In: Karakter, Tijdschrift voor de Wetenschap. Nummer 50, 2015, 22-24

Big Data, de revolutie voorbij de hype

Bettina Berendt, Departement Computerwetenschappen, KU Leuven
Mathias Verbeke, Sirris
Email: voornaam.achternaam@cs.kuleuven.be

"Big Data" – dat amalgaam van Google, Facebook, de Large Hadron Collider en de NSA – "zal onze manier van leven, werken en denken veranderen" – zo wordt ons verteld. Maar wat is "Big Data" (BD) nu eigenlijk? Twee recente boeken – "Big Data" door Mayer-Schönberger en Cukiers en"The Data Revolution" door Kitchin – proberen ons hierop een antwoord te verschaffen. In deze tekst zullen we deze twee boeken kort beschrijven en vervolgens hun visie op BD vergelijken. Hiertoe zullen we een voorbeeld uit het eerste boek deconstrueren met de hulp van ideeën uit het tweede. (Dit laatste is uitgewerkt op de webpagina beneden.) We besluiten deze tekst met een aanbeveling voor uw leeslijst en een oproep tot actie. Als computerwetenschappers in de artificiële intelligentie en data mining met een interdisciplinaire visie, is deze recensie geschreven vanuit twee perspectieven: wetenschap en onderwijs enerzijds, en de bedrijfspraktijk anderzijds.

Alle citaten zijn eigen vertalingen vanuit het Engels; de pagina's verwijzen naar de hieronder vermelde uitgaven.

- Mayer-Schönberger, V. & Cukier, K. (2013). Big Data. A Revolution That Will Transform How We Live, Work and Think. London: John Murray (Publishers).
- Kitchin, R. (2014). The Data Revolution. Big Data, Open Data, Data Infrastructures & Their Consequences. London: Sage.

De originele citaten, een gedetailleerde deconstructie van Mayer-Schönberger, V. & Cukier's tekstpassage over *predictive policing*, en een verzameling van relevante referenties vindt de lezer in de uitgebreide, Engelstalige versie van deze recensie op http://people.cs.kuleuven.be/~bettina.berendt/Reviews/BigData.pdf .

1. Wat is Big Data uberhaupt?

BD zijn, eerst en vooral, data, of letterlijk vertaald: gegevens. Maar zijn data echt *gegeven* door de fenomenen die gemeten worden? In de praktijk verwijst data echter meestal naar "de elementen die *genomen* worden: geëxtraheerd door observatie, berekeningen, experimenten, en het behouden van records", "geselecteerd door de wetenschapper naargelang zijn [of haar] doel" (Kitchin, p.2). Dus in feite zou "capta" hier meer op zijn plaats zijn dan "data".

Dit betekent ook dat data nooit "voor zich zelf spreken", dat zij ons *geen* directe toegang geven tot de echte natuur der dingen, dat het gebruik van data helemaal *niet* het tegendeel is van de subjectiviteit van kwalitatieve onderzoeksmethodes die mensen altijd in de zingeving betrekken – hoewel dit maar al te vaak beweerd wordt. Kitchin waarschuwt ons voor de naïeve aanname dat data objectief zijn, die aan de basis ligt van het huidige enthousiasme rond (big) data, en de oorsprong is van vele misvattingen hierrond. In plaats daarvan zouden we ons kritisch moeten afvragen waar gegevens vandaan komen, met welke bedoelingen zij verzameld en verwerkt zijn, en met welke methodes dit gebeurd is. Uiteraard zijn ook deze vragen en alle interpretaties van data en analyseresultaten een product van een mens die over de data spreekt en over de inferenties eruit, en dit op basis van een achterliggende agenda – ook hier spreken de data niet voor zichzelf.

"Big Data" erven deze algemene eigenschappen van data. Daarenboven worden vaak de 5 Vs aangehaald om Big Data te definiëren:

- *volume*: vandaag, terabytes of petabytes van data;
- *velociteit* (snelheid): data worden gecreeërd in (bijna-)real time;
- *variëteit*: data zijn gestructureerd en ongestructureerd, en vaak tijdelijk en ruimtelijk geïndexeerd;
- *juistheid*: de kwaliteit van de data, die ook de kwaliteit van de analyse ervan beïnvloed;
- *waarde*: de toegevoegde waarde die voorkomt uit (een geschikte analyse van de) data.

Kitchin bevestigd de eerste 3 eigenschappen. Bovendien voegt hij de volgende eigenschappen toe (p. 68):

- *Exhaustief*, met als bedoeling om hele populaties of systemen vast te leggen, of tenminste veel grotere steekproeven te gebruiken dan in traditionele "small data" studies;
- Van hoge *resolutie*, met als bedoeling zo gedetailleerd mogelijk te zijn, en indexicalisch zodat de beschreven objecten, mensen, dieren, … uniek geïdentificeerd kunnen worden;
- *Flexibel* en *schaalbaar*, zodat eenvoudig meer detail (attributen) en/of meer beschreven objecten (records) toegevoegd kunnen worden.

Dit is gelijkaardig aan de beschrijvingen die ook in andere werken te vinden zijn. Merk echter op dat dit slechts één definitie is; BD is en blijft een buzzword (zie Dutcher, 2014).

Naast het beschrijven van de data an sich, wordt de term "Big Data" wordt ook gebruikt om te verwijzen de analyses erop en de toepassingen die voortkomen uit de analyseresultaten. Deze analyses bepalen op hun beurt de eigenschappen van de data en de mogelijke toepassingen.

2. "Big Data" (Mayer-Schönberger en Cukier)

Hoewl het boek van Mayer-Schönberger en Cukier (voortaan MS/C) een populariserende tekst is, wordt het ook gebruikt door wetenschappers, vooral buiten computerwetenschappen, als inleiding tot het thema. Zoals veel andere populariserende teksten werkt MS/C met de ruimere betekenis van BD: veel gegevens, analyses erop, en de toepassing van de resultaten. De "capta"-eigenschappen van data worden grotendeels genegeerd; de (impliciete) aanname dat data neutraal en objectief zijn loopt doorheen de tekst.

Een boek recenseren dat op amazon.com alleen al meer dan 300 recensies kreeg, dat in toonaangevende kranten is besproken, en in de bestseller lists van de New York Times en de Wall Street Journal heeft gestaan is niet eenvoudig. We zullen daarom beginnen met een samenvatting van argumenten uit bestaande reviews die aansluiten bij onze mening over het boek (bronnen en hyperlinks: zie webpagina-versie van deze recensie):

- Het boek presenteert een breed gamma aan voorbeelden van "positieve" (commercieel succesvolle en sociaal gewenste) toepassingen van BD, en doet dit op een eenvoudig te begrijpen en onderhoudende manier.
- Het boek beschrijft ook een aantal belangrijke "risico's" van BD die tegenwoordig bediscussieerd worden, waaronder de implicaties voor privacy.
- Zowel de "positieve" also de "negatieve" voorbeelden kunnen goede startpunten vormen voor discussies rond BD, o.a. in het onderwijs.

Analytisch is het boek echter niet echt sterk. Het stelt voor om BD door drie elementen te karakteriseren:

- BD hoeft niet echt "groot" te zijn, zolang het maar geen steekproef is van de data, maar "alle" data bevat (bv. alle werknemers van een bepaald bedrijf, in plaats van een steekproef ervan).
- BD is rommelig (*messy*): onvolledig, slecht gerepresenteerd, …
- Werken met BD betekent op zoek gaan naar correlaties, eerder dan naar verklaringen (causaliteit).

Geen van deze drie eigenschappen is echt nieuw: zo bestaan er al dataverzamelingen die op bepaalde wijze "volledig" zijn, en in alle empirische wetenschappen wordt meestal ook met "rommelige gegevens" en correlatieve analyses gewerkt.

Bovendien zijn deze aannames te sterk om algemeen geldig te zijn. Zoals Kitchin in zijn boek aantoont is de aanname van volledigheid een te grote vereenvoudiging. Zelfs een dataset van alle zoekopdrachten in Google of alle posts van Facebookgebruikers is slechts een representatie én een steekproef. Een representatie omdat het een specifieke manifestatie is van wat mensen willen weten of wat zij denken, en een steekproef omdat het enkel de mensen bevat die dit platform gebruiken, de dienst slechts een deel van de data publiek maakt,of omdat het is opgelegd door het API-beleid. (Zo geeft Twitter via zijn openbare API maar toegang tot "ca. 1%" van de tweets, zonder dat we weten hoe deze 1 procent gekozen wordt.)

In verband met het tweede punt (BD is rommelig) wordt vaak, en ook door MS/C, beweerd dat "meer beter overtroeft": Data mogen dan wel rommelig zijn, als er maar genoeg data is, wordt de slechte kwaliteit teniet gedaan. Maar hier worden statistische modellen en vragen van datakwaliteit door elkaar gegooid. Een van de oudste spreekwoorden in computerwetenschappen stelt echter dat "garbage in, garbage out" – onafhankelijk van de hoeveelheid rommelige data.

Verradelijker is de bewering dat "Correlatie causaliteit kan vervangen". Het staat vast dat correlatie vaak nuttig en causaliteit niet altijd nodig is – *dit* is niets nieuws. MS/C volgt echter – op een relatief onkritische manier – de hype dat correlatie *alleen* voldoende is, dat correlatie het causale denken overbodig maakt. Ze benadrukken dit met minachtende formuleringen zoals "de maatschappij zal iets van zijn bezetenheid voor causaliteit laten vallen" (p.7). Dit is polemisch, en de aanname dat correlatie causaliteit kan vervangen vanuit een wetenschappelijk perspectief gewoon onzin (zoals Kitchin in bijzonder groot detail uitlegt in zijn boek).

Het boek is ook vaag als het gaat om de vraag gaat hoe analysten inzicht kunnen verwerven uit BD, en nieuwe analytische technieken zoals data mining methodes worden niet beschreven. Hetzelfde is, helaas, het geval in MS/C's omgang met ethische vragen: deze worden vermeld, maar de analyse is noch diep noch kritisch. In plaats ervan wordt een marktgeoriënteerde aanpak op basis van zelfregulatie naar voor geschoven, zonder een kritische uiteenzetting van zijn voor- en nadelen.

3. "The Data Revolution" (Kitchin)

Hoewel dit boek duidelijk academischer dan MS/C, is het nog steeds zeer toegankelijk. Bovendien is het een goed geïnformeerd en weldoordacht boek dat veelomvattend in inhoud is. Het is gestructureerd in elf hoofdstukken waarvan de titels op zich de encyclopedische breedte aantonen. Kitchin illustreert de concepten van de hoofdstukken met de hulp van

voorbeelden (minder dan MS/C dit doet). De gestructureerde aanpak die hierbij gebruikt wordt is echter beter geschikt voor een diep begrip van BD dan de trefwoorden gekozen door MS/C (bv., “Now”, “More”, “Messy”, “Correlation” voor hoofdstukken 1-4). De hoofdstukken zijn (wij vertalen deze niet omdat de meeste concepten in het Nederlands met hun Engelstalige woorden benoemd worden):

1. Conceptualising Data
2. Small Data, Data Infrastructures and Data Brokers
3. Open and Linked Data
4. Big Data
5. Enablers and Sources of Big Data
6. Data Analytics
7. The Governmental and Business Rationale for Big Data
8. The Reframing of Science, Social Science and Humanities Research
9. Technical and Organisational Issues
10. Ethical, Political, Social and Legal Concerns
11. Making Sense of the Data Revolution .

Dit boek is recenter dan MS/C en heeft, o.a. om deze reden, tot op heden minder recensies ontvangen, maar deze zijn heel positief. Zoals voor MS/C zullen wij eerst de punten uit bestaande reviews waarmee wij akkoord gaan samenvatten: een gebalanceerde presentatie en het vermijden van - elders in dit thema typische - overdrijvingen in taal en inhoud. Verder benadrukt een van deze recensies dat Kitchin naast de motivering (*rationale*), ook de implicaties voor governance, management, en zelfs ons begrip van wetenschap en kennis onderzoekt. Deze implicaties motiveren de “nood aan een kritischer en filosofischer engagement”. Kitchin geeft een gedetailleerd en genuanceerd overzicht van de literatuur om zo’n engagement af te bakenen, waarvan we de cruciale argumenten hier graag even samenvatten.

Argumenten 1. en 2. zijn hierboven reeds beschreven: “data” zijn in feite niet gegeven maar genomen en al daardoor niet neutraal. De volledigheid die BD *nastreeft* is over het algemeen onmogelijk, aangezien ook BD slechts een representatie en een steekproef zijn.

Argument 3. is dat ook de technologie die BD onderbouwt niet neutraal is: gegevensbanken en infrastructuur (data infrastructures) zijn niet zomaar neutrale technische methodes voor het verzamelen en delen van data, maar veeleer een verzameling van processen die “*contingent*” (mogelijk maar niet noodzakelijk, en van de omstandigheden afhankelijk) en relationeel (verbonden met andere processen) zijn. Gegevensbanken en infrastructuur zijn complexe sociotechnische systemen, d.i. systemen van mensen en machines die veelvoudig met elkaar interageren, en zij maken deel uit van een groter institutioneel landschap van onderzoekers, instellingen en kapitaal (Kitchin, p.21).

Argument 4. is dat deze BD-onderbouwende technologie decontextualisatie aanmoedigt, die misleidend en gevaarlijk kan zijn: gegevensbanken ontkoppelen data-analyse van de data doordat ze complexe aanvragen en berekeningen mogelijk maken zonder dat degenen die deze analyses doorvoeren de data moeten bewerken of zelf begrijpen hoe de data samengesteld en georganiseerd werden (Kitchin, p.22).

BD is het kernthema van hoofdstukken 4-11. Hoofdstuk 4 focust op definities (zie boven). Hoofdstuk 5 geeft een grondig overzicht van het gamma van apparatuur en gedragswijzen die samen de niet-aflatende stroom aan data creëren. Ook hoofdstukken 6-11 bieden omvattende overzichten van hun respectievelijke onderwerpen. Bijzonder belangrijk is de passage over de

misvattingen van het empiricisme in hoofdstuk 8. Dit is aanbevolen leesstof voor iedereen die gelooft dat "correlatie causaliteit kan vervangen". Kitchin toont ons hoe data-analyse altijd doordrongen is met voorkennis, aannamen van causatie, en de erop gebaseerde interpretaties. Dit is niet alleen een theoretisch en wetenschappelijk probleem, maar een heel concrete bedreiging van fundamentele rechten (zie Solove, 2011).

Aan het einde van zijn boek schetst Kitchin een alternatieve visie van "een datagedreven wetenschap dat aspecten van abductie, inductie en deductie verenigt" (p. 133). Als een bijdrage tot deze visie willen wij tenslotte nog toevoegen dat zo'n datagedreven wetenschap ook *concreet* moet zijn. Dit is volgens ons het enige echte zwakke punt van dit verder excellente boek: soms is de concept-per-concept presentatie te abstract en mist het de kans om de technieken die het boek voorstelt in dienst van de deconstructie van de BD-hype te stellen.

Een eerste voorbeeld betreft Kitchins beschrijving van de problemen die bij het ontkoppelen van gegevens en analyse onstaan. Dit argument blijft vrij abstract. Een concreet voorbeeld wordt gegeven door boyd en Crawford (2012): er is op dit moment een grote interesse in de sociale netwerken van mensen. De grafen van "friends" of "followers" die gebruikers op sociale netwerksites vormen worden vaak beschouwd en geanalyseerd als dé sociale netwerken (in een sociologische zin). Maar zowel intuïtie als onderzoek tonen duidelijk dat deze niet hetzelfde zijn.

Een tweede voorbeeld is een observatie die wel correct is, maar zonder de bijkomende informatie die kan aantonen dat het beschreven doel ook zonder BD – en dan beter – kan bereikt worden. Kitchin beschrijft de surveillatie van "vroeger informeel opgevolgde" processen zoals het ophalen van huisvuil. Met de hulp van RFID chips die aan de vuilnisbakken vastgemaakt zijn wordt het mogelijk volume te meten en huishoudens naargelang hun effectief volume van vuilnis te laten betalen (p. 90). Dit doel (betalen op basis van volume) is misschien wel sociaal gewenst, maar het kan ook anders. In plaats van enkel de eerste vage *privacy by design* principes van Cavoukian te citeren (pp. 173ff.), kan dit voorbeeld ook gebruikt worden om deze principes in de praktijk te illustreren. Het Belgische systeem voor de ophaling van huisvuil is hiervan een goed voorbeeld: een huishouden mag zoveel volume laten afhalen als zij officiële vuilniszakken (van 30l, 60l, ...) gekocht heeft. De zakken worden in de winkel gekocht en zijn daardoor (aangenomen dat contant betaald werd) ook niet aan een huishouden te linken, en daardoor een privacy-respecterende manier om te betalen voor het ophalen van huisvuil in functie van het volume.

Een derde voorbeeld is dat Kitchin het verstrekken van data via sociale media of *quantified-self* toestellen (bv. armbanden die je snelheid tijdens het joggen of de duur van je slaap meten) als puur vrijwillig classificeert (pp. 93ff.). Kitchin benadrukt in hoofdstuk 7 dat "de gebruikte discoursen hun boodschap als gezond verstand laten uitschijnen om mensen en instellingen van hun logica te overtuigen, om hen op deze manier volgens deze logica te laten gedragen" (p. 113). Als men deze discoursanalyse toepast dan komen er toch twijfels of deze *data donations* zo vrijwillig zijn. Zo controleren werkgevers al nu regelmatig de LinkedIn profielen van potentiele werknemers om zich te overtuigen van de kwalificaties van sollicitanten, en ziekteverzerkingen belonen het gebruik van *quantified-self* toestellen.

We denken dat het koppelen van conrete voorbeelden van BD met concrete voorbeelden van BD-kritiek noodzakelijk is, niet alleen om meer kritische perspectieven op BD aan te moedigen maar ook om het kritische programma zelf te veranderen zodat het actiegerichte aanbevelingen produceert – want abstracte begrippen zoals "discours" zijn moeilijk te vatten voor veel lezers. Een uitgebreid voorbeeld van zo'n koppeling vindt de lezer in de bij deze

tekst horende webpagina (zie sectie 1). Het beschreven voorbeeld analyseert een passage in MS/C over *predictive policing* (het gebruik van BD om voor te spellen waar misdaden waarschijnlijk zijn, of door wie misdaden waarschijnlijk zijn, en politiereacties erop, zoals de concentratie van politieagenten in een bepaalde streek op een bepaalde tijd).

We hopen u met deze tekst een aanzet gegeven te hebben om *The Data Revolution* toe te voegen aan uw leeslijst of die van uw studenten – zij het in computerwetenschappen, rechten, of andere domeinen, in het bijzonder de humane, sociale en gedragswetenschappen. We kijken uit naar het horen van uw reacties op deze tekst en het boek, en uw ervaringen in het (aan)leren van Big Data!

Referenties:

boyd, d. & Crawford, K. (2012). Critical questions for Big Data. Information, Communication & Society, 15:5, 662-679,
Dutcher, J. (2014). What is Big Data? datascience@berkeley Blog. http://datascience.berkeley.edu/what-is-big-data/
Solove, D. (2011). Nothing to hide. The false trade-off between privacy and security. Yale University Press.