

Correlatie (samenhang tussen A en B) is niet hetzelfde als causaliteit (A leidt tot B). Een samenhang is niet per definitie een causale relatie. Met name bij grote hoeveelheden data is er de kans dat er sterke correlaties zijn tussen variabelen die in werkelijkheid geen relatie hebben. Met name bij big data, waarbij veel data betrokken zijn, is dit een aandachtspunt. Voorkomen moet worden dat op basis van onjuiste analyses interventies worden ontwikkeld.

De toegankelijkheid van data wordt gekenmerkt door twee aspecten. In de eerste plaats zijn lang niet alle big data openbaar toegankelijk (open data). Dit wil zeggen dat er geen auteursrecht of rechten van derden op berusten. Nog maar een klein deel van de Nederlandse gemeenten (veertien gemeenten) heeft bijvoorbeeld zelf een open data portal (De Kruijck et al., 2015). In de tweede plaats is lang niet alle big data door software op een computer te gebruiken (*machine-readable*). Big data in PDF-bestanden kunnen bijvoorbeeld wel openbaar zijn, maar in dit formaat zijn ze niet *machine-readable*.

Bij big data mag dan in veel gevallen sprake zijn van grote hoeveelheden data, maar dat betekent niet dat er altijd voldoende data zijn. Over sommige onderwerpen, zoals het gedrag op sociale media, zijn relatief veel data aanwezig, maar voor andere onderwerpen is dat niet het geval. Bovendien kan het voorkomen dat er weliswaar over een onderwerp voldoende data zijn, maar dat die betrekking hebben op een relatief kleine groep. Dit is bijvoorbeeld het geval als er veel data zijn over sportaccommodaties in één regio en nauwelijks tot geen data over accommodaties in een andere regio. Er zijn dan weliswaar veel data beschikbaar, maar die maken het niet mogelijk om een landelijk beeld te schetsen. Tot slot kunnen data onjuist zijn omdat ze te oud zijn en/of onderhevig aan manipulatie.

Aansluitend op het voorgaande punt: er kan ook sprake zijn van een overvloed aan gegevens waardoor door de bomen het bos niet meer zichtbaar is. Het risico dat dan kan ontstaan, is dat waardevolle data of verbanden niet worden opgemerkt door diegenen voor wie het van waarde is. Dit fenomeen wordt 'dark data' genoemd (Millington & Millington, in voorbereiding). Experts wijzen al op een toekomstige uitdaging: hoe komen we over enkele jaren van grote hoeveelheden (onbruikbare) data af?

Het gebruik van persoonsgegevens (burgers, consumenten) is gelimiteerd: de privacyregels die in Nederland nageleefd dienen te worden, zijn vastgesteld door het College Persoons Bescherming.¹⁰ Elementaire rechten mogen niet worden geschonden. Zo is het voor burgers bijvoorbeeld onwenselijk dat medische gegevens in handen van een zorgverzekeraar vallen.¹¹ Bovendien betekent de

¹⁰ Op de website van het College Persoons Bescherming zijn de privacyregels voor Nederland te vinden: <https://autoriteitpersoonsgegevens.nl/>.

¹¹ Het gebruik van data kan worden gestimuleerd als personen er de regie over houden, profiteren van het gebruik van de data, ze vooraf worden geïnformeerd, ze de

beschikbaarheid van (open) data niet dat die voor alle (commerciële) doeleinden gebruikt mogen worden.

Een risico of belemmering is het optreden van de zogenaamde transparantieparadox (o.a. CDV, 2014). Dit houdt in dat profielen en gedrag van burgers en consumenten voor de overheid en het bedrijfsleven transparanter worden, terwijl aan de andere kant overheden en bedrijven en de manier waarop zij analyses doen en toepassen voor burgers en consumenten niet duidelijk zijn. Als dit zich voordoet - waar voldoende aanwijzingen voor zijn - wekt dat onder burgers en consumenten mogelijk gevoelens van twijfel en argwaan op. Dit kan resulteren in een rem op medewerking of het delen van informatie en een verminderd vertrouwen in organisaties. Dit wordt in de hand gewerkt doordat mensen niet altijd goed in de gaten hebben waarvoor goedkeuring wordt gegeven, zoals bij het accepteren van *cookies*.¹²

gelegenheid wordt gegeven de gegevens in te zien (en eventueel te kunnen wijzigen) en zorg te dragen voor een goede beveiliging.

¹² *Cookies* zijn data die een server naar de browser stuurt en opgeslagen worden. Bij een volgend websitebezoek wordt dit weer naar de server teruggestuurd waardoor de server de browser herkent en bijhoudt wat de gebruiker eerder online heeft gedaan.

3. Sportonderzoek en big data

In het vorige hoofdstuk hebben we de algemene ontwikkelingen van big data beschreven. In dit hoofdstuk beschrijven we de betekenis en de ervaring met big data voor sociaalwetenschappelijke vraagstukken rond sport. Een strakke definitie van big data hanteren we daarbij niet. Voor de beschrijving van de toekomstige rol van big data voor sportonderzoek is het echter zinvol om het onderwerp enigszins af te bakenen. Een belangrijke inperking betreft onze focus op breedtesport. Hoewel het verbeteren van (top)sportprestaties veelbelovend is voor big data analyses, valt dat buiten dit onderzoek. De relevantie van big data voor sportmarketing staat ook buiten kijf, maar komt ook niet aan de orde.

Voor de ordening van ons onderwerp gaan we uit van vier bronnen van big data die belangrijk zijn voor sportonderzoek of dat kunnen worden:

- Persoonlijke meters en apps (§ 3.1): informatie van individuen via (beweeg)meters en (beweeg)apps.
- Uitingen via media (§ 3.2): informatie van sociale media, traditionele media en internet.
- Dataregisters voor sportonderzoek (§ 3.3): informatie van databases zoals registers.
- Openbare observaties (§ 3.4): informatie van observaties in ruime zin zoals gsm-signalen, doorgaans vinden de waarnemingen in de openbare ruimte plaats.



Per onderwerp bespreken we de ontwikkelingen met daarbij de (verwachte) waarde voor sportonderzoek.

3.1 Persoonlijke meters en apps

In deze paragraaf bespreken we data die afkomstig zijn uit draagbare apparaten om lichaamsfuncties te meten en (beweeg)apps. Met apparaten bedoelen we smartphones voorzien van zaken als gps-positiebepaling en stappentellers, maar ook andere draagbare apparaten zoals *activity trackers* ('polsbandje'), sporthorloges en hartslagmeters. De apps hebben betrekking op de functies die de apparaten hebben. In onze bespreking gaan we uit van de functies die de draagbare apparaten/apps kunnen hebben. We realiseren ons dat er vele combinatiefuncties mogelijk zijn.

Metten van inspanning

Onder de categorie draagbare meters om inspanning te meten, vallen onder andere hartslagmeters, stappentellers en accelerometers. Deze laatste meten versnellingen en verplaatsingen in drie dimensies en worden door wetenschappers vaak gebruikt in onderzoek naar de mate van inspanning. Meerdere smartphones kunnen afgelegde afstanden meten (via GPS of stappenteller). Hartslagmeters en accelerometers zijn te vinden op smartphones, smartwatches en *activity trackers*.

De nieuwsgierigheid van gebruikers in hun eigen gezondheidswaarden of sportprestaties heeft inmiddels een hoge vlucht genomen (*quantified self movement*, Nafus et al, 2014). Volgens de Smart health monitor 2016 van Multiscope gebruikt 34 procent van de Nederlanders apps, *wearables* of meters om de gezondheid te monitoren. Twintig procent monitort sport- en beweeggedrag. Overigens is 13 procent ex-gebruiker. Dat suggereert dat voor een deel van de gebruikers de interesse niet blijvend is (Multiscope, (2016). Van de Nederlandse hardlopers maakt één op de vijf van een app gebruik (GfK, 2014). Van de deelnemers aan hardloopevenementen ligt dit percentage waarschijnlijk een stuk hoger (Dallinga et al, 2016). In de Verenigde Staten maakt meer dan de helft van de lopers gebruik van een GPS-functionaliteit (The Atlantic, 2015). Ander onderzoek uit de VS wijst echter ook uit dat de helft van de bezitters van een draagbaar meetapparaat deze niet meer gebruikt (Endeavour Partners, 2014).

Appbouwers hebben diverse functies gecombineerd zoals gezondheidsmeting, vergelijking met andere gebruikers, informatie over populaire routes of snelheidsmeting. Bekende apps zijn Strava (vooral wielrennen/mountainbike), Runtastic, Runkeeper (hardlopen) en Endomondo (fitness/sportmotivatie). Een opvallende Nederlandse app is de Fietstelapp (www.fietstelweek.nl). Via een webtool kunnen gebruikers hun data uploaden naar de appbeheerders. De bedrijven stellen (een deel van) de geüploade informatie voor algemeen gebruik beschikbaar. De informatie is te gebruiken voor de analyse van de populariteit van wandel-, hardloop- of fietsroutes. Op individueel niveau is informatie beschikbaar over de variatie in routes, tijdstippen, duur en intensiteit.

Naast apparaten worden ook rfid-chips (electronische sensoren) ingezet om onder meer afstanden te meten. Al sinds 1993 is het bedrijf Mylaps actief met de ChampionChip om deelnemers van hardloopevenementen te voorzien van een objectieve tijdregistratie via een chip in het rugnummer. Enquêtes onder deelnemers zijn daardoor aan de loopprestaties te koppelen.

Voor onderzoek naar sport- en bewegen is een meter belangrijk die de mate van inspanning eenduidig meet. De inspanning valt op verschillende manieren te meten zoals via het aantal stappen, de hartslag of de hoeveelheid zweet. De interpretatie is niet altijd helder. Een hartslagmeting is alleen te interpreteren als je de patronen van een individu kent en is op zich onvoldoende om bijvoorbeeld de mate van verbruikte energie te achterhalen. Hetzelfde geldt voor de hoeveelheid zweet. De vorm van de beweging (traplopen, wandelen, fietsen, hardlopen) is soms al wel goed uit meetgegevens af te leiden. Maar meters die driedimensionale bewegingen vastleggen, hebben soms moeite met het

registreren van fietsen als activiteit als deelnemers de beweegmeter op de heup dragen. De heup staat namelijk relatief stil tijdens het fietsen (Hansen et al, 2014). Hills et al (2013) komen in een overzichtsartikel tot de conclusie dat energiegebruik het beste is vast te stellen door hartslagmeting en accelerometing te combineren.

Doorgaans experimenteren jongere, gezonde mensen met (beweeg)meters waardoor via appgebruik geen betrouwbaar beeld van de bevolking ontstaat. Behalve deze selectiviteit moeten gebruikers ook bereid zijn informatie te delen. De meeste mensen zullen vooral positieve uitkomsten naar een website uploaden.

Specifiek gebruik inspanningsmeters voor onderzoeksdoeleinden

De informatie van beweegmeters is vooral waardevol in combinatie met een aanvullende vragenlijst. In Nederland is van deze combinatie al veel gebruikgemaakt (o.a. door de universiteiten van Maastricht en Utrecht). Ook het Mulier Instituut heeft deze combinatie toegepast. Voor het onderzoek School, Bewegen en Sport gebruikten we een accelerometrie die in drie dimensies meet (Actigraph). Daarbij zijn de beweegdata van bijna vierhonderd leerlingen verzameld (Stuij et al, 2011). Samen met een vragenlijst die de leerlingen invulden, konden gegevens worden gebruikt om na te gaan of de schoolomgeving van invloed was op de wijze van transport van en naar school. De meting duurde een week en legde om de tien seconden het aantal beweegeenheden vast. Door de grote hoeveelheid meetgegevens, was veel bewerking nodig voor de beantwoording van de onderzoeksvraag. Na 2010 zijn honderden artikelen verschenen waarin deze meters zijn gebruikt (Troiano et al, 2010).

Voor bevolkingsonderzoek kan in de toekomst een meter/app beschikbaar komen die inspanning op een eenduidige wijze kan meten. Als deze meter/app op veelgebruikte smartphones of een prettig draagbare smartwatch is geïnstalleerd, dan helpt dit het onderzoek naar sport- en beweeggedrag. Deze informatie is te koppelen met meer algemene vragen naar achtergrondkenmerken, sport-/beweeggedrag en gewoonten.

Metten andere lichaamseigenschappen

Veel informatie over het lichaam is traditioneel via (duur) medisch onderzoek te verkrijgen. Het meten van lichaamsfuncties valt echter steeds goedkoper uit. Zeker als het relatief makkelijk te meten verschijnselen betreft. In de jaren zeventig is in laboratoria bijvoorbeeld veel (psychologisch) onderzoek verricht naar zweetresistentie. Zweet is een aanwijzing voor emotionele stress (*skin resistance measurement* of *galvanic skin response*). Dit soort meters zijn en worden onder andere gebruikt voor leugendetectors. Door efficiëntere technieken is de DNA-analyse sterk in opkomst. Door voortgaande efficiëntie zijn deze analyses zowel goedkoper als eenvoudiger geworden. Dit opent de weg om dergelijke technieken in te zetten als leverancier van informatie over de beweegaanleg van een individu (wikipedia.org/wiki/Rapid_DNA).

Andere lichaamsinformatie is te verkrijgen via technieken zoals *eye-tracking*. *Eye-tracking* was in de beginperiode in de jaren negentig veelbelovend, maar krijgt nu minder aandacht. Met deze techniek zijn via een computersysteem nauwkeurig oogbewegingen te achterhalen. Zo kunnen adverteerders zien hoe mensen bijvoorbeeld een krant lezen en welke aandacht advertenties krijgen. De betekenis van de duur dat mensen naar een specifiek punt kijken, blijkt echter lastig te bepalen. Voor (top)sportonderzoek is de techniek bruikbaar voor bijvoorbeeld de analyse van het nemen van strafschoppen (Timmis et al, 2014). Welke kijkbewegingen leiden tot welke beslissingen bij doelman en

schutter? Een ander voorbeeld is het gebruik voor de Formule 1 (Skysports, 2016). Hoe snel kunnen coureurs reageren op waarnemingen, ook al is het vanuit hun ooghoeken?

Apps om bewegen te stimuleren

Enkele beweegapps hebben speciaal als doel om mensen tot sport en bewegen aan te zetten, bij wijze van persoonlijke coach. Een succesvol voorbeeld is 'Hardlopen met Evy' (www.hardlopenmetevy.com). Feitelijk is het een soort hardloopcursus waar de vorderingen via de app bijgehouden worden. Afhankelijk van de vorderingen volgen aanwijzingen. In Nederland zijn zowel sporthogescholen (Fontys, Hogeschool van Amsterdam) als universiteiten (TU Eindhoven) voor dit doel met apps bezig. De terugkoppeling van beweeginformatie kan stimuleren om gestelde dag- of weekdoelen te halen. Anderzijds kan deze informatie ontmoedigen als een doel te ambitieus is. Zie voor een overzichtsartikel Dute et al (2016).

Lector Steven Vos van Fontys Sporthogeschool voert met negentien andere partijen het project 'Voor iedereen een app?' uit. Doelstelling van dit project is bepalen welke sportapp past bij welke gebruiker. Het betreft apps op vlak van sport-/beweegstimulering, trainingsbegeleiding, gezondheidsbevordering, *community building* en monitoring. De nadruk ligt op de implementatie van allerhande beweegapps en beweegmeters. Samen met TU Eindhoven voert Vos ook een onderzoeksprogramma uit met als doel om ongebonden sporters duurzaam te laten hardlopen. Het project heeft de producten Inspirun (gepersonaliseerde hardloopcoach) en Run! (interactieve hardlooproute) opgeleverd (Vos, 2016).

Sportlector Marije Baart de la Faille van de Hogeschool van Amsterdam doet onderzoek naar de stimulans van apps om meer te gaan bewegen (zie o.a. www.hva.nl/urban-vitality). Daarnaast onderzoekt Baart de la Faille effecten van *beacons* als stimulans bij het hardlopen (Bewegen in Amsterdam met Beacons, Bambea). *Beacons* zijn bakens in de omgeving die informatie geven aan smartphones via bluetooth (www.ibeacon.com).

Via positiebepalingsapparatuur van bijvoorbeeld Garmin of met de GPS-functie van een smartphone kunnen deelnemers Geocachen. Het apparaat/app bevat informatie over de locatie van kleine eenheden die in het landschap of een stad zijn verstopt. Als met behulp van het apparaat de *cache*/eenheid gevonden is, kan dat op de website worden gemeld. Feitelijk is het een vorm van spoorzoeken en kan bewegen stimuleren voor mensen die anders niet wandelen (www.geocaching.com). Het voorbeeld van de zomer van 2016 is de opkomst van Pokémon Go. Massaal gaan vooral jongeren op jacht naar objecten die in het landschap zijn 'geprojecteerd'. Wellicht stimuleren apps die wandelroutes weergeven het lopen van deze routes.

Overige functies apps

Een andere toepassing is dataverzameling via een app met als (hoofd)doel onderzoek. Het is dan een alternatieve mogelijkheid om vragen aan groepen mensen te stellen, naast bijvoorbeeld een online of een papieren vragenlijst. Voor enquêteurs is een app ook handig. Soms vragen onderzoekers respondenten een app te downloaden waarop vragen zijn te beantwoorden. Deze methode zal een vrij lage respons opleveren gezien de extra stappen die respondenten moeten nemen voor hun deelname aan een onderzoek (ook beschikbaarheid wifi speelt een rol). Het is ook mogelijk om apps in te zetten voor registraties door het publiek. Zo heeft het Mulier Instituut de app Sportdrempelvrij ontwikkeld om vast te stellen of sportaccommodaties toegankelijk zijn met een beperking (www.sportdrempelvrij.nl).

Meerdere apps hebben als hoofddoel het informeren van sporters zoals de app van Uniek Sporten (www.unieksporten.nl/apps). Zij helpen mensen met een beperking om aangepaste sportaanbieders te zoeken. Andere sportinformatie-apps richten zich op club of speel informatie en soms ook op het algemene sportaanbod. Dit soort apps leveren nauwelijks (big) data op.

Een andere functie van apps is het gebruik bij sportevenementen (onder meer door Golazo). De aanmelding kan bijvoorbeeld via een app gebeuren. Als met de aanmelding een aantal gegevens van de deelnemers worden geregistreerd, kan dat uiteraard meer informatie opleveren. Zeker als deze zijn te combineren met de individuele prestaties van het sportevenement.

3.2 Uitingen via media

In deze paragraaf gaan we in op informatie die we via diverse media kunnen verkrijgen. We onderscheiden de nieuwe mediavormen naast de traditionele media zoals radio, televisie, kranten, tijdschriften en boeken. Van de eerste categorie bespreken we sociale media en internet.

Achtereenvolgens bespreken we:

- Uitingen op sociale media: 3.2.1
- Informatie van internet zelf: 3.2.2
- Internetzoekopdrachten en websitebezoek van individuen: 3.2.3
- Informatie uit kranten, boeken, radio- en televisie-uitzendingen: 3.2.4

3.2.1 Uitingen op sociale media

Voor media zoals kranten, radio en televisie is het meten van de ‘emotionele temperatuur’ op sociale media een geliefde bron om informatie met persoonlijke reacties in te kleuren. Sociale media kunnen daarnaast ook een bron zijn voor (sport)onderzoek.

Onder sociale media verstaan we toepassingen waar directe onderlinge communicatie mogelijk is. Dit betreft dus vooral Facebook, Instagram en Twitter. Whatsapp beschouwen we niet als sociale media aangezien de berichten niet openbaar zijn. Persoonlijke bijdragen in de vorm van blogs, vlogs en YouTube-filmpjes zien we als internetuitingen (zie verder). Voor de analyse van sociale media is het van belang dat de uitingen openbaar zijn. De meeste Twittergebruikers delen hun tweets met iedereen, maar de accounts van veel Facebook- en Instagramgebruikers zijn vaak afgeschermd. In de praktijk betekent dat alleen openbare tweets voor analyse in aanmerking komen.

Voor de analyse van sociale media zijn kant en klare tools ontwikkeld om verbanden op te sporen. In Nederland zijn onder andere Coosto en Buzzcapture hiermee actief. Mogelijkheden zijn het tellen van de frequentie van woorden en woordcombinaties, netwerkanalyses van personen en het vaststellen van het affect (positief of negatief). De rapportage vindt onder andere plaats met woordwolken en de weergave van de uitingen in de tijd. Andere aanbieders zijn Radian6 en socialmention.com dat gratis is. Een belangrijke vraag bij de keuze van een aanbieder is hoe lang zij tweets/sociale media berichten bewaren. Twitter bewaart bijvoorbeeld niet alle tweets. Na verloop van tijd verdwijnen deze van hun website.

Een voorbeeld van een Twitteranalyse over sport is een artikel over vier WK-voetbalwedstrijden van de VS (Billings et al., 2015). In bijna 7.500 wedstrijd tweets van Amerikanen vonden zij geen verschil in de genoemde succes- en faalfactoren voor het eigen team en het team van de tegenstander. Deze uitkomst

is in tegenstelling met ander onderzoek. Met gebruikmaking van vrijwel dezelfde data hebben Yang Yu et al (2015) vooral de verandering van emoties bij Amerikanen gemeten, voor en na het scoren van een goal. Angst en boosheid waren de sterkst gemeten emoties. Deze namen af als de VS scoorde. De vraag is of veel mensen op echt spannende wedstrijdmomenten door blijven twitteren.

Voor de Olympische Spelen in Londen 2012 is een uitgebreide analyse van de sociale media uitingen uitgevoerd (Hemment et al., 2012). Het doel was om een grafisch instrument te ontwikkelen om de emoties in een gebied over het evenement weer te geven. Via deze software kreeg elk sentiment in een tweet een score van -6 'erg negatief' tot +6 'erg positief' (Salience Engine van Lexalytics).

Andere onderzoeksdomeinen kunnen inspireren om vraagstellingen met sociale media te beantwoorden. Recent is bijvoorbeeld een onderzoek naar tweets met #gelukszoeker gepubliceerd (Verhaar et al., 2016). Zij maakten gebruik van een combinatie van een netwerk- en een linguïstische analyse om 560.000 tweets te analyseren. Het resultaat is een grafiek waarin zowel taalkundige aspecten als de verbinding tussen personen staat afgebeeld. Het thema vluchtelingen stond in 2015 centraal vanwege de komst van grote groepen vooral Syrische vluchtelingen die via Griekenland en Oost-Europa richting West-Europa trokken. Het resultaat laat zien dat tegenstanders vooral geconcentreerd zijn rond de tweets van een centraal persoon en dat de voorstanders rond verschillende belangrijke personen zijn gegroepeerd.

Het Nederlandse bedrijf Ynformed voorspelt het gedrag van individuen via onder andere sociale media. Zij ontwikkelden een model voor het voorspellen van schooluitval door het taalgebruik op Twitter aan schoolsucces te koppelen (<http://ynformed.nl>).

Bruikbaarheid sociale media voor onderzoek

De analyse van sociale media stuit op meerdere methodologische problemen. Zo kunnen alleen mensen met een (openbaar) Twitteraccount reageren. Mensen met veel volgers kunnen een Twitterdebat sterk beïnvloeden. De eenheid van onderzoek zijn doorgaans de tweets, de intensieve twitteraars tellen daardoor vaker mee. Kantelt de mening van twitteraars na een afwijkende mening van iemand met veel volgers of iemand die veel wordt geretweet?

Tot slot rijst de vraag of er voldoende uitingen zijn te vinden voor een specifieke vraagstelling. Met name voor sport is dat soms een thema. Een voorbeeld is de start van de Tour de France in Utrecht begin juli 2015. Er zijn veel berichten over de eerste touretappe, maar de duur is lang omdat het een individuele tijdrif betrof. Daarnaast zijn binnen deze periode minder specifieke hoogtepunten waar mensen over twitteren (zoals een doelpunt bij voetbal).

Wat zijn vragen voor de sportsector die in aanmerking komen voor analyse van sociale media?

- Reacties op mediaberichten. Hoe reageren mensen op nieuws over FIFA en de doping in de atletiek? Beoordelen zij deze sporten negatiever of zien zij dit los van de sport zelf?
- Beleving van (topsport)evenementen. Bijvoorbeeld voor alle voetbalwedstrijden in een competitiejaar bestuderen welke clubs het populairst zijn, tegen welk team supporters het meest opzien om tegen te spelen etc.;
- Wat melden individuen vooraf en achteraf over de wedstrijd van hun club? Effecten van euforie en teleurstelling vergelijken;

- Sportmoment. Laten sporters zich beïnvloeden door het weer? Hoe reageren ze als sportmaatjes afzeggen?

3.2.2 Informatie van internet zelf

In deze paragraaf gaan we in op internet als bron van informatie via website-artikelen, vlogs, blogs en YouTube-filmpjes. Op internet opgeslagen informatie van traditionele media bespreken we hier niet (zie paragraaf 3.2.4). We richten ons daarnaast niet op meta-informatie afkomstig van zoekopdrachten of websitetrackers (zie paragraaf 3.2.3).

Veel informatieverzameling op internet vindt via deskresearch plaats. Voor deze zoektochten is internet ideaal: een zoekopdracht, selectie van informatie, vaststellen van de kwaliteit van de informatie en vervolgens vastleggen van het resultaat in een rapport, een scriptie of op een op andere website. Ook wetenschappelijk literatuuronderzoek vindt steeds vaker via internet plaats. Een voorbeeld van informatie die relevant is voor de sport, is het verzamelen van de jaarverslagen van sportbonden. Door hieruit informatie in een centraal bestand op te nemen, zijn voor meerdere disciplines over langere periode trends over bijvoorbeeld bondsafdracht per lid en de financiële situatie af te leiden.

Naast handmatige zoektochten zijn systemen ontwikkeld om internet ‘af te grazen’. De systemen hebben namen als spiders, zoekrobots, webcrawlers of botnets (*spiders internet search engines*). De toepasbaarheid van geautomatiseerde zoekmachines neemt steeds verder toe. Te denken valt aan een systeem dat de entreetarieven van zwembaden in Nederland bijhoudt. Het Nederlandse bedrijf Owlín pikt daarnaast snel (financiële) berichten op uit twee miljoen websites. Het betreft informatie die aanvankelijk kleinschalig wordt verspreid via bijvoorbeeld blogs en steeds belangrijker wordt. Zij zijn daardoor in staat trends snel te zien, bijvoorbeeld over een tegenslag bij een bedrijf die gevolgen heeft voor de winstverwachting.

De informatie van internet is afhankelijk van de betrouwbaarheid van de basisbronnen. Als een instelling gegevens over de openingstijden van zwembaden verzamelt, dan moet bekend zijn of de eigenaar van het zwembad zelf de informatie verspreidt of dat dat via andere websites gebeurt die op hun beurt in een eerdere fase allerlei informatie op internet heeft verzameld. Actualiteit is daardoor vaak een knelpunt. Het eindresultaat kan leiden tot informatie waar een deel zonder meer betrouwbaar zal zijn en een ander deel minder zeker is.

Een ander issue is of informatie daadwerkelijk via zoekmachines op een snel uitdijend internet te vinden is. Internetzoekmachine kunnen bij wijze van ‘service’ informatie die vaker is geraadpleegd bovenaan in een resultaatlijst te zetten. Onbekende informatie komt dan minder snel bovendrijven. Om tunnelvisies te voorkomen, kunnen verschillende zoekmachines worden ingezet.

3.2.3 Internetzoekopdrachten en websitebezoek van individuen

In dit deel richten we ons op de meta-informatie over het internetgebruik. Hoe zoeken mensen informatie, waar zoeken zij naar en welke websites bezoeken zij? Instrumenten hiervoor zijn informatie van zoekmachines (Google Trends) en van trackers zoals *cookies*.

Trackers houden bij welke websites en sociale media gebruikers bezoeken en hoe lang. Zij werken met de unieke ip-adressen van pc’s en smartphones. *Cookies* zijn een vorm van tracking. Deze *cookies* zijn

behalve voor het geven van service (locatie) vooral bedoeld om informatie te verzamelen. *Third party cookies* kijken voor website A naar internetgedrag op website B.

Met voldoende informatie van trackers kunnen bedrijven de zogenaamde *customer journey* van een individu volgen (Marketingfacts, 2016). Het is een term uit de marketing die weergeeft hoe consumenten naar een aankoop toegroeien. Dergelijke *customer journeys* zijn vooral relevant voor een op-een marketing van bedrijven. Zorgen dat precies op tijd reclame over een bepaald product voorhanden is. Bij wijze van spreken nog voordat iemand zelf heeft nagedacht of hij of zij een product wil hebben. Overigens heeft de *customer journey* niet alleen betrekking op informatie via internet. Het kan ook bestaan uit gesprekken met mensen in de omgeving en bestaande gebruikers of via informatie via de traditionele media zoals televisie en radio. Gericht kunnen adverteren op interesse dat uit internetsurf- en zoekgedrag blijkt, is een goudmijn voor marketeers.

Een voorbeeld van een *internet customer journey*: een individuele vrouw krijgt een baby. Eerst valt via zoekgedrag op dat zij uitgaat. Vervolgens bezoekt ze af en toe een website met informatie hoe het is om kinderen te hebben. Weer later vallen bezoeken aan websites over zwangerschap op. Daarna volgt de behoefte aan informatie over de bevalling. Bedrijven willen bijvoorbeeld weten wanneer zij behoefte heeft aan informatie over soorten en merken luiers. Meestal blijkt dit pas te zijn als het kind geboren is. De inrichting van een babykamer is tijdens de zwangerschap al voorbereid.

Voor sport kan het denken vanuit *customer journeys* interessant zijn. Denk bijvoorbeeld aan het proces van toegroeien naar een specifieke sport. Eerst kennismaking met een sport via vrienden op Facebook. Vervolgens zoeken welke spullen nodig zijn en wat deze kosten. Daarna kan zoekgedrag naar sportvoorzieningen volgen.

Met meta-informatie van zoekmachines zoals Google valt algemene informatie over zoekgedrag op te zoeken (via Google Trends). Bekend is het voorbeeld dat het aantal keer zoeken naar het woord griep verband zou houden met de daadwerkelijke opkomst van een griepepidemie (*Google Flu Trends*). Het zoekgedrag geeft een sneller signaal dan de registratiesystemen van de artsen die de ziektes van hun patiënten bijhouden. Het bureau Duodecim stelde met dit soort informatie een paar jaar geleden de Populariteitsindex sport samen. De makers combineerden belangstelling en beoefening van sport in een index (<http://inmotion-sportsconsultancy.nl/Downloads/Populairiteitsindex%20sport%20-%202011.pdf>).

Deze vergelijkingen in de tijd zijn lastig vanwege de sterke groei van zowel het gebruik van internet als het aanbod van websites. In de tussentijd kunnen immers nieuwe/andere gebruikersgroepen gebruik zijn gaan maken van internet en kan content zijn toegevoegd die tot ander zoekgedrag/gebruik leidt. Daarnaast kan de keuze van de zoekterm invloed hebben (hardlopen, joggen of running). Als meerdere valide zoektermen beschikbaar zijn, leidt dit tot een onderschatting. Bepaalde groepen kunnen ander zoekgedrag hebben (frequentie, woordkeus). Tot slot maakt uit in welke mate behoefte bestaat aan informatie voor een sport. Zoeken deelnemers bijvoorbeeld naar speelschema's op internet of is dat voor een specifieke sport niet nodig.

3.2.4 Informatie uit kranten, boeken, radio- en televisie-uitzendingen

Veel informatie is te halen uit beeld, geluid en geschreven tekst afkomstig van kranten, radio, televisie en boeken. Internet is hiervoor tegenwoordig een belangrijke verzamelplaats, zelfs voor boeken. Instellingen zoals Google Books en de Koninklijke Bibliotheek digitaliseren boeken en maken deze via internet bereikbaar. Met hulp van NWO biedt Clariah data plus toegangsapplicaties (www.clariah.nl).

Inhoudsanalyse is het instrument voor onderzoek naar de traditionele media. Dit betreft ook media-analyses over sport. Vragen kunnen zijn hoeveel aandacht een specifieke sport in de media krijgt of wat de positie van vrouwen in de sport is. Observatie en interpretatie van videobeelden is ook voor sportonderzoek belangrijk. In veel voetbalpraatprogramma's worden wedstrijdfragmenten verrijkt met cirkels en pijlen rond spelers in de 'wat-als-analyses'. Voor topsporters zijn videobeelden al lange tijd onmisbaar om prestaties te verbeteren. Video-observaties en inhoudsanalyses vinden steeds vaker plaats met geautomatiseerde analysetechnieken van tekst, geluid en beelden. We bespreken een aantal ontwikkelingen.

Ontwikkeling analyse tekst, geluid en beeld

Inhoudsanalyses van sportberichtgeving met traditionele middelen bestaan al lang. Zo publiceerden bijvoorbeeld McCarthy et al. in 2003 een artikel over de positie van zwarte spelers in de media. Knoppers en Elling (1999) deden vergelijkbaar onderzoek naar gender en etniciteit. Ruim tien jaar later zijn technieken om beeld en geluid te coderen en om in teksten te zoeken volop in ontwikkeling. De technieken kunnen informatie uit beelden, geluiden en teksten destilleren in een vorm die door onderzoekers te interpreteren is. Met deze technieken vallen inhoudsanalyses aanzienlijk goedkoper uit maar is het aantal onderzochte artikelen ook nog eens aanzienlijk groter.

Van beeld, geluid en geschreven tekst is de laatste het makkelijkst te onderzoeken. Om onderzoek in Europa te bevorderen, verzamelt FutureTDM informatie over tekst en datamining (www.futuretdm.eu). Het is een initiatief van de Europese Unie voor tekstanalyse van gedigitaliseerde teksten. Bij de Koninklijke Bibliotheek gaat een onderzoeker via een geautomatiseerd systeem aan de slag met de analyse van sprookjes. Haar vraag is of software semantische rollen in verhalende teksten herkent. Daarbij speelt vooral of de op de Engelse taal gerichte software voor Nederlandse tekst is in te zetten (E-data & research Jaargang 10 nummer 3, juni 2016).

Eagleman et al. (2014) beschrijven een onderzoek waarbij zij een inhoudsanalyse uitvoerden op bijna 4.000 artikelen van zes nieuwswebsites van zes verschillende landen tijdens de Olympische Spelen in Londen van 2012. Voor artikelen over de sport op de spelen, bekeken zij de inhoud op sekse, landenvoorkeur en sportvoorkeur. Zij zagen hierop voor internet nauwelijks verschillen terwijl deze voor gedrukte media juist vrij groot zijn. Dit zou betekenen dat internet op deze onderwerpen neutraler is.

Nog grootschaliger is een onderzoek van Shor et al (2014). Hij onderzocht in krantenartikelen de verschillen in de beoordeling van mannen en vrouwen. De onderzoekers maakten gebruik van het systeem Lydia. Dit systeem verzamelt sinds 2004 meer dan 25 miljoen Engelstalige artikelen van 3.000 kranten en stelt aan onderzoekers informatie beschikbaar. Uit dit systeem selecteerde Shor zes bladen die verschillen naar politieke kleur. Voor sport vonden zij dat slechts een klein deel van de sportberichtgeving over vrouwen ging, onafhankelijk van de signatuur van de krant.

Veel televisieprogramma's kennen de mogelijkheid om spraak als leestekst onder het scherm te projecteren. Deze tekstjes vormen aldus een goede basis om het gesproken woord te bestuderen. Een studie van de nieuwsprogramma's van 140 televisiekanalen in de Verenigde Staten gedurende zes maanden maakt hiervan voor een inhoudsanalyse gebruik (Castillo et al., 2013). De onderzoekers gebruikten een geautomatiseerd analysesysteem bestaande uit verschillende stappen. De eerste stap is het scheiden van nieuwsprogramma's van andere programma's. Elke zin van de ondertiteling is vervolgens gecategoriseerd en voorzien van een sentimentscore. Deze score kwam tot stand door het

tellen van positieve en negatieve woorden in de zin. De onderzoekers pasten hiervoor een methode toe om computers met mensen te laten communiceren (*natural language processing* ofwel NLP). Voor sportonderwerpen ontdekten zij dat zinnen korter waren en dat een laag opleidingsniveau volstond om deze te begrijpen. De uitkomsten komen overeen met eerdere inhoudsanalyses met traditionele middelen. Met deze methode komen ook andere onderwerpen in bereik. Door uit te gaan van een kleine basiseenheid kunnen zij de stijl van een nieuwslezer analyseren, televisiekanalen typeren en zelfs vaak in nieuwsonderwerpen genoemde personen beoordelen.

Video-analyse van bijvoorbeeld een nieuwsuitzending bestaat ook. Zo gebruikte Lin et al (2002) rond een item een combinatie van tekst en kenmerken van een beeld (lucht, snelle wisselingen van beelden etc.) om betekenis aan de beelden te geven (*Support Vector Machine* methode).

De analyse van het *gesproken woord* is een ander veld waar veel onderzoek zich op richt. Een flink deel van de inspanningen hiervoor zijn gericht op de communicatie tussen mens en machine/computer. Enerzijds zijn hierop stapjes gezet via *machine learning* technieken. Anderzijds vindt vanuit linguïstische kennis/theorievorming vooruitgang plaats. De uitdagingen blijven echter groot omdat weinig gesproken woord perfect is. We zien complicaties in grammaticale fouten, tongval, half uitgesproken zinnen en variabele spreek snelheid (*Neural Information Processing Systems*, NIPS 2015).

De beschreven technieken leiden vooralsnog tot een relatief oppervlakkige analyse van de inhoud van media. Deze is uiteraard te verlevendigen met een aantal citaten. Het voordeel van deze nieuwe technieken is dat een schatting is te maken in welke mate de artikelen zijn gelezen/gezien door het tellen van de *views* en downloads op internet. Traditionele inhoudsanalyse is niet in staat om het bereik te bepalen. Zo telt een amper gelezen ‘verstopt’ bericht even sterk mee als voorpaginanieuws.

3.3 Dataregisters voor sportonderzoek

Veel informatie ligt verscholen in databases met registraties. Sommigen zijn openbaar, voor andere leidt betaling soms tot toegang. We bespreken kort enkele databases die relevant voor sportonderzoek kunnen zijn. Eerst van redelijk openbare bronnen zoals van overheid en stichtingen, vervolgens van bedrijven waar je informatie kunt kopen en tot slot data die in principe afgeschermd zijn, maar waar via samenwerking of opdrachtonderzoek mee gewerkt kan worden.

Sportinformatie van overheden en sportorganisaties

Er zijn diverse registraties die voor sportonderzoek relevant zijn, zowel van sportorganisaties, bedrijven als (semi)overheden. We sommen een aantal voorbeelden op.

In de sport zijn voorbeelden van organisaties die hun sportdata toegankelijk maken voor andere belangstellenden (open data). De Amerikaanse NBA (*National Basketball Association*) heeft een online database ontsloten met nagenoeg alle statistieken die er zijn. Dit is interessante informatie voor fans die erin kunnen grasduinen en er met andere fans of derden over kunnen spreken. Een ander voorbeeld is dat van Manchester City. Deze Engelse voetbalclub heeft alle data van spelers en teams uit de competitie geopenbaard op een website en heeft fans, data-experts en andere geïnteresseerden opgeroepen om op basis van deze cijfers met goede ideeën bij de voetbalclub aan te kloppen (Millington & Millington, in voorbereiding).

Bij informatie van (semi)overheidsorganisaties valt te denken aan het fietsknooppuntennetwerk (www.pdok.nl) en wandelroutes (www.openstreetmap.nl).

Verder is gedetailleerde financiële data van de uitgaven en inkomsten van gemeenten beschikbaar (IV3 data). Het Mulier Instituut verzamelt sinds 2010 de informatie over sport. Met deze informatie zijn rapporten geschreven over bezuinigingen, gemeentelijke investeringen in accommodaties en algemene trends in de uitgaven voor sport (zie o.a. Hoekman et al., 2015). Per gemeente is informatie voorhanden zoals de uitgaven aan sport per inwoner en het aandeel van sport in de begroting. Daarnaast zijn uitgaven te onderscheiden zoals de uitbesteding van het beheer van accommodaties, de uitgaven aan sportstimulering en de kosten voor de accommodaties zelf.

Een andere mogelijkheid bestaat om via de Dienst Uitvoering Onderwijs in Groningen (DUO) eindexamenresultaten te koppelen aan kenmerken van sportdeelname volgens de Gezondheidsenquête. Heeft lichamelijke activiteit van leerlingen bijvoorbeeld een relatie met de (latere) examencijfers? Of hebben kenmerken van de opleiding invloed op het voldoen aan de beweegnorm op latere leeftijd?

De Jeugd Gezondheidszorg screent van 13- en 14-jarigen diverse facetten van hun leven. In principe krijgt iedere leerling deze vragen voorgelegd. Het betreft ook vragen over sport en bewegen (12 vragen in Utrecht). Ook de (school)carrière in een latere fase is te volgen, aangezien het burgerservicenummer bekend is. Met deze data kan een GGD aldus verbanden leggen tussen gezondheid, sport-/beweegdeelname en schoolcarrière. Een gestandaardiseerde landelijke vragenlijst voor het onderzoek ontbreekt overigens.

Bepaalde databestanden zijn vooral relevant om gegevens uit enquêtes of andere bronnen te verrijken. Dit betreft data van de kwaliteit buurten (Leefbaarometer), het aantal inwoners per postcode en locatiedefinities van kaarten (GIS voor Accommodatiedatabase). Voor enkele van deze bestanden is een bewerking noodzakelijk.

De Sport Data Valley wil een bemiddelaar zijn voor diverse instanties die op zoek zijn naar data (www.sportdatavalley.nl). Het platform is vooral ook bedoeld om data afkomstig van bedrijven aan onderzoekers aan te bieden. Algemene databestanden van de overheid komen steeds vaker beschikbaar (data.overheid.nl).

De stichting Dans van KNAW stelt veel data van sociaalwetenschappelijk onderzoek en humaniora beschikbaar (www.dans.knwa.nl). Ook sportonderzoekers deponeren data bij deze instelling. Onder andere de data van de Nationale Sportonderzoeken (NSO) van het Mulier Instituut zijn hier ondergebracht. Het onderzoek van het CBS is alleen via een speciale afgeschermd procedure te analyseren (microdata-analyse).

Kopen van informatie over sport

We werken hier drie voorbeelden van data uit die voor sport van belang kunnen zijn.

Over wedstrijden zijn veel statistieken te verzamelen. De winnaars, het aantal verzamelde punten of andere meetbare prestaties. Daarnaast worden soms de wedstrijden zelf geanalyseerd. Op welk moment vielen de punten, wat zijn de kansen bij een achterstand in punten of prestaties? In Nederland heeft Gracenote (voorheen Infostrada) veel van dit soort wedstrijd informatie in handen. Voor sportmedia zijn zij een belangrijke leverancier voor aanvullende informatie. Toepassingen voor sociaalwetenschappelijk

(sport)onderzoek naar sport zijn er niet direct voor een database met wedstrijd informatie. Interessanter voor sportonderzoekers is bijvoorbeeld de spreiding van kopers van seizoenkaarten van voetbalclubs dat is verwerkt door het dagblad Tubantia (www.tubantia.nl/extra/datajournalistiek/voetbalkaart).

Voor informatie over (sport)tijdschriften heeft het Nationaal Onderzoek Multimedia verkoopinformatie beschikbaar (nommedia.nl, voorheen via HOI).

De stichting Kijkonderzoek (SKO, www.kijkonderzoek.nl) heeft informatie over de aandacht die de diverse sporten krijgen. Deze gegevens zijn gebaseerd op registraties in panelonderzoek. Strikt genomen zijn dit dus geen echte registraties omdat zij gebaseerd zijn op statistisch onderzoek. Per gezinslid is bekend waar iemand naar kijkt en hoelang. Deze informatie is vooral gepubliceerd in de verschillende jaargangen van de Rapportage Sport van het Sociaal en Cultureel Planbureau. Daarnaast is ook veel informatie te vinden op de website van stichting Kijkonderzoek.

Afgeschermd informatie van bedrijven en sportbonden

Met ledenadministraties van sportbonden/-verenigingen en verkoopinformatie van bedrijven zijn talloze databestanden opgebouwd. Vanwege privacybescherming of (financiële) belangen zijn veel van deze registraties niet of slechts beperkt toegankelijk. Soms komen data onder bepaalde voorwaarden beschikbaar op geaggregeerd niveau. We noemen enkele voorbeelden van bestaande, maar lastig te verkrijgen data.

Op het gebied van onderwijs bestaan kansen voor de analyse van leerlingvolgsystemen. Voor de sport zijn deze systemen relevant voor het volgen van het gymnastiekonderwijs. Deze data zijn in principe niet openbaar. De eigenaren zijn uitgevers die de systemen aan de scholen verkopen.

NOC*NSF beheert de ledeninformatie van de aangesloten sportbonden in Kiss. Met het systeem is te achterhalen waar de leden wonen, wat het geslacht en de leeftijd is, welke sport zij beoefenen, hoe lang ze lid blijven en van welke andere sportbonden zij ook lid zijn. NOC*NSF heeft de ambitie om een model van sportdeelname in het algemeen te maken (dus meer dan alleen lidmaatschap van een sportvereniging). Door dit soort koppelingen probeert NOC*NSF de betekenis van de data te vergroten.

Data afkomstig van enquêtes

Hoewel discussie kan ontstaan in welke mate data van enquêtes als big data zijn te beschouwen, willen we aandacht vragen voor informatie over sport uit bestaande (bevolkings)onderzoeken. Met deze vaak grote, landelijke onderzoeken zijn secundaire analyses uit te voeren. In bijlage III hebben we een paar voor sport belangrijke onderzoeken genoemd.

3.4 Openbare observaties

Via openbare observatie bestaan vele mogelijkheden voor het verzamelen van informatie. Met observatie bedoelen we niet alleen kijken, maar ook het verzamelen van gegevens via metingen met apparatuur zoals camera's. Zo heeft het CBS verkeersstellingen vervangen door informatie te verzamelen via duizenden verkeerslussen in de wegen. Zij weten daardoor hoeveel auto's op welk moment passeren. Een andere registratiemethode betreft de metingen van de verkeersdruk met mobiele telefoons. Door te meten hoeveel mobiele apparaten actief zijn op de wegen, is de lengte van files te bepalen. Uiteraard is ook geo-informatica als relevante big data te benoemen.

Geo-informatie

Een andere belangrijke vorm van big data zijn locatiegegevens gecombineerd met gebiedskaarten, de geo-informatie. Met deze kennis zijn lokale uitkomsten zoals bijvoorbeeld de sportdeelname snel te vergelijken. Geo-informatie is belangrijk voor de weergave van informatie, maar ook voor afstandsanalyses en andere ruimtelijke analyses. Hiermee zijn bronnen onderling te combineren zodat nieuwe inzichten ontstaan over (lokale) samenhangen. Google Maps is een belangrijke bron voor kaarten die wereldwijd gebruikt worden om informatie toe te voegen. De kaarten worden daarnaast massaal gebruikt om de weg te vinden. Aan de basis van Google Maps liggen satellietbeelden. In Nederland is de Basisregistratie Adressen en Gebouwen (BAG) belangrijk. Daarnaast bestaan talloze bestanden met geografische informatie (zie o.a. www.nationaalgeoregister.nl).

Het Kadaster stelt mede met satellietbeelden de functie van een gebouw of gebied vast. Dit gebruikt het Mulier Instituut voor de controle van locatie van sportvoorzieningen (Databestand Sportaanbod) van het Mulier Instituut. Veel informatie van de publicatie Sportaccommodaties in Nederland: kaarten en kengetallen (Van der Poel et al., 2016) is op deze database gebaseerd.

Meting bezoekersaantal sportevenementen

Bij grote sportevenementen in de openbare ruimte is het niet eenvoudig om het aantal bezoekers te tellen. De cijfers daarover zijn meestal opgaven van de organisatoren die deze cijfers meestal matig onderbouwen. Aannemelijk is dat het overschattingen zijn. We bespreken in het vervolg methoden om te tellen via gsm-signalen, wifi, bluetooth en een camerasysteem.

Recent is een alternatief voor het tellen van toeschouwers ontwikkeld en toegepast bij de Giro-etappes van 6-8 mei 2016 in Gelderland. Het aantal mobiele telefoons (van Vodafone) binnen een bepaald gebied en tijd geeft een indicatie van de belangstelling voor dit evenement. Met deze methode is overigens ook te bepalen waar de bezoekers wonen. Voor deze analyse is speciale zorg besteed aan de anonimiteit van de abonnees van de telecomprovider.

Vragen hierbij zijn of bekend is wat het gemiddelde aantal bezoekers langs het parcours is als er geen wielerevenement is zodat de netto bezoekers zijn te registreren. Daarnaast moet uiteraard bekend zijn hoeveel procent van de toeschouwers een abonnement bij de betreffende provider heeft en of deze verdeling ook opgaat voor wielervanliefhebbers. Het weer kan ook invloed hebben op dit soort verhoudingen. Het was op het betreffende weekend (6 en 7 mei 2016) volgens het KNMI 25-27 graden Celsius in het parcoursgebied (Gelderland, meetpunt Deelen). Vrij warm voor deze tijd van het jaar. Tot slot zal een deel van de bezoekers direct aan het parcours wonen waardoor het onzeker is of zij als bezoekers meegeteld kunnen worden en zal soms verkeer opgehouden worden om de wielerkaravaan te laten passeren (Decisio). Toch kan deze methode van betekenis zijn om het aantal toeschouwers van (sport)evenementen beter te tellen gezien het feit dat de toeschouwers over grote afstand zijn verdeeld.

Een vergelijkbare methode is het meten van bluetooth-signalen. In Vlaanderen hebben ze daarmee de bezoekers aan de Ronde van Vlaanderen geteld (Versichele et al., 2012). Ze maten daarvoor het percentage bezoekers met een bluetooth-sigitaal. De insteek van het onderzoek was niet alleen het tellen van bezoekers, maar ook het handhaven van de veiligheid.

De derde methode is gebaseerd op het tellen van wifi-signalen. Het bedrijf Locatus telt passanten van winkels en winkelgebieden in haar onderzoek Passanten 24/7. Zij registreren in winkelgebieden mobiele

apparaten waarvan de wifi aanstaat. Ook hiervoor bepalen zij met vooronderzoek welk percentage van de voorbijgangers een smartphone met een actieve wifi heeft. Mensen die niet willen deelnemen aan de meting kunnen zich centraal afmelden bij PrivacySIG. Zij registreren het IP-adres van het apparaat waarna tellingen de wifi van dit adres overslaan.

Een andere methode is het tellen van passanten/deelnemers met behulp van camera's (CrowdDynamics van Visense). Het systeem is in 2007 door Vinotion in samenwerking met de TU/Eindhoven ontwikkeld. Over een breedte van achttien meter kunnen zij simultaan passanten tellen. Het systeem is vooral bruikbaar voor tellingen in gebieden waar de camera's staan. Wellicht zijn drones met deze camerasystemen te ontwikkelen. In tegenstelling tot tellingen via bluetooth of wifi is niet vast te stellen hoe lang bezoekers blijven en of zij eventueel terugkomen.

Beeld en geluidsanalyse

In paragraaf 3.2.4 stipten we naast de mogelijkheden van tekstanalyse ook de kansen voor beeld- en geluidanalyse van radio- en televisieopnamen aan. Bij een succesvolle ontwikkeling hiervan ontstaan mogelijkheden om opnamen van bijvoorbeeld gymnastieklessen via dergelijke systemen te analyseren. Dit betekent dan een geautomatiseerd observatie-instrument. Voor bewakingscamera's bestaat al de mogelijkheid om afwijkend gedrag in mensenmassa's vast te stellen (zoals dronken jongeren). Voor voetbalwedstrijden op televisie maakt het Enschedse bedrijf Scisports gebruik van speciale beeldanalysetechnieken om een wedstrijd te analyseren.

4. Combinatie bronnen big data

Echte meerwaarde kan ontstaan door verschillende bronnen van big data te koppelen. Een voorbeeld hiervan ligt buiten de sport. Bij de presidentsverkiezingen van de Verenigde Staten in 2008 zette het campagneteam van de democraten sterk op big data in. Zij verzamelden statistieken met wijkprofielen en informatie hoe deze samenhangen met de stemvoorkeuren. Met deze gegevens konden ze twijfelende kiezers in belangrijke staten voorzien van gepersonaliseerde e-mails. Zij maakten ook op een slimme manier gebruik van sociale media door vooral op weblocaties waar veel twijfelende kiezers kwamen boodschappen achter te laten (nl.linkedin.com/in/timothyprescott, Obama for America).

De verbinding van de verschillende bronnen is de specialiteit van *datascience*. Dit vakgebied bevindt zich op het snijvlak van analysetechniek, specifieke domeinkennis zoals van sportdeelname en (journalistiek) speurwerk. Dat laatste heeft betrekking op zaken zoals deskresearch maar ook op *spidering* en zelfs *hacking*. De combinatie van verschillende bronnen is dus een belangrijk kenmerk van het succes van big data.

De Amerikaanse organisatie voor publieke opinie-onderzoek AAPOR concludeert dat vooral de combinatie van nieuwe technieken die big data genereren met meer traditionele methoden zoals vragenlijsten de toekomst hebben (AAPOR big data taskforce, 2015). Zij geven aan: *‘A preferred strategy is to use a combination of new and traditional data sources to support research, analytics, and decision making, with the precise combination depending on the demands of a given situation’*.

In het vervolg bespreken we enkele sportgerelateerde mogelijkheden voor de koppeling van data om informatie te verrijken.

Koppelen aan weerinformatie

Een interessante mogelijkheid om (ad hoc) gedrag te verklaren is om data te koppelen aan weersinformatie. Het KNMI biedt deze op haar website aan per weerstation. Interessanter is het als weerinformatie per uur beschikbaar is zoals bijvoorbeeld buienradar.nl doet. Deze informatie is vervolgens te koppelen aan de momenten dat mensen gaan wielrennen of hardlopen via Strava en Runkeeper. Vragen zouden kunnen zijn: in welke mate zijn mensen met vaste patronen door het weer te beïnvloeden en hoe zit dat met incidentele lopers/fietsers?

In de publicatie *Met het oog op de tijd* (Van den Dool & Tiessen-Raaphorst, 2013) is geprobeerd de KNMI weersinformatie te verbinden met dagboekgegevens van de tijdsbesteding per kwartier uit 2010. In het onderzoek (tijdsbestedingsonderzoek TBO) houden mensen gedurende een week bij welke activiteiten zij verrichten. De publicatie ging met name in op uitgaansgedrag (zoals dagjes uit, terrasbezoek, sporten, winkelen en bioscoopbezoek). Een sterk verband met het actuele weer was niet vast te stellen. Waarschijnlijk komt dit mede doordat de informatie op een dag was gebaseerd. Een ochtend heftige regen kan gevolgd worden door een goede middag om op een terras te zitten. Ook voor sporten bleek er geen speciaal verband met het weer te zijn. Een ander probleem was dat onbekend is bij welke weerkenmerken (temperatuur, wind, neerslagverwachting) mensen besluiten op een terras te gaan zitten. Uiteraard is daarbij niet alleen het actuele weer, maar ook de verwachting relevant. Hoe meer organisatie voor een activiteit nodig is, hoe meer de verwachting een rol zal spelen. Met een collega op een terras zitten, vraagt nauwelijks planning. Met een paar vrienden naar een pretpark is vaak ingewikkelder vanwege de planning van beschikbare vrije dagen. Voor met name dit soort activiteiten zal de verwachting een relatief grote rol spelen. Buiten Buienradar om bestaan wel websites die een

deel van de historische weersverwachtingen bewaren, maar de organisatie verzamelt deze niet systematisch (medewerker Buienradar via e-mail).

Invloed van sportprestaties op de maatschappij

Het Mulier Instituut heeft in 2010 geprobeerd om grote sportprestaties in een bepaald jaar te verbinden met een forse reeks algemene statistieken van CBS Statline (bloei economie etc.). Het idee was dat trots op sportprestaties of juist teleurstelling zichtbaar is in maatschappelijke kengetallen. Dit leverde vooralsnog weinig aanknopingspunten op. Wellicht omdat de mogelijke trots door sportprestaties niet lang genoeg beklijft (Elling et al., 2014) voor invloed op zaken als economische groei.

Kavetsos en Szymanski bestudeerden de relatie tussen het organiseren van grote sportevenementen en het optreden van gevoelens van geluk in de organiserende landen gedurende een periode van dertig jaar (2008, in: Cornelissen & Maennig, 2010). Hiervoor koppelden zij bestaande data over evenementen aan data over niveaus van nationaal geluk. Zij vonden een positief effect op de toename van geluk in een land bij de organisatie van een WK voetbal, maar niet voor de organisatie van Olympische Spelen en EK voetbal.

Koppelen persoonsgegevens aan uitkomsten enquêtes

Voor bijvoorbeeld vraagstukken over de samenhang van bewegen met cognitie bestaan er interessante koppelingen voor onderzoekers. Deze verbindingen zijn via het CBS mogelijk in geanonimiseerde vorm. Van de enquêtes van het CBS zoals de Gezondheidsenquête is bekend welk individu dat betreft. Daardoor is deze informatie te koppelen aan de eindexamenuitslagen die geregistreerd staan bij de Dienst Uitvoering Onderwijs (DUO) in Groningen of aan het medicijngebruik via Vektis in Zeist. De eindexamencijfers zijn te vergelijken met de informatie over de hoeveelheid sport en bewegen van kinderen die eerder aan de enquête deelnamen. Andersom is ook te zien wat de cijfers waren van mensen met bepaalde kenmerken (te weinig bewegen etc.). Andere geanonimiseerde koppelingen zijn mogelijk met inkomensgegevens, huizenprijs (Kadaster), persoonsverhuizingen (Sociaal Statistisch Bestand) en met voertuigregistratie. Koppelingen met bijvoorbeeld OV-chipkaart informatie over reizen zijn niet mogelijk.

Arbeidsmarktonderzoek sport: de toekomst koppelen databestanden?

Via de Dienst Uitvoering Onderwijs (DUO) is ook bekend welke opleiding mensen hebben gevolgd en voltooid. Deze gegevens zijn in theorie te koppelen aan gegevens over de werksituatie die bij de belastingdienst bekend zijn. Voorwaarde is uiteraard dat de werkgevers die werknemers bij hun belastingaangifte opgeven gecategoriseerd zijn zodat bekend is welke werkgever gelieerd is aan sport. Deze categorisering kan plaatsvinden via de registratie bij de Kamer van Koophandel. Lastig aspect blijft dan de verschillende spelwijzen van een naam die bij een specifieke werkgever kunnen worden gebruikt.

Van recent afgestudeerden aan een sportopleiding is via de beschreven methode globaal na te gaan in welke sector zij werken, welk inkomen zij hebben en wat hun achtergronden zijn (leeftijd, geslacht, burgerlijke staat, locatie etc.). Niet bekend is of zij daadwerkelijk een sportfunctie bij een sportorganisatie hebben. Daarnaast kunnen mensen zich bezighouden met een sportfunctie bij een werkgever die niet speciaal actief is in de sport. Voor dit soort detailinformatie en vragen over bijvoorbeeld tevredenheid is aanvullend (vragenlijst)onderzoek noodzakelijk. Via het CBS is het in

theorie mogelijk om specifieke personen voor onderzoek te selecteren. Het CBS zal dan zelf ook het belang van het onderzoek moeten inzien.

De beschreven opzet van koppeling van gegevens van afgestudeerden met belastinggegevens is vooralsnog een theoretische mogelijkheid. Het CBS zelf baseert haar arbeidsmarktonderzoek voorlopig nog op enquêtes (Enquête Beroepsbevolking, EBB). Desalniettemin vinden wel verkenningen plaats van de beschreven mogelijkheden (o.a. door KBA te Nijmegen).

Functie van locaties bruikbaar voor accommodatiedatabase

Het Nederlandse Kadaster probeert het land in kaart te brengen door precies de grenzen van percelen te registreren. Zij probeert daarbij zo goed mogelijk de functie van een perceel vast te leggen in de Basisregistratie Topografie. Voor sport onderscheiden zij de volgende gebouwen: sporthallen, overdekte ijsbanen, stadions, maneges en zwembaden. In het open veld registreren zij: crossbaan, golfterrein, jachthaven, kartingbaan, renbaan, skibaan, sportterrein, sportcomplex, tennispark, ijsbaan, zwembadcomplex. De uitkomsten zijn gebruikt voor de controle op volledigheid van het Databestand Sportaanbod (DSA) die het Mulier Instituut beschikbaar heeft. Daarnaast is deze informatie bruikbaar om vast te stellen wat de afstanden tot de voorzieningen zijn, de onderlinge afstanden maar ook de afstanden tot de wooncentra.

Locaties sportvoorzieningen gecombineerd met enquêtedata

In München heeft Wicker et al (2013) van een groot aantal mensen vastgelegd welke sportvoorzieningen in de buurt aanwezig zijn. Via een enquête heeft zij vervolgens het gebruik daarvan achterhaald. In de analyse hebben zij vastgesteld dat er een verband is tussen sportdeelname en de aanwezigheid van sportvoorzieningen. In Nederland vonden Van de Heuvel et al (2011) en Hoekman et al (2016) een dergelijk verband overigens niet.

Relatie voetbaluitslagen en ongewenst gedrag in het Verenigd Koninkrijk

Voor dit rapport hebben we een case uitgewerkt waarin we onderzoeken of er een relatie is tussen de uitslagen van voetbalwedstrijden en registraties van ongewenst gedrag. We hebben voor een concrete case gekozen om te laten zien welke beslissingen nodig zijn om verschillende vrij beschikbare databronnen te gebruiken. We hebben de uitslagen van voetbalwedstrijden van het seizoen 2012/2013 gekoppeld aan de registraties van ongewenst gedrag en criminaliteit in het gebied waar de fans van een voetbalclub wonen. De uitkomsten laten zien dat winst of verlies (enige) invloed heeft op de mate van ongewenst gedrag (zie bijlage II).

Uit de case leren we dat voor de combinatie van databestanden compromissen noodzakelijk zijn waardoor de onderzochte relatie verwatert. De lastigste vraag die we daarbij hebben moeten beantwoorden, is hoe we bepalen waar het thuispubliek van een club woont. Daarnaast waren de registraties van het ongewenste gedrag slechts per maand beschikbaar terwijl de meeste incidenten rond de wedstrijddag zelf plaats zullen vinden. Voor een beschrijving van de resultaten, de methode en een discussie zie bijlage II.

5. Reflectie: betekenis big data voor sport

In dit hoofdstuk bespreken we onze centrale vraag wat de betekenis is van big data voor sportonderzoek. We beginnen deze reflectie met enkele algemene plus- en minpunten van big data in een methodologisch perspectief (paragraaf 5.1). Vervolgens gaan we in op de kansen voor de sectoren die we in de twee eerdere hoofdstukken hebben besproken (paragraaf 5.2). Tot slot geven we de ambities van het Mulier Instituut voor big data weer (paragraaf 5.3).

5.1 Methodische implicaties big data voor sportonderzoek

In de beschrijving van big data (hoofdstuk 2) hebben we de kansen en beperkingen van big data besproken. In deze paragraaf bestuderen we deze aspecten in perspectief van het sportonderzoek. We richten ons daarbij specifiek op onderzoeksmatige aspecten.

Beperking big data: representativiteit

In sociaalwetenschappelijk onderzoek ligt een sterke nadruk op representatief onderzoek. Met steekproeven proberen onderzoekers een goed beeld van een populatie als geheel te verkrijgen. Apps die sportgedrag zoals wielrennen en hardlopen bijhouden, zijn doorgaans niet representatief voor alle lopers in een land/periode. Veel mensen lopen hard of fietsen zonder dit bij Strava of Runkeeper te registreren. Mensen die dat wel doen hebben niet altijd zin om een smartphone mee te dragen bijvoorbeeld vanwege voorspeld slecht weer of omdat het een dagelijkse tocht betreft die weinig extra informatie oplevert. Niet-sporters en niet-gebruikers van apps en beweegmeters blijven geheel uit beeld. Discussie of beweegmeters vragenlijsten kunnen vervangen is door dit soort selectief gedrag voorlopig nog niet aan de orde. Wel zijn goede metingen van beweeggedrag in het kader van een onderzoek makkelijker geworden.

Ook uitingen op sociale media zijn selectief. Actievere gebruikers worden daarbij meerdere keren geteld. De uitspraken over Twitter gaan vaak niet over de Twittersaars in het algemeen, maar over het aantal tweets over een onderwerp. Actieve gebruikers zijn veel vaker dan gemiddeld opinieleiders, politici en journalisten.

Andere vormen van big data die we beschreven, hebben minder last van problemen met de representativiteit. Verkeerslussen zijn redelijk objectief in het tellen van het aantal voertuigen en het verzamelen van uitslagen van wedstrijden is uiteraard ook betrouwbaarder dan het publiek te ondervragen over de uitslag van hun laatste bezochte voetbalwedstrijd. Ook bij het meten van zoekgedrag op internet is representativiteit geen kwestie.

Bij big data mag dan vaak sprake zijn van overvloedige hoeveelheden data, maar dat betekent niet altijd dat er voldoende data zijn. Voor sociale media onderwerpen betreft dat bijvoorbeeld kleine of lokale gebeurtenissen. Bovendien zijn de data sterk scheef verdeeld. Denk aan bijvoorbeeld data over sportaccommodaties waar van één regio niet of nauwelijks data beschikbaar zijn en van een andere veel.

Beperking big data: verkrijgbaarheid

De theoretische potentie van big data is groot. Door informatie te koppelen ontstaat veel informatie over bijvoorbeeld een individu. Hoewel veel big data verkrijgbaar zijn, blijft ook veel onbereikbaar en

ontoegankelijk omdat het privacygevoelig is (reisdata van de OV-chipkaart) of omdat bedrijven en instellingen het vanwege strategische redenen niet willen afgeven/verkopen. In het grijze gebied liggen data die gekocht kunnen worden en data waar veel inspanning voor nodig is om deze te verkrijgen of waar veel bewerking voor noodzakelijk is. Data die momenteel redelijk eenvoudig verkrijgbaar zijn, zoals van beweegapp Strava, hoeven dat in de toekomst vanwege commerciële redenen niet meer te zijn. Overheden kunnen privacygevoelige informatie van burgers beter gaan beschermen. Ook dat kan tot verminderde verkrijgbaarheid van data leiden. Het gebruik van persoonsgegevens is gereguleerd: de regels zijn vastgesteld door het College Persoons Bescherming (zie autoriteitpersoonsgegevens.nl). Volgens de WRR (2016) bestaat een achterstand in de wetgeving voor analyse en gebruik. Bij het gebruik van buitenlandse websites (Google, Facebook, etc.) gelden Nederlandse wetten dikwijls niet.

Voor veel onderzoekers is een groot deel van de big data dus niet bereikbaar. Voor medewerkers van Google en Facebook is de situatie wezenlijk anders. Voor hen is het internetzoekgedrag of wat mensen op Facebook zetten een ware goudmijn. Universiteiten met voldoende contacten met deze bedrijven kunnen wel data krijgen.

Beperking big data: relevantie uitkomsten

Het beleid heeft logischerwijs de voorkeur voor effectieve maatregelen. Kennisinstituten onderzoeken potentiële maatregelen daarop. Ook big data analyses vinden in principe plaats om de grootte van een effect vast te stellen. De uitkomsten van deze analyses geven soms echter vooral kleine effecten aan. Voor marketingdoeleinden is sprake van een duidelijke verbetering als twee in plaats van één procent van de websitebezoekers op een internetadvertentie klikt. Voor beleid gelden echter andere maatstaven. Uiteraard kan een analyse met big data serieuze verbanden opsporen, maar met veel data zijn ook minieme relaties te vinden die slechts zelden relevant zijn. De mate van verklaring is dus ook bij big data een belangrijk vraagstuk.

Voor onderzoekers is het vaak “veel zeven voor het vinden van een klompje goud”, aldus onderzoeker Piet Daas van het CBS. Hij maakt zich ook zorgen over de reproduceerbaarheid van data. “Big data vertalen naar statistieken is dé grote uitdaging,” zo verklaarde hij in een interview (Relatiemagazine CBS, zomer 2013). Voor het CBS is het belangrijk om toegevoegde waarde uit zowel gestructureerde als ongestructureerde data te halen. Daarom zetten zij op big data in. Onderzoeker Jurjen Iedema van het Sociaal en Cultureel Planbureau meldt in een mailwisseling: “Vroeger leerden wij dat ‘*unobtrusive measures*’ ideaal waren: bijvoorbeeld dat met de slijtage van het tapijt voor schilderijen was af te leiden welke schilderijen populair waren. Dit is natuurlijk heel leuk. Maar bij big data ligt iets vaak zeer voor de hand (met mooi weer, meer sport) en als er iets bijzonders is, blijft de vraag: we vinden iets, maar wat betekent dat nou eigenlijk?”

Voor big data zijn dus zelden ‘ideale’ data beschikbaar. Dit betekent compromissen wat betreft de validiteit en soms genoegen nemen met data die een bepaald aspect globaal beschrijven. Daarin zijn deze data niet anders dan het gebruik van een gestructureerde dataverzameling zoals interviews of experimenten. Ook hiermee zijn veel vraagstukken niet (perfect) te beantwoorden.

Wegstrepn beperking: big data het einde van de theorie?

In de omgang met big data lijken twee tradities te bestaan. Aan de ene kant zijn dat met name wiskundigen en econometristen die zich richten op het ‘kraken van data’. Met methoden zoals *machine learning* zoeken zij naar verbanden in data. Een vraagstelling of theorie is daarvoor minder relevant.

Aan de andere kant staan vooral sociaalwetenschappelijke onderzoekers met kennis van een bepaald veld die uitgaan van theorieën en deze vervolgens toetsen via onderzoek en analyse. Deze onderzoekers stellen vaak eisen aan big data en zijn geschoold om verklaringen voor uitkomsten in de context te zoeken. Een belangrijke woordvoerder van deze traditie was Karl Popper (1959) met eisen voor falsifieerbaarheid van onderzoeksresultaten.

Wat zijn nu de plus- en minpunten van deze twee tradities? Wat kunnen zij van elkaar leren? Onderzoek zonder theorie of inbedding in een onderzoeksveld kan resulteren in losse flodders. Wellicht halen deze de media, maar de informatie beklijft minder goed omdat deze niet opgenomen is in de kennis-infrastructuur van het sportonderzoek. Uiteraard kan een onbekende partij een frisse blik op een onderzoeksveld werpen. Voor de ‘traditionele’ onderzoekers kan speuren in big data zonder enige achtergrondgedachte waarde hebben om met de gevonden resultaten hypothesen te vormen (exploratief onderzoek). In een nieuw onderzoek of met een andere dataset zijn de nieuwe veronderstellingen te toetsen. In wezen gebeurt het voortdurend dat wetenschappers verklaringen zoeken voor onverwachte of tegenstrijdige uitkomsten en vervolgens nieuwe theorieën ontwikkelen. Exploratief onderzoek hoort bij de wetenschapsontwikkeling (zie het werk van Tukey (1977) hierover).

De ‘toevallige’ uitkomsten van *machine learning* zijn ook te testen met andere data of in een andere periode. Blijken de uitkomsten consistent, dan zet dat wetenschappers aan tot het zoeken van verklaringen. Jin et al (2015) reppen overigens van situaties die te complex zijn voor een theoretische beschrijving. De vraag komt op wat te doen met uitkomsten die een wezenlijk voordeel opleveren, zoals de ontwikkeling van medicijnen door vele chemische stoffen tegelijkertijd te onderzoeken, maar die theoretisch niet goed onderbouwd zijn? Het zijn vragen die de wetenschapsleer de komende tijd bezig zullen houden. Het einde van de theorie zoals Anderson (2014) deze voorspelt, valt daarbij overigens niet te verwachten. Zonder deskundige interpretatie zijn data waardeloos en zonder theorie is wetenschap richtingloos. Aarnout Brombacher van TU Eindhoven (lid van het Topteam Sport) gaf in ons gesprek over big data aan dat data zonder kennis niet te begrijpen is.

Onderzoek zonder big data?

Voor sociaalwetenschappelijk sportonderzoek zijn er nog volop mogelijkheden voor onderzoek via een vragenlijst of een experiment buiten big data om. Dit geldt ook voor beleidsonderzoek waar theorietoetsing doorgaans geen grote rol speelt. De schaal van de gewenste informatie biedt kansen voor ‘gewoon’ onderzoek. Een internationale analyse van de beoordeling van mannen en vrouwen in het televisienieuws is anders dan een analyse van de strategie van de sportbonden in Nederland of de vaststelling van de effectiviteit van buurtsportcoaches. Onderzoek met big data zou vooral interessant kunnen zijn om snel en met een beperkt budget een indicatie te krijgen van het te onderzoeken onderwerp. Daarnaast komen onderwerpen in aanmerking die minder makkelijk met traditionele methoden zijn te onderzoeken (zoals weerregistraties). Exclusief voor traditionele methoden lijken onderzoeken waarvoor een goede representativiteit is vereist en waarbij informatie van individuen essentieel is. De combinatie van big data en traditionele methoden zal terrein winnen.

5.2 Kansen big data voor sportonderzoek

In hoofdstuk 2 lasen we al dat de term big data niet eenduidig is en vele vormen kent. Dit geldt ook voor het gebruik voor sportonderzoek. Daarom is het zinvol om het nut voor sportonderzoek naar bronnen te differentiëren. We hanteren hiertoe de eerder geschetste indeling om per doel kort de kansen en beperkingen te bespreken. Zoals eerder aangegeven geldt ook hier dat de kracht van big data

vooral in de combinatie van bronnen is gelegen. Deze stelling onderschrijft ook lid van het Topteam Sport Aarnout Brombacher in ons gesprek over big data.

(Beweeg)meters en beweegapps

Hoewel beweegmeters om de mate van inspanning vast te stellen steeds belangrijker worden, blijft de interpretatie vooralsnog lastig. Daarnaast bestaat momenteel een grote mate van selectiviteit van het gebruik (niet alle sportbeoefenaars en niet op alle momenten). De informatie van apps van bijvoorbeeld Strava en Runkeeper kennen een te beperkt gebruik om waardevol te zijn. Het blijft steken in stukken van routes die vaker worden gebruikt. Informatie die een waarnemer ter plekke ook ziet. Interessanter wordt het als het onderdeel is van een experiment waarin groepen worden vergeleken, maar dat lijkt meer op traditioneel onderzoek met aangepaste technische middelen.

Een app kan wellicht een rol spelen om mensen aan het sporten en bewegen te krijgen. Een vraag die daarbij opkomt is waarom een app meer zou stimuleren tot sporten en bewegen dan een trainer of een sociaal verband. Een app is wat kosten betreft wel een stuk voordeliger dan een trainer. Daarnaast kunnen gebruikers op hun eigen momenten trainen. Een ander vraagstuk is of mensen die onvoldoende sporten of bewegen handig zijn met apps op smartphones. Of het daadwerkelijk leidt tot een grotere beweegdeelname is nog onduidelijk. Niet-sporters aanzetten tot sport en bewegen is uiteindelijk een gecompliceerd proces (Van den Dool, 2015).

(sociale) media

Het media-onderzoek is het meest ongestructureerd vanwege de vele bronnen die beschikbaar zijn. De analyse van de inhoud van de traditionele media zoals boeken, kranten, radio en televisie bestaat al lange tijd. Maar door nieuwe technieken voor analyse en dataverwerking is of wordt het onderzoek hiernaar gaandeweg steeds effectiever. Of het daardoor een grotere plaats in het domein van sociaalwetenschappelijk sportonderzoek krijgt, is de vraag.

Sociale media zijn een relatief nieuwe bron. Doordat alleen Twitter redelijk openbaar is, vinden hier de meeste analyses mee plaats. Het veel belangrijkere Facebook is slechts voor een klein deel openbaar. Twitter is vooralsnog uitstekend te benutten voor het inkleuren van mediaverhalen. Onduidelijk blijft of hiermee serieuze vragen zijn te beantwoorden over de maatschappelijke positie van sport. Opinieleiders zijn duidelijk oververtegenwoordigd onder de Twittergebruikers. Voor minder grote thema's zoals specifieke voetbalwedstrijden laat het aantal tweets te wensen over. Op landelijke schaal hangen tweets sterk samen met het economische humeur van de Nederlander (Daas et al., 2013). Een nadeel van sociale media voor onderzoek is dat niet alle berichten van bijvoorbeeld Twitter bewaard blijven. Het bedrijf Coosto heeft daarom zelf het initiatief genomen om zo veel mogelijk tweets te bewaren.

Internet is behalve een belangrijke opslagplaats voor documenten (documentanalyse) een bron van informatie om rapporten van onderzoek in te kleuren (deskresearch). Gegevens over zoekgedrag naar voor sport interessante termen levert weer nieuwe informatie op. Uiteraard beïnvloeden deze zaken het sportonderzoek in positieve zin.

Via *webcrawlers* is zoeken op internet te automatiseren. Hiermee ontstaan nieuwe dataverzamelingen die bijvoorbeeld het Databestand Sportaanbod van het Mulier Instituut kunnen versterken.

Dataregisters

Data uit registers zijn sterk gestructureerd doordat ze doorgaans een administratief doel hebben (zoals bijhouden van autokentekens in Nederland). Ingewikkelder wordt het als meerdere registers

gecombineerd dienen te worden. Het gebruik is vooral afhankelijk of de data vanwege privacykwesties toegankelijk zijn. Daarnaast zullen bedrijven met interessante verzamelingen hun kostbare data niet zonder meer afstaan. Voor overheidsbestanden bestaat het streven om steeds meer data voor algemeen gebruik ter beschikking te stellen. Voor sportonderzoek levert dat soms waardevolle data op zoals van de gemeentelijke financiën (IV3). Daarnaast kunnen registers enquêtes verrijken zoals bijvoorbeeld een RIVM-analyse van enquêtedata over beweeggedrag (Gezondheidsenquête CBS) en registerdata van medicijngebruik (Vektis). Overigens gelden voor deze datakoppeling strenge eisen aangezien het om gevoelige informatie gaat.

Observaties in de openbare ruimte

In principe zijn metingen in de openbare ruimte heel divers van karakter zoals tellen van voertuigen over een meetlus of het tellen van mobiele telefoons in een bepaald gebied. De openbare metingen zullen in de toekomst belangrijker worden. Voor sportonderzoek lijkt de inzet voor het tellen van deelnemers aan evenementen en de bepaling van hun herkomst belangrijk te worden. Dit geschiedt via de meting van smartphonezenders van mensen in de buurt van het evenement. Na meerdere metingen kan worden vastgesteld wat de haken en ogen van deze methode zijn.

5.3 Big data ambities Mulier Instituut

Samenvattend stellen we dat big data als bijzonder betekenisvol gelden voor de verdere doorontwikkeling van het (breedte-)sportonderzoek. Tegelijk signaleren we ook dat ook big data duidelijke grenzen kennen, bijvoorbeeld vanwege de representativiteit van de data en de verkrijgbaarheid. Er zijn dikwijls meerdere, complexe bewerkingsslagen nodig om tot resultaten te komen en zonder contextuele en theoretische kennis is ook bij big data geen betekenis aan de uitkomsten van de analyses te geven. De kansen van big data hebben we uiteengezet met als uitgangspunt vier bronnen die we in dit rapport beschrijven. De bronnen zijn persoonlijke meters/apps, (sociale) media, dataregisters en (openbare) observaties. Als Mulier Instituut zien we de kracht van big data vooral waar dergelijke gegevens worden ingezet in combinatie met andere dataverzamelingen (zoals een enquête).

In de aankomende jaren zal het Mulier Instituut verdere investeringen plegen in het werken met big data. De ambitie is om in 2020 voldoende veelsoortige ervaring met big data te hebben opgedaan om op projecten waar dat relevant voor is big data zinvol te kunnen inzetten. Voor de aankomende twee jaren staan hiertoe de volgende projecten op stapel:

1. *Sportuitgaven gemeenten.* Doorontwikkeling van de analysemogelijkheden op big data over overheidsuitgaven ('IV3 staten') en koppeling daarvan aan andere gegevensbronnen zoals over sportdeelname en gebruik van sportvoorzieningen.
2. *Metten sportklimaat.* Voor het monitoren van (on)gewenst gedrag is het 'real-time' vastleggen van de vele kleine handelingen die samen de sfeer maken belangrijk. Uit een pilotstudie bleek dat schriftelijke verslaglegging te wensen overlaat omdat er te veel tegelijk gebeurt. Met de gebruiksvriendelijke SportklimaatApp van het Mulier Instituut neemt een onderzoeker meer (on)sportieve handelingen waar. In samenwerking met hogescholen willen we een database opbouwen met de beoordeling van honderden wedstrijden. Hiermee kunnen we analyseren naar sporttak, geslacht en leeftijd van de sporters, de speelregio en het scoreverloop.

3. *Benutten Databestand Sportaanbod (DSA)*. In de Nationale Fiets Telweek in september 2016 hebben ruim 40.000 Nederlanders hun fietsgedrag geregistreerd met de Fietstelapp. Hiermee weten we welke routes Nederlanders benutten, hoe snel/wanneer ze fietsen en waar fietsen tot oponthoud leidt. We willen de fietsroutes die beginnen en eindigen bij een sportaccommodatie uit de DSA analyseren. Vragen zijn hoelang een rit duurt, op welke momenten wordt gefietst en naar welk type accommodaties zij fietsen.

Een tweede activiteit is de inzet van zogeheten ‘webcrawlers’ voor het permanent en automatisch actualiseren en aanvullen van DSA over bijvoorbeeld tariefontwikkeling.

4. *Volgen van sportevenementen*. Ten eerste is er de ambitie om ervaring op te doen met het schatten van het aantal bezoekers van gratis toegankelijke sportevenementen in de openbare ruimte aan de hand van het meten van smartphone-activiteit. Ten tweede is er het voornemen om de beleving van sportevenementen (deelnemers/volgers) in beeld te brengen met sociale media en/of apps. Ten derde is het interessant om kijkcijfers via televisie of internet van sportevenementen te koppelen aan de kenmerken van deze kijkers.
5. *Verenigingsinformatie*. Het Mulier Instituut verzamelt als kenniscentrum zoveel mogelijk informatie over sportverenigingen onder andere via de Verenigingsmonitor. Het streven is om aan dit onderzoek data te koppelen zoals de kantine-informatie via de kassasystemen van Le Credit Sportif en financiële gegevens van de Stichting Waarborgfonds Sport.
6. *Maatschappelijke betekenis sport*. We bezien waar big data kunnen bijdragen om de betekenis van sport te onderzoeken, zoals bij de koppeling van eindexamenresultaten van middelbare scholieren aan het latere beweeggedrag (relatie tussen lichamelijke activiteit en cognitie).
7. *Werkzaamheden Mulier Instituut voor Sia-Raak project* “Voor iedereen een app? Populaire sportapps en hun mogelijkheden en beperkingen voor breedtesporters én professionals”. (www.fontys.nl/Sportfolio/Projecten/Voor-iedereen-een-App.htm).
8. *Kennisuitwisseling en kennisdeling*. Het Mulier Instituut is voornemens om zijn eigen ervaringen met big data en die van anderen te delen in een serie bijeenkomsten met deskundigen en geïnteresseerden (en daar verslag van uit te brengen via meer en minder traditionele rapportagevormen).

Bijlage I. Hulpmiddelen verzamelen big data

Behalve concrete verzoeken aan eigenaren van websites, zijn data ook te verkrijgen via een geautomatiseerd systeem. De methode bestaat uit het contact zoeken met een website (API). Een tweede methode bezoekt een website en zoekt daar specifieke gegevens op (*Spiders*). Dit laatste is relevant voor bijvoorbeeld het bijhouden van entreetarieven van sportaccommodaties.

Databron API

Veel data zijn via dit communicatieprotocol tussen websites te benaderen. Onderliggend aan een website van bijvoorbeeld Facebook loopt een API mee (*application programming interface*). Informatie hoe dit in zijn werk gaat bij een specifieke website is met een verzoek aan de website te downloaden. Met deze informatie is het verzoek te verfijnen. Het is daarmee dus vooral een communicatiemethode. Grote websites hebben speciale pagina's over de beste benaderwijze via API's. Zie voor Twitter de website <https://dev.twitter.com/> of voor Strava <http://labs.strava.com/developers/>. Hiermee zijn Twitterberichten op een website te zetten of sportdata te downloaden. Websites hebben twee mogelijkheden om de te downloaden informatie te beperken. De meeste websites beperken het aantal verzoeken per tijdseenheid. Is het aantal verzoeken bereikt, dan gaat een wachttijd in. Een tweede manier is door slechts een deel te sturen (*pagination*). Beide methoden zijn bedoeld om te voorkomen dat websites vastlopen. Op www.github.com delen veel ontwikkelaars methoden om via API met websites te communiceren.

In principe is API vanuit het statistische programma R aan te sturen. R is gratis te downloaden en kent veel mogelijkheden om big data te verwerken. R dient als cliënt in API. In R moet een *package* aangemaakt worden om de API te maken. Via <http://metro.strava.com/> zijn Strava-data beschikbaar.

Databron Spiders

Deze methode schraapt inhoudelijke informatie van websites. Strikt genomen verloopt dat ook met de hiervoor beschreven API's (het is een verzamelterm). Google maakt gebruik van deze methode om een index van alle onderwerpen te maken. Door Google als zoekinstrument te gebruiken, krijgen we een lijst met websites waar de informatie te vinden is. Dezelfde methode is te gebruiken om specifieke informatie van websites te destilleren en wel op een geautomatiseerde wijze.

Naast de naam *spiders* worden namen als *crawlers*, *botnets* en *scrapers* gebruikt. Zie voor een voorbeeld www.parsehub.com (Turn dynamic websites into APIs). Een specifieke webpagina/site is te *scrapen* maar ook een serie websites die aan bepaalde voorwaarden voldoen.

Hoewel de meeste websites zichtbaar willen zijn via een zoekopdracht van Google zijn sommige websites ook niet gediend van het overnemen van content. Een voorbeeld van ongewenste overname is een nieuwswebsite die nieuws van andere nieuwswebsites overneemt om een samenvatting van het nieuws te maken. *Spiders* lezen in eerste instantie zogenaamde metatags en zien ook hoever zij een site mogen doorzoeken (*index-follow tag*). Een methode om informatie minder toegankelijk te maken is het veranderen van de hyperlink. Bij het doorzoeken van de nieuwe tarieven van een zwembad kan dit natuurlijk hinderlijk zijn.

Bijlage II. Voorbeeldstudie: voetbal en ongewenst gedrag

In hoofdstuk 4 hebben we de mogelijkheden van de combinatie van databestanden besproken om maatschappelijke vraagstukken te bestuderen. Daarvan hebben we hier een voorbeeld uitgewerkt. We hebben daarbij gebruikgemaakt van enkele openbare databronnen. De eerste betreft voetbaluitslagen van de Engelse Premier League en het tweede bestand zijn criminaliteitscijfers van de Engelse politie. Leidt een bepaalde uitslag tot meer of minder ongewenst gedrag in de steden waar de fans vandaan komen?

Inleiding en vraagstelling

Hoewel identificatie met sportteams met positieve aspecten wordt geassocieerd zoals een grotere zelfwaarde (Branscombe et al., 1991) blijkt ook dat de betrokken fans minder controle hebben over hun gedrag (agressie) dan fans die zich minder sterk met een sportteam identificeren (Dimmock et al., 2005). Mogelijk hangt een verminderde controle over het eigen gedrag samen met de prestaties van het eigen voetbalteam. Hieruit volgt de vraag die we met beschikbare data willen beantwoorden: hoe hangen de prestaties van een voetbalteam samen met de mate van ongewenst gedrag dan wel criminaliteit?

De verwachting is dat slechtere prestaties van een voetbalteam leiden tot een toename van antisociaal gedrag door de fans van dat team. Sommige teams hebben echter op voorhand minder kans op succes. Fans van dergelijke teams kunnen wellicht beter met verlies omgaan dan fans van teams waar op voorhand veel van verwacht wordt. De hypothese is dus tweeledig:

- A. Fans van verliezende teams zullen meer antisociaal gedrag vertonen dan fans van niet-verliezende teams (winnen of gelijk spelen) en
- B. deze relatie is sterker voor fans van op voorhand hoog aangeschreven teams dan voor fans van op voorhand laag aangeschreven teams.

Gebruikte datasets en bewerkingsmethode

Drie datasets hebben we voor deze studie gebruikt:

- Maandelijks registratiecijfers van verschillende soorten criminaliteit en antisociaal gedrag in Engeland en Wales zijn gevonden op www.ukcrimestats.com. Deze zijn op verschillende niveaus beschikbaar, onder andere per 'district'.
- Voetbaluitslagen van de Premier League zijn gehaald van www.football-data.co.uk. Zij registreren voor elke gespeelde wedstrijd gegevens zoals het aantal gescoorde goals, rode/gele kaarten en overtredingen.
- Van Engeland en Wales zijn in de periode augustus tot en met december 2012 geotagged tweets verzameld over voetbalteams uit de Premier League van het seizoen 2012-2013. De cijfers zijn toegankelijk gemaakt door onderzoekers en geven weer hoeveel getweet is over een team in een district. Hiermee bepalen we waar het thuispubliek woont.

Van de voetbaldata zijn de wedstrijdgegevens van een club per maand opgeteld. Hierdoor sluiten deze gegevens aan bij de beschikbare registraties van criminaliteit en antisociaal gedrag. De voetbaldata zijn vervolgens zo geordend dat voor elke wedstrijd een uit- en thuisteam geïdentificeerd werd en voor beide clubs een aantal wedstrijdgegevens werd vastgelegd. Voor elke club is uiteindelijk het aantal goals per maand meegenomen, evenals het aantal gele en rode kaarten per maand. Als indicator voor de

prestaties kreeg elke winstpartij een 1, elk gelijkspel een 0 en elke verliespartij een -1. Per maand zijn deze clubprestaties weer opgeteld. Clubs die vier wedstrijden in een maand hebben gespeeld, kunnen dus een maximale prestatie van +4 hebben en minimale prestatie van -4.

Voor onze hypothese is het belangrijk vast te stellen waar het thuispubliek woont. Omdat in met name grote steden meer voetbalteams in de Premier League spelen, bleek het onmogelijk om alleen met een kaart de woonlocatie van het thuispubliek vast te stellen. Dit hebben we opgelost door gebruik te maken van Twitterberichten in de periode 2012-2013 die per district beschikbaar zijn. Per bericht is vastgesteld over welke club is getwitterd en in welk district de Twitteraars woonden. De aanname is dat mensen fan zijn van een club uit de plaats waar ze wonen.

Voor gebieden met één actieve Premier League club is de thuisbasis meestal duidelijk. De Engelse regio's bestaan uit 'areas' en daarbinnen 'districts'. Meestal beslaat een area de stad waar de club speelt plus een gebied eromheen. Een aantal clubs ligt meer in een (groter) landelijk gebied. Voor deze clubs zijn met Google Maps visueel de meest verstedelijkte districten geselecteerd.

Er zijn vier steden waarin meerdere clubs in de Premier League actief zijn: Liverpool, Manchester, Birmingham en Londen. Van deze steden is eerst het grootstedelijk gebied als uitgangspunt genomen. Elk district is vervolgens gekoppeld aan de club waar het meest over getwitterd is. In Londen is echter geen enkel district waarin Fulham populairder is dan de andere Londense clubs. Fulham is om die reden niet meegenomen in de analyse.

Vervolgens zijn per district de maandelijkse criminaliteitscijfers gekoppeld aan de thuisbasis van de clubs. De cijfers van alle districten die bij een club horen, zijn voor de analyse samengenomen en vervolgens gekoppeld aan de maandelijkse voetbalgegevens van een club. Zo hebben we een bestand waar de onderzoekseenheid een speelmaand per Premier League club is. Per eenheid weten we de maandslag van de club en de hoeveelheid ongewenst gedrag in het fangebied van de club.

Voor hypothese B is tenslotte een indicatie gemaakt van het verwachtingspatroon van elke club voor dat seizoen. Aangenomen is dat fans van clubs die in het vorige seizoen beter presteerden dat voor dit seizoen ook van hun club verwachten. De ranking van het seizoen 2011-2012 is daarom genomen als indicator voor de verwachtingen van het seizoen 2012-2013, resulterend in een schaal van 1 tot 20 waarbij 20 de kampioen van het vorige seizoen is (Manchester City) met de hoogste verwachting.

We hebben voor het seizoen 2012/2013 gekozen omdat van alle drie de databronnen informatie over deze periode beschikbaar is. Het voetbalseizoen in de Premier League loopt van augustus tot mei.

We gebruiken de termen criminaliteit en antisociaal gedrag. Onder antisociaal gedrag verstaat UK crimestats gedrag dat hinder geeft aan personen en/of omgeving. Criminaliteit is een samenvattend begrip. Te denken valt aan diefstal, inbraak, zakkenrollen, schade aan infrastructuur/gebouwen, brandstichting, drugsgerelateerde overtredingen, openbare ordeverstoring en geweldsmisdrijven.

Voor hypothese A en B zijn elk twee OLS regressie-analyses uitgevoerd waarbij criminaliteit respectievelijk antisociaal gedrag voorspeld werd uit de prestaties van een club. Voor hypothese B is een moderatie-effect van de resultaten van de club met de verwachtingen van de fans opgenomen. Elk team komt tien keer voor in de data (tien maanden). In verband hiermee is gecontroleerd voor clustering binnen teams. De analyses zijn verricht met Stata.

Resultaten

De voetbalslagen correleren laag, maar positief met criminaliteit (.20) en antisociaalgedrag (.18). Deze resultaten zijn net als de regressiecoëfficiënten weergegeven in tabel 1. Er is verder een aanwijzing dat slechtere maandelijkse voetbalresultaten van een club leiden tot een toename van criminaliteit in de thuisbasis (p -1sided=.0495). Ook antisociaal gedrag neemt iets toe als de voetbalresultaten afnemen (p -1sided=.0515). Criminaliteit en antisociaal gedrag worden niet beïnvloed door een negatief spelbeeld (weinig doelpunten, veel overtredingen of rode kaarten). De gevonden effecten verdwijnen wanneer we controleren voor verwachtingen van fans, plus wanneer we een moderatie-effect toevoegen van de verwachtingen met de maandresultaten.

Tabel 1. OLS regressie coëfficiënten voor hypothese a en b (N=190).

	Hypothese A				Hypothese B			
	Criminaliteit		Antisociaal gedrag		Criminaliteit		Antisociaal gedrag	
	B	SE	B	SE	B	SE	B	SE
Resultaat	482.46	277.33	137.27	79.97	237.81	374.44	123.82	142.02
Goals	89.57	163.19	1.49	46.65	54.24	103.81	-6.52	30.58
Overtredingen	-29.34	43.81	-10.20	13.12	-25.88	35.62	-9.21	11.11
Rode kaarten	-269.28	434.98	-73.30	148.57	-381.60	425.70	-103.44	126.73
Verwachtingen					84.11	235.90	24.09	64.07
Verwachting x resultaat					11.02	28.17	-1.86	9.72
Constante	6157.84	2013.43	2238.60	591.04	5302.68	1877.24	2009.98	633.78
R ²	.047		.044		.053		.051	
F(df)	1.07(4,18)		1.40(4,18)		1.07(6,18)		1.08(6,18)	

In alle modellen is de F-waarde niet significant. Er is dus geen reden om aan te nemen dat deze modellen een verbetering zijn ten opzichte van het null-model.

Discussie

Door drie datasets te combineren zijn twee hypothesen getest met betrekking tot ongewenst gedrag en voetbalresultaten. De modellen bleken geen verbetering ten opzichte van het null-model. Een aantal factoren kunnen hier een verklaring voor zijn.

Vaststellen thuisbasis clubs. De voetbalberichten op Twitter betroffen enkel berichten die gepaard gingen met de officiële *hashtag* van een voetbalclub. Tweets die deze *hashtag* niet hadden zijn dus niet meegenomen in de telling. Daarnaast is van slechts een klein deel van de twitteraars de geografische locatie zichtbaar. De schatting ligt tussen 1% en 3% van de twitteraars, maar vormen hoe dan ook weinig mensen. Bovendien maken twitteraars slechts een klein deel van de bevolking uit. Ten slotte is het niet duidelijk of mensen die over een voetbalclub twitteren fan zijn van de club. Engelsen staan erom bekend uiterst kritisch te staan tegenover hun eigen club, maar ook tegenover rivaliserende clubs. Veel voorkomende reacties op fora over deze Twitterdata is dan ook dat deze weinig waarheidsgetrouw zou zijn.

Onze analyse heeft geprofiteerd van onderzoekers die voetbalfandata van Twitter ingedeeld hebben naar district. Wanneer dergelijke informatie zelf wordt bewerkt, kan deze informatie ook ingedeeld worden in andere gebieden, zoals wijken, waardoor deze meer bruikbaar is voor onze onderzoeksvraag. Dit vergt echter een flinke (tijds)investering.

Grilligheid districten. De gekozen districten zijn in enkele gevallen vrij willekeurig gekozen. Het betreft soms een gebied dat ver van een stadscentrum ligt. Voor een deel is dit ingegeven door de indeling van het Britse postsysteem: een stadscentrum hoort niet altijd tot het district. Om het gehele stadscentrum te beslaan moeten meerdere districten samengenomen worden, waardoor buitenwijken en voorsteden vanzelf worden meegenomen. Onze analyses zouden baat hebben bij een methode om kleinere gebieden te identificeren zodat een gebied voor elke stad op een zelfde manier is te bepalen. Een oplossing is om wijken te selecteren, maar hiervan zijn geen Twittergegevens voorhanden.

Behoefteweekcijfersongewenstgedrag. Hoewel de criminaliteitscijfers gedetailleerd zijn, namelijk per maand per district, is het voor de huidige analyse mogelijk niet gedetailleerd genoeg. Omdat dit maandcijfers zijn, analyseren we ook de voetbalresultaten per maand. Het is echter mogelijk dat de maandelijkse resultaten van een voetbalclub geen goede voorspeller zijn van crimineel en antisociaal gedrag. Hoewel de resultaten een effect laten zien van voetbaluitslagen op het eigen (agressief) gedrag is het aannemelijk dat deze effecten vlak na een wedstrijd sterker zijn. Dat zal tot minder variantie van de criminaliteitscijfers leiden wat het verwerpen van het null-model van deze analyse kan verklaren.

Data niet fijnmazig genoeg. Het uiteindelijke doel van onze analyse is in feite niet mogelijk met deze datasets. Om te kunnen analyseren of crimineel gedrag dan wel antisociaal gedrag varieert als gevolg van een verminderde zelfcontrole met de voetbaluitslagen van iemands favoriete club, moet er een dataset zijn die deze gegevens verbindt aan een unieke case. Hier is enkel informatie op macroniveau gebruikt, namelijk het aantal Twitterberichten per district en de criminaliteit per district. Een causaal verband is zo niet vast te stellen.

Conclusie. De analyse laat zien dat big data in potentie veel mogelijkheden bieden voor (sport) sociologisch onderzoek. Gegevens die erg tijdrovend en kostbaar zijn om te verzamelen zoals de criminaliteit in honderden districten zijn nu eenvoudig te verkrijgen uit openbare datasets. Aan de andere kant bieden dergelijke data enkel een macroperspectief. De gegevens zijn niet te herleiden tot een persoon. Er kan dus niet achterhaald worden in welke mate iemand fan is van een bepaalde club, waar deze woont en hoe crimineel gedrag verandert met de resultaten van de club. In een survey is dit wel mogelijk, maar de opzet zou ingewikkeld en daarmee duur uitvallen terwijl onzeker is of survey's wel altijd alle vragen kunnen beantwoorden. Het gebruik van big data is door de besparing op veldwerkkosten aanzienlijk goedkoper, maar deze data geven in onze analyse slechts beperkt antwoord op de vraag. Een vooropgezet onderzoek zoals een survey heeft meer mogelijkheden omdat vooraf precies nagedacht kan worden hoe een vraag is te beantwoorden.

Bijlage III.

Enquêtedata voor sportonderzoek

Hoewel discussie kan ontstaan in welke mate data van enquêtes als big data zijn te beschouwen, willen we aandacht vragen voor informatie over sport uit bestaande (bevolkings)onderzoeken. Met deze vaak grote, landelijke onderzoeken zijn secundaire analyses uit te voeren.

Het onderzoek dat centraal staat is de Gezondheidsenquête van het CBS (2001-2015). Dit is een onderzoek onder duizenden Nederlanders naar zorggebruik en leefstijl. Op wekelijkse basis is informatie over sport en bewegen beschikbaar. Daarnaast gaat het onderzoek in op zaken zoals roken en alcoholconsumptie. Via een speciale, beveiligde analyse zijn koppelingen met andere databestanden mogelijk aangezien het CBS de burgerservicenummers van de respondenten heeft. Te denken valt aan informatie van de Dienst Uitvoering Onderwijs (opleidingsdata scholieren) of autobezit (Rijksdienst voor het wegverkeer).

Voor sport is voorts het onderzoek Vrijetijdsomnibus (VTO) van het Sociaal en Cultureel Planbureau relevant. Om de twee jaar vindt een meting plaats (beschikbaar/gepland 2012, 2014 en 2016). De inhoud heeft betrekking op sport en cultuur. Ook aan deze data valt externe informatie te koppelen (via BSN).

De data van de Leefstijlmonitor - aanvullend van RIVM/CBS - (in 2015 uitgevoerd en voor 2017 gepland) zijn ook voor analyses te benutten. Bij het RIVM is daarnaast data beschikbaar van de Gezondheid Monitor en de Jeugd Gezondheid Monitor. De Gezondheids Monitor is in 2012 uitgevoerd door de GGD's in Nederland en is grotendeels gebaseerd op de vragen van de Gezondheidsenquête. In het najaar van 2016 vindt het veldwerk van de tweede meting plaats. Het voordeel van deze data is dat informatie van bijna 400.000 (volwassen) Nederlanders is verzameld waardoor selecties onder specifieke groepen mogelijk zijn.

Andere relevante onderzoeken hebben soms een incidenteel karakter. Het betreft data van de Eurobarometer die af en toe vragen over sport bevat (www.gesis.org/eurobarometer-data-service). Het voordeel van het gebruik van dit onderzoek is dat een vergelijking met andere Europese landen mogelijk is. De vragen van het Nationaal Sportonderzoek van het Mulier Instituut vallen op door de variatie van de onderwerpen waarover in de loop der jaren data zijn verzameld. Het is een onderzoek onder de bevolking. NOC*NSF voert een vergelijkbaar onderzoek onder de bevolking uit onder de naam SportersMonitor.

De stichting Dans van KNAW stelt veel data van sociaalwetenschappelijk onderzoek en humaniora beschikbaar (www.dans.knwa.nl). Ook sportonderzoekers deponeren data bij deze instelling. Onder andere de data van de Nationale Sportonderzoeken (NSO) van het Mulier Instituut zijn hier ondergebracht. Het onderzoek van het CBS is alleen via een speciale afgeschermd procedure te analyseren (micro data analyse).

De data van internationaal onderzoek zoals de EU Eurobarometer zijn ook beschikbaar voor onderzoekers (www.gesis.org/eurobarometer-data-service/home/).

Literatuur

AAPOR big data taskforce (2015). AAPOR Report on Big Data.

Anderson, Chris (2014). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Science, 23 juni 2008. Verkregen via: www.wired.com/2008/06/pb-theory/

Belt, J. van den & Nillessen, C. (2014). Kritische succesfactoren voor de inzet van Big Data: is groot wel anders? *Maandblad voor Accountancy en Bedrijfseconomie*, pp. 147-156.

Billings, A.C., L.M. Burch, M.H. Zimmerman (2015). Fragments of us, fragments of them: social media, nationality and US perceptions of the 2014 FIFA World Cup. *Soccer & Society*, 2015.

Bongers, F., Jager, C.J., Velde, R. ter (2015). *Big Data in onderwijs en wetenschap. Inventarisatie en essays*. Utrecht: Dialogic innovatie interactie.

Branscombe, N.R., & Wann, D.L. (1991). The positive social and self concept consequences of sport team identification. *Journal of Sport and Social Issues*. 15(2). 115-127.

Castillo, C., De Francisci Morales, G., M. Mendoza (2013). Says who? Automatic Text-Based Content Analysis of Television News. arXiv, augustus 2013. Doi 10.1145/2513549.2513558

Cornelissen, S. & Maennig, W. (2010). On the Political Economy of 'Feel-good' Effects at Sport Mega-events: Experiences from FIFA Germany 2006 and Prospects for South Africa 2010. *Alternation* 17,2 (2010) pp. 96 - 120.

CDV (2014). *Biopolitiek: de macht van Big Data*. Amsterdam: Boom Tijdschriften.

Centraal Bureau van de Statistiek, CBS (2016). Transport en mobiliteit 2016. Centraal Bureau van de Statistiek: Den Haag.

Collard, D.C.M., Boutkan, S., Grimberg, L., Lucassen, J.M.H. & Breedveld, K. (2014). Effecten van sport en bewegen op de basisschool. Voorstudie naar de relatie tussen sport en bewegen op school en schoolprestaties. Utrecht: Mulier Instituut.

Daas, P., Offermans, M., Tennekes, M., Priem, A., P. van den Hurk (2013). Big Data en het CBS. Enkele voorbeelden (presentatie). Centraal Bureau van de Statistiek: Den Haag.

Dallinga, J., Janssen, R., Bie, van der, J., Nibbeling, N., Krose, B., Goudsmit, J., Megens, C., Baart de la Faille-Deutekom, M., Vos, S. (2016). De rol van innovatieve technologie in het stimuleren van sport en bewegen in de steden Amsterdam en Eindhoven. In: *Vrijetijdsstudies*, nummer 2, jaargang 34.

Decisio (verwacht, 2016) <http://decisio.nl/nieuwsblog/2016/bezoekersanalyse-giro-ditalia-2006-aantal-en-herkomst-in-beeld/>

DeltIQ Viewpoint (2013). *Zes standpunten over Big Data*. Breukelen: DeltIQ Viewpoint.

Dimmock, J.A., & Grove, J.R. (2005). Relationship of fan identification to determinants of aggression. *Journal of Applied Sport Psychology*. 17(1). 37-47.

Dool, R. van den en A. Tiessen-Raaphorst (2013). Recreatieve en ontspannende activiteiten. In: Met het oog op de tijd. Red. M. Cloin. Den Haag: Sociaal en Cultureel Planbureau.

Dool, R. van den (2015). Gedragsverandering bij 'moeilijke' groepen. Kansen om sport en bewegen te stimuleren? Utrecht: Mulier Instituut.

Dute D.J., Bemelmans W.J.E., Breda J. (2016). Using Mobile Apps to Promote a Healthy Lifestyle Among Adolescents and Students: A Review of the Theoretical Basis and Lessons Learned. *JMIR Mhealth Uhealth* 2016;4(2):e39. DOI: 10.2196/mhealth.3559

Eagleman, A., L. M. Burch, R. Vooris (2014). A unified version of London 2012, new media coverage of gender, nationality, and sport for olympics customers in six countries. *Journal of Sport Management*, 2014, 28, 457-470. Doi 10.1123/jsm.2013-0151.

Elling, A., I. van Hilvoorde, R. van den Dool (2014). Creating or awakening national pride through sporting success: A longitudinal study on macro effects in the Netherlands. *International Review for the Sociology of Sport* 2014, Vol. 49(2) 129-151. DOI: 10.1177/1012690212455961

Endeavour Partners (2014). Inside Wearables. How the Science of Human Behavior Change Offers the Secret to Long-Term Engagement. <http://endeavourpartners.net>

European Commission (2016). *The EU Data Protection Reform and Big Data*. European Commission.

Gartner (2011). *Gartner Says Worldwide Enterprise IT Spending to Reach \$2.7 Trillion in 2012*. URL: <http://www.gartner.com/newsroom/id/1824919>. Geraadpleegd op 15 juli 2016.

Hansen, H.B., Børtnes I., Hildebrand M., Holme I., Kolle E., Anderssen S.A (2014). Validity of the ActiGraph GT1M during walking and cycling. *Journal of Sports Science*, 32, 6. Doi: 10.1080/02640414.2013.844347.

Hement, D. (2013) 'emoto - Visualising the online response to London 2012' in Cleland, K., Fisher, L. & Harley, R. (Eds.) *Proceedings of the 19th International Symposium of Electronic Art, ISEA2013*, Sydney. <http://ses.library.usyd.edu.au/handle/2123/9475>

Heuvel, M. van den, Hoekman, R. & Poel, H. van der (2011). Sport in de stad. Over de maatschappelijke, ruimtelijke en economische rol van sport in de stedelijke context. Nieuwegein/'s-Hertogenbosch: Arko Sports Media/W.J.H. Mulier Instituut.

Hiemstra, J., Dijkman, M., Minderhoud, M., Meerman, M., Bij, W. van der (2014). *Big data in de publieke sector. Nationaal Big Data Congres voor de overheid. Essenties Nationaal Big Data Congres*. Utrecht 2014.

Hills, A.P., N. Mokhtar, N. M. Byrne (2014). Assessment of physical activity and energy expenditure: an overview of objective measures. *Frontiers in nutrition*, june 2014, volume 1. Doi: 10.3389/fnut.2014.00005

Hoekman, R., van den Dool (2015). Gemeentelijke uitgaven aan sport. Een overzicht van de ontwikkelingen (2010-2015). Utrecht: Mulier Instituut.

Hoekman, R.H., Breedveld, K., Kraaykamp, G. (2016) Sport participation and the social and physical environment: explaining differences between urban and rural areas in the Netherlands. *Leisure Studies*, 2016. <http://dx.doi.org/10.1080/02614367.2016.1182201>

IBM (2011). *IBM Big Data Success Stories*. IBM.

Jin, X., Wah, B.W., Cheng, X., Wang, Y. (2015). Significance and Challenges of Big Data Research, *Big Data Research*, 2, pp. 59-64. Elsevier. Doi.org/10.1016/j.bdr.2015.01.006

Kamer van Koophandel (2014). *(Big) data. Groeikansen voor ondernemers*. Kamer van Koophandel.

Kool, L., Timmer, J., Est, R. van (2015). De datagedreven samenleving. Den Haag: Rathenau Instituut.

Knoppers, A., Elling, A. (1999). Gender, etniciteit en de sportmedia. Een inventarisatie van 'reguliere' en olympische sportberichtgeving. Utrecht: Universiteit Utrecht / Tilburg: Katholieke Universiteit Brabant.

Kruik, E. de, Schoenmakers, F., Dijkman, M., Gremmen, M., Minderhoud, M. (2015). *Rol en positie van gemeenten in een samenleving van Big en Open Data*. Den Haag: Kwaliteitsinstituut Nederlandse Gemeenten.

Lin, W. H., A. Hauptmann (2002). News video classification using SVM-based multimodal classifiers and combination strategies. In: *Procedures of Association of Computer Machinery Multi Media (ACM MM)*, 323-326.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Marketingfacts (2016).

www.marketingfacts.nl/images/specials/Customer_Journey.pdf?utm_source=website&utm_medium=banner&utm_campaign=CJ (maart 2016).

McCarthy, D., R.L. Jones, P. Potrac (2003). Constructing images and interpreting realities. The Case of the Black Soccer Player on Television. *International review for the sociology of sport* 38/2(2003) 217-238.

Millington, B. & Millington, R. (in voorbereiding). 'The Datafication of Everything': Towards a Sociology of Sport and Big Data, *Sociology of Sport Journal*.

Multiscope (2016). <http://www.multiscope.nl/diensten/marktcijfers/smart-health-monitor.html>

Nafus, D., Sherman, J. (2014). The Quantified Self Movement as an Alternative Big Data Practice. *International Journal of Communication* 8 (2014), 1784-1794.

Neville, J., A. Wixted, D. Rowlands, D. James (2010). Accelerometers, An Underutilized Resource in Sports Monitoring.

NIPS 2015. Samenavtting aankondiging. <https://nips2015.sched.org/event/4GPi/machine-learning-for-spoken-language-understanding-and-interactions>

NWO (2016). Nationale Kennisagenda Sport en Bewegen. Van traplopen tot podium. Den Haag: NWO.

Poel, H. van der, Wezenberg-Hoenderkamp, K., Hoekman, R., Bakker, S., Davids, A., Hoffmans, W., Lindert, C. van, Lucassen, J.M.H., Rewijk, D., Schadenberg, B., Scholten, V. (2016). Sportaccommodaties in Nederland : kaarten en kengetallen. Utrecht / Nieuwegein: Mulier Instituut / Arko Sports Media.

Popper, K. (1959). *The Logic of Scientific Discovery*. Basic Books, New York.

Ranck, J. (2012). *The wearable computing market: a global analysis*. Gigoam.

Romijn, D., J. van Kalmthout, J. Broerse (2016). Resultaten pilot SportklimaatApp. Mulier Instituut: Utrecht.

Shor, E., A. van de Rijt, C. Ward, S. Askar, S. Skiena (2014). Is there a political bias? A computational analysis of female subjects coverage in liberal and conservative newspapers. *Social Science Quarterly*, 95, 5, 2014. DOI: 10.1111/ssqu.12091

Skysports (2016). <http://www.skysports.com/f1/news/24227/10328011/f1-driver-eye-tracking-nico-hulkenberg-tests-out-reactions>

Stuij, M., Wisse, E., Mossel, G. van, Lucassen, J.M.H., Dool, R. van den (2011). School, bewegen en sport. Onderzoek naar relaties tussen de school(omgeving) en het beweeg- en sportgedrag van leerlingen. Nieuwegein/'s-Hertogenbosch: Arko Sports Media/W.J.H. Mulier Instituut.

The Atlantic (2015). The Technology That Created a New Generation of Runners. How fitness-tracking apps have made the sport more accessible, *The Atlantic*, 22 november 2015, URL: <http://www.theatlantic.com/technology/archive/2015/11/the-technology-that-created-a-new-generation-of-runners/417126/>, geraadpleegd op 27 november 2015.

Timmis, M. A., Turner, K., van Paridon, N. (2014). Visual Search Strategies of Soccer Players Executing a Power vs. Placement Penalty Kick. *PLoS ONE* 9(12): e115179. doi 10.1371/journal.pone.0115179

Troiano, R.P., J.J. McClain, R.J. Brychta, K.Y. Chen (2014). Evolution of accelerometer methods for physical activity research. *British Journal of Sports Medicine* 2014;48:1019-1023. Doi:10.1136/bjsports-2014-093546

Tukey, J.W. (1977). *Exploratory Data Analysis*. Boston: Addison-Wesley. ISBN 0-201-07616-0.

United Nations (2014). A world that counts. Mobilising the data revolution for sustainable development. United Nations.

Verhaar, P., Schäfer, M.T. (2016). Onderzoek naar tweets met #gelukzoeker (Utrecht Data School). E-data & Research, juni 2016.

Versichele, M., T. Neutens, S. Goudeseune, F. van Bossche, N. van de Weghe (2012). Mobile Mapping of Sporting Event Spectators Using Bluetooth Sensors: Tour of Flanders 2011. *Sensors* 2012, 12, 14196-14213; DOI:10.3390/s121014196

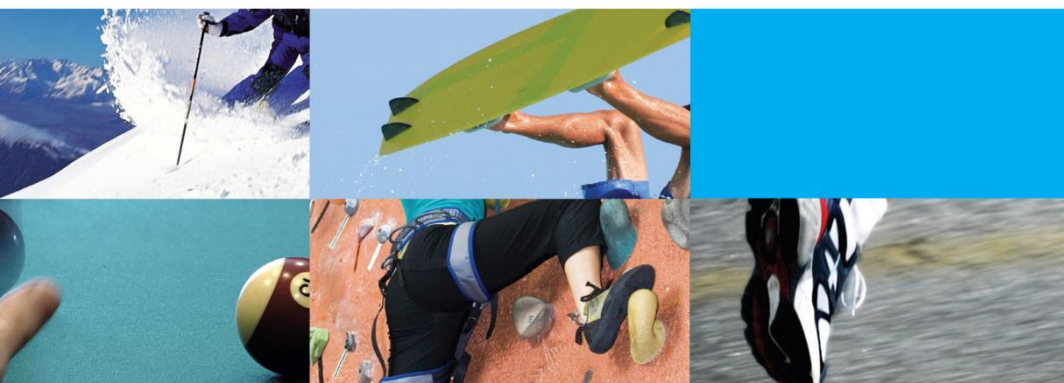
Vos, S.B. (2016). Designerly solutions for vital people. Inaugurele rede TU Eindhoven 7 oktober 2016.

VSO (2014). Big data. De hype voorbij. Nijmegen: VSO.

Wicker, P., K. Hallmann, C. Breuer (2013). Analyzing the impact of sport infrastructure on sport participation using geo-coded data: evidence from multi-level models. *Sport Management Review* 16 (2013) 54-67. doi.org/10.1016/j.smr.2012.05.001

WRR (2016). Big Data in een vrije en veilige samenleving. Den Haag: WRR.

Yang Yu, Xiao Wang (2015). World cup 2014 in twitter world: A big data analysis of sentiments in U.S. sports fans' tweets. *Computers in Human Behavior*, 48, 2015, 392-400. DOI:10.1016/j.chb.2015.01.075



Herculesplein 269 | 3584 AA Utrecht | Postbus 85445 | 3508 AK Utrecht
T +31 (0)30 721 02 20 | info@mulierinstituut.nl | www.mulierinstituut.nl