Big data en de onderzoeker: een gesprek met Michel van Eeten

van der Voort, Haiko; de Bruijn, Hans

DOI
10.5553/Bk/092733872016025001007

Publication date
2016

Document Version
Final published version

Published in
Bestuurskunde

Big data en de onderzoeker: een gesprek met Michel van Eeten*

Haiko van der Voort & Hans de Bruijn

Om meer inzicht te krijgen in de praktijk van big data hebben we een aantal mensen gesproken voor wie big data niet alleen theorie, maar ook dagelijkse kost is. We spraken bestuurskundig onderzoeker *Michel van Eeten*, die big data heeft omarmd voor zijn onderzoek naar aspecten van internetsecurity. Hij doet onderzoek naar patronen in de wereldwijde netwerken van gehackte websites en besmette computers – de zogenaamde *botnets*. Tevens onderzoeken hij en zijn team de mogelijkheden en onmogelijkheden van verschillende partijen – zoals Internet service providers – om aanvallen tegen te houden. Aan hem vroegen we wat big data gedreven onderzoek precies inhoudt. Vervolgens vroegen we hem om consequenties van big data voor onderzoekers en voor de maatschappij.

Een omkering van vraag en aanbod

Er is een overproductie aan data. *Machine learning* mag misschien een hype worden genoemd, maar er zijn wel degelijk overal apparaten die als hoofd- of bijproduct data produceren. De kosten om data te produceren worden steeds lager. Zo ontstaat er een enorm aanbod aan data en proberen onderzoekers er bijpassende vragen bij te stellen.

Dat is fundamenteel anders dan bij niet-big data gedreven onderzoek. Instanties als het Centraal Bureau voor de Statistiek hebben ook veel data, maar die is kostbaar om te verzamelen, dus gaat de vraag vooraf aan het aanbod, met als motto: 'Welke informatie gaan we genereren om de door ons van tevoren vastgestelde vragen te beantwoorden?' Big data gedreven onderzoek heeft een ander motto, namelijk: 'Ik heb een enorme hoeveelheid data en wat ga ik ermee doen?' Daarmee betekent het een omkering van vraag en aanbod van data.

Een multi-disciplinair onderzoeksteam

Het aanbod aan relevante data is overvloedig, maar nog steeds beperkt. De beperking wordt echter meer dan voorheen bepaald door eigenaars van data. Overheden streven vaak open data na, maar een meerderheid van de data is gegenereerd door partijen die er een doel mee hebben, zoals het commercialiseren ervan. Dit bepaalt de volgorde van onderzoeksactiviteiten.

* Dr. H.G. van der Voort is universitair docent aan de Technische Universiteit Delft. Prof. dr. J.A. de Bruijn is hoogleraar aan de Technische Universiteit Delft.

De eerste activiteit is namelijk het verwerven en toegang verkrijgen tot datasets. Deze datasets zijn weliswaar enorm, maar toegang tot alle datasets is niet mogelijk. Een tweede activiteit is het doorgronden welke vragen met de datasets beantwoord kunnen worden. Een derde activiteit is het herkennen en matigen van discrepanties. Bijvoorbeeld: als de patronen aangeven dat in Italië veel minder cyberaanvallen plaatsvinden dan in alle andere Europese landen, dan kan dit betekenen dat er in Italië minder cyberaanvallen plaatsvinden. Het is ook mogelijk dat het verschil is te verklaren door de techniek. Wellicht zit er een geografische *bias* in de incidenten die door de technieken geobserveerd kunnen worden. Het herkennen en erkennen van discrepanties is daarmee essentieel voor de kwaliteit van de onderzoeksresultaten.

Dan volgt een zoektocht door talloze *plots*, waarbij alles tegen elkaar wordt afgezet, om te begrijpen welke patronen erin schuilgaan. Het genereren van die plots en doorsnedes gebeurt automatisch, omdat de hoeveelheid data te groot is om dit 'met de hand' te doen. Het onderzoeksteam bevat daarom niet alleen sociaalwetenschappers, maar ook data-analisten met een ICT-achtergrond. Het onderzoeksteam van big data gedreven onderzoek is multi-disciplinair.

Monnikenwerk

Big data gedreven onderzoek is geen vrijblijvend onderzoek. Allereerst zijn de eindeloze analyses van de data en de gevonden plots te kwalificeren als monnikenwerk. Daarbovenop geldt dat de data en de data-analisten niet altijd antwoord kunnen geven op de belangrijkste vragen. Het onderzoek van Michel van Eeten is sociaalwetenschappelijk en richt zich uiteindelijk op de actoren: zij die de cyberaanvallen uitvoeren en zij die ze kunnen tegenhouden. Om goede uitspraken te kunnen doen is een koppeling van IP-adressen (of andere technische *identifiers*) en concrete organisaties nodig. Een dergelijke connectie tussen een technisch artefact en een actor bestond nog niet. Een van de universitair docenten heeft een tool ontwikkeld om een dergelijke koppeling te maken. Dit betekende een halfjaar fulltime werk.

Witte ruis

Een overgroot gedeelte van de data wordt verzameld door partijen die er een doel mee hebben. Dit heeft aldus effect op de beschikbaarheid van data voor onderzoek, maar heeft ook bredere consequenties. Er zijn heel veel private bedrijven die data genereren en er producten van maken om te verkopen. In de Verenigde Staten hebben bepaalde winkelketens een contract met online advertentienetwerken als die van Google. In de fysieke winkels klinkt witte ruis uit de *speakers*. Die is niet voor voorbijgangers te horen, maar kan wel worden opgepikt door hun smartphones, die dat doorgeven aan de advertentienetwerken. Op deze manier kan een koppeling worden gelegd tussen een online advertentie en een bezoek aan een fysieke winkel. Een ander voorbeeld is dat bedrijven profielen maken van

Bestuurskunde 2016 (25) 1
doi: 10.5553/Bk/092733872016025001007

individuen die in *real time* worden gekoppeld aan een locatie waar een melding is gedaan. Die informatie wordt verstuurd naar de politieagenten die op weg zijn naar de melding.

De mogelijkheden zijn eindeloos, de drijfveer om met telkens nieuwe dataproducten te komen is groot. De belangen van het individu sneeuwen hier echter onder. Wie representeert hen in de *arm's race* om nieuwe, betere dataproducten? Er zijn nauwelijks *checks and balances* tussen de internationale en vaak onzichtbare producenten van data en het individu. De Europese Unie en ook andere supranationale overheden hebben geïnvesteerd in regels voor *informed consent*: data kunnen worden gebruikt als het individu er expliciet toestemming voor heeft gegeven. Maar dit model is kapot. De meeste mensen weten niet waarvoor ze toestemming geven. Daarvoor is big data te complex en zijn de bedrijven die data commercialiseren te ongrijpbaar. Het is van groot belang om een alternatief voor *informed consent* te ontwikkelen. Dit zal echter niet voorkomen dat het individu soevereiniteit kwijtraakt.