BIG DATA

Uitgeverij Academia Press
Ampla House
Coupure Rechts 88
9000 Gent
België

www.academiapress.be

Uitgeverij Academia Press maakt deel uit van Lannoo Uitgeverij,
de boeken- en multimediadivisie van Uitgeverij Lannoo nv.

ISBN 978 94 014 8022 2 – D/2021/45/405 – NUR 740

Johan Decorte
Big data. Een revolutie ontrafeld
Gent, Academia Press, 2021, 96 p.

Vormgeving cover: Studio Lannoo
Vormgeving en zetwerk binnenwerk: Studio Lannoo

VIZIER

Johan Decorte

BIG DATA

Een revolutie ontrafeld

ACADEMIA PRESS

Voor Nore, Rhune en Lars in de hoop dat zij kunnen opgroeien in een wereld van open en faire data.

INHOUD

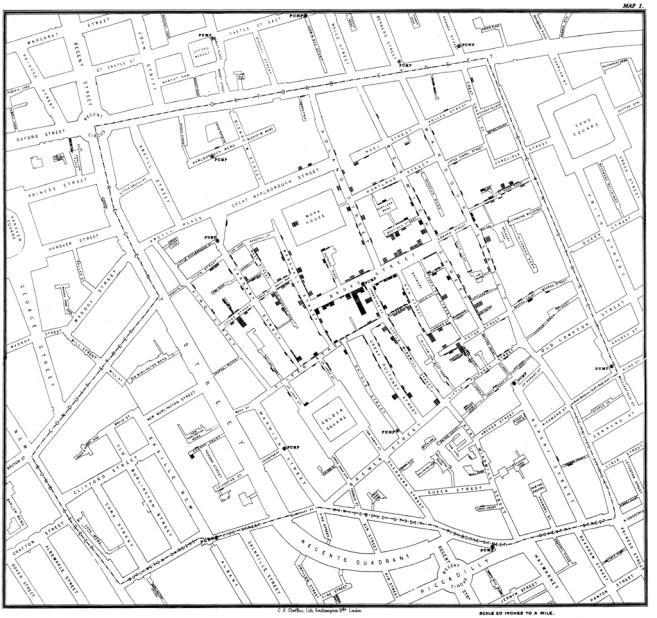
1 WAT ZIJN BIG DATA?

Een dokter in Londen

In 1854 viel de Londense wijk Soho ten prooi aan een cholera-epidemie. Er was toen nog maar weinig bekend over infectieziekten en er deden diverse theorieën de ronde over de manier waarop iemand besmet kon raken. Dokter John Snow, die in een van de armste stadsdelen werkte, vermoedde dat iemand cholera kreeg door contact met water dat besmet was met afval van andere cholerapatiënten. Samen met een dominee bracht hij het aantal choleragevallen in dit stadsdeel – letterlijk – in kaart, zoals te zien is op **figuur 1**.

Hij stelde vast dat het zwaartepunt van de besmettingen te situeren was in de buurt van een waterpomp, waar de buurtbewoners, bij gebrek aan stromend water, zich kwamen bevoorraden. Hij concludeerde daaruit dat het drinken van het besmette water de oorzaak was van de uitbraak, maar hij kon de besmetting niet eenduidig aantonen omdat de wetenschap daarvoor nog in de kinderschoenen stond. Weliswaar daalde na het uitschakelen van de pomp het aantal sterfgevallen in dat stadsdeel drastisch, maar volgens dokter Snow kwam dat wellicht ook doordat een groot aantal bewoners de wijk inmiddels ontvlucht waren.

Dokter John Snow, die beschouwd wordt als een van de grondleggers van de epidemiologie, was een van de eersten die op zoek gingen naar schijnbaar verborgen patronen in data om onderliggende fenomenen te detecteren. Veel landen gebruiken vandaag,



Figuur 1 Op het stratenplan dat dokter Snow gebruikte, is de lengte van elk zwart balkje een maat voor het aantal cholerагevallen op dat adres.

(© Wikimedia Commons, licentie: https://commons.wikimedia.org/wiki/File:Snow-cholera-map-1.jpg).

met uitgebreide digitale middelen en met wisselend succes, dezelfde technieken om lokale besmettingshaarden van de coronapandemie in kaart te brengen en in te dijken.

Dit is slechts een van de vele voorbeelden die in dit boek aan bod komen en die illustreren hoe data een schat aan verborgen informatie kunnen bevatten. Het komt erop aan om, via de juiste technieken, deze informatie te ontsluiten en te interpreteren. Dit voorbeeld illustreert ook dat je via data wel een correlatie tussen

twee grootheden kunt aantonen – er is een duidelijk verband tussen het aantal sterfgevallen op een adres en de nabijheid van de pomp – maar geen oorzakelijk verband of causaliteit. De ingezamelde data toonden niet onomstotelijk aan dat het water van de pomp besmet was en al evenmin dat de bewoners door het drinken ervan overleden waren, maar er was wel een zeer sterk vermoeden.

Twee studenten aan Stanford

> ***'It is our mission to organize the world's information and make it universally accessible and useful.'***
>
> — **Mission statement van Google**

We maken nu een grote sprong in de tijd naar 1996. Sergey Brin en Larry Page, twee doctoraatsstudenten in de computerwetenschappen aan de universiteit van Stanford, werkten aan een project om op het internet te speuren naar webpagina's die aansluiten bij de zoekopdracht van een gebruiker. Zij legden daarmee de basis voor de succesvolle zoekmachine Google, waarvan Page en Brin tot op vandaag de grootste aandeelhouders zijn. Het lukte hen niet alleen om in een razendsnel tempo webpagina's te vinden waarin de zoekwoorden van de gebruiker prominent aanwezig waren, maar dankzij een ingenieus systeem konden ze deze pagina's bovendien rangschikken zodat de 'belangrijkste' bovenaan staan. Hoe ze precies bepalen welke pagina's bovenaan moeten staan, is nog steeds niet volledig duidelijk en de strategie wijzigt ook af en toe, maar cruciaal daarbij is dat een pagina als 'belangrijk' wordt beschouwd als er veel links van andere pagina's naar deze pagina

verwijzen en deze andere pagina's bovendien eveneens belangrijk zijn. Essentieel daarbij is dat er geen menselijk oordeel aan te pas komt om te bepalen of een pagina al dan niet belangrijk is, maar dat deze analyse puur gebaseerd is op de inhoud en de structuur van de webpagina in combinatie met de rest van het internet. Dit *PageRank*-algoritme, genoemd naar de ontwerper Larry Page, illustreert hoe algoritmen – stap-voor-stapinstructies geschikt voor uitvoering op een computer – stilaan ons leven zijn gaan bepalen. Ondertussen is een hele bedrijfstak aan adviesbedrijven ontstaan die andere bedrijven adviseren over hoe ze hun website kunnen organiseren zodat ze in de zoekresultaten van Google bovenaan verschijnt.

Google is ongetwijfeld het bedrijf dat als een van de eersten het potentieel van grote hoeveelheden gegevens heeft ingezien en ten volle heeft benut. Toen ze in 1997 hun zoekmachine lanceerden, vroeg iedereen zich af wat het achterliggende businessmodel van deze gratis dienst kon zijn. Dat werd al snel duidelijk toen Google Adwords (nu Google Ads genoemd) ten tonele verscheen. Bij je zoekresultaten krijg je steevast advertenties die aansluiten bij je zoektermen en waarin je mogelijk dus geïnteresseerd bent. Een ingenieus veilingsysteem, waarbij enkel de advertenties van de meestbiedende adverteerders zichtbaar zijn, biedt een oplossing voor het plaatsgebrek op een webpagina en laat de kassa bij Google rinkelen.

Ondertussen ging Google onverdroten verder met het lanceren of overnemen van tientallen aanverwante producten zoals Gmail, Chrome, Maps, Photos, Meet (online vergaderen), Nest (kamerthermostaten en rookdetectoren) en YouTube om er slechts enkele op te noemen. Het gebruik van al deze producten is slechts mogelijk als je een Google-account aanmaakt. Als je gemakshalve overal hetzelfde account gebruikt, combineert Google alle data die je in

deze toepassingen achterlaat om zo een geïntegreerd beeld te krijgen van je verplaatsingen, beroep, hobby's, contacten enzovoort. Dit beeld gebruiken ze dan om je zoekresultaten – en de bijhorende advertenties – te optimaliseren.

Acht V's

> ***'As one Google Translate engineer put it, "when you go from 10000 training examples to 10 billion training examples, it all starts to work. Data trumps everything".'***
>
> — Garry Kasparov, voormalig schaakgrootmeester en auteur van het boek *Where Machine Intelligence Ends and Human Creativity Begins*

Het verhaal van Google in het vorige hoofdstuk illustreert meteen ook de belangrijkste eigenschappen van wat we vandaag als big data beschouwen en die in het Engels vaak aangeduid worden als de 8 V's.

Volume

De meest voor de hand liggende eigenschap is vanzelfsprekend de omvang – of het ***volume*** – van de gegevens. Dit begrip is evenwel zeer relatief. De grootte van een hoeveelheid computergegevens drukken we uit in *bytes*, wat op zijn beurt een verzameling is van 8 *bits*. Een bit, afkorting voor *binary digit*, kan de waarden 0 of 1 aannemen. Het is de elementaire data-eenheid bij de digitale computers. De toekomstige generatie computers, de zogenaamde *quantum computers*, die wellicht in de komende tien jaar op ons afkomen, gebruiken daarentegen *qubits*, die de waarden 0, 1, beide of alles daartussenin kunnen aannemen. Hier beperken we ons

tot de klassieke digitale computers die het moeten stellen met 0 en 1. Net zoals de eenheden uit de natuurkunde, zoals meter, gram of watt, geven we veelvouden van bytes aan met kilo, mega, giga enzoverder. Concreet is 1 kilobyte, afgekort tot kB, gelijk aan 1000 bytes, 1 megabyte (of MB) gelijk aan 1000 kilobytes, of één miljoen bytes enzoverder.

De capaciteit van de harde schijf van een moderne standaard desktop- of laptopcomputer heeft een grootteorde van 1 terabyte (TB) of 1000 gigabytes (GB) of één miljoen megabytes. Klassieke dataverwerkingssystemen kunnen gegevensverzamelingen van een dergelijke omvang nog vrij eenvoudig verteren en hier spreken we dan ook meestal niet over big data, althans niet op basis van het criterium volume. Ons ongebreidelde gebruik van sociale media daarentegen is een goede illustratie van de gigantische berg gegevens die we tegenwoordig produceren. Zo delen we in 2021 met zijn allen *elke minuut* 500.000 Facebook-commentaren, versturen we bijna 70 miljoen WhatsApp-berichten en voegen we 500 uur video toe aan YouTube. Dergelijke collecties beslaan al snel meerdere petabytes (1 petabyte = 1000 terabytes) of zelfs exabytes (1 exabyte = 1000 petabytes). Ook de tegenwoordig alom aanwezige sensoren produceren massale hoeveelheden data. De sensoren van een motor van een Boeing 737 produceren per vlieguur 20 TB met meetgegevens over bijvoorbeeld temperatuur, druk of chemische samenstellingen. Als we weten dat een 737 over twee motoren beschikt, dat een binnenlandse vlucht in de Verenigde Staten gemiddeld zes uur duurt en dat er dagelijks 28.537 dergelijke vluchten plaatsvinden, dan kunnen we gemakkelijk berekenen dat deze Boeings jaarlijks 20 TB × 2 × 6 × 28.537 × 365 = 2,5 miljard terabytes of 2,5 miljoen petabytes of 2500 exabytes produceren. Het vergt speciale opslaginfrastructuur om dergelijke dataverzamelingen op te slaan.

Het is belangrijk op te merken dat het bij big data niet enkel gaat over gestructureerde data zoals sensorgegevens of boekhoudkundige cijfers, maar steeds meer over ongestructureerde data zoals teksten of semigestructureerde data zoals spreadsheets. Deze ongestructureerde gegevens verdubbelen naar schatting elke achttien maanden in omvang.

Ook de ongeveer tien miljard webpagina's, van elk gemiddeld meerdere megabytes, die Google in zijn zoekmachine beheert, zijn een voorbeeld van ongestructureerde gegevens.

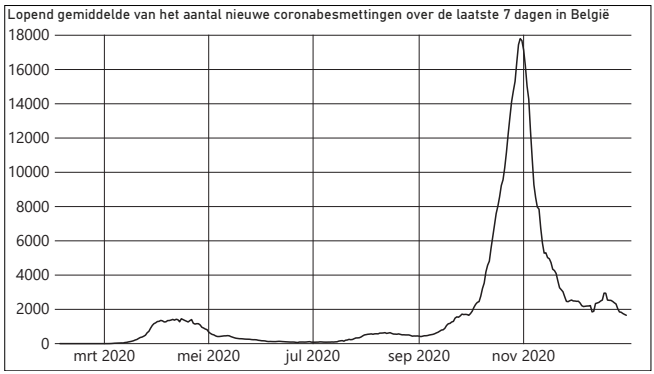
Velocity

In tegenstelling tot wat de term big data laat vermoeden, is volume zeker niet het enige criterium dat bepaalt dat we kunnen spreken over big data. De snelheid – of ***velocity*** – waarmee deze gegevens kunnen ontstaan en moeten worden verwerkt is vaak minstens even belangrijk. Het heeft immers geen enkele zin om bijvoorbeeld de data die tijdens de vlucht van een vliegtuig ontstaan pas na de landing te analyseren. Het kan hier gaan over cruciale vluchtgegevens waar bepaalde patronen een indicatie kunnen zijn van nakend onheil. Om ongevallen te voorkomen is het dus cruciaal om deze data, of minstens een deel ervan, nagenoeg onmiddellijk na het uitvoeren van de meting te analyseren.

Veracity

Een derde, niet te versmaden eigenschap van big data is de waarheidsgetrouwheid (in het Engels ***veracity***). Big data zijn vaak een nevenproduct van bepaalde processen. Aangezien bij de creatie van deze gegevens absolute correctheid niet onmiddellijk de grootste bezorgdheid is, ontstaan op die manier dikwijls gebrekkige of rommelige data. 'Het is een weekendeffect', 'Wij tellen de doden anders dan in het buitenland', 'Onze teststrategie is veranderd', 'Er

is een vertraagde registratie bij sommige Waalse rusthuizen' en 'We hebben slechts fragmentarische toegang tot de data van de contactopsporing' zijn slechts enkele van de uitspraken die we tijdens de coronacrisis vaak hebben gehoord uit de mond van de Vlaamse virologen en biostatistici. Samengevat gaat het over gegevens die onnauwkeurig, vaag, onzeker, onvolledig of inconsistent zijn. De cijfers van de coronabesmettingen zijn bijvoorbeeld onderling inconsistent omdat de teststrategie in de meeste landen tussen de eerste en de tweede golf totaal veranderd is, waardoor de golven onderling onmogelijk te vergelijken zijn, zoals **figuur 2** illustreert.



Figuur 2 Bij de eerste golf in België ontdekten de laboratoria naar schatting een dertigste van de werkelijke coronabesmettingen, bij de tweede golf naar schatting een derde. Dat geeft de verkeerde indruk dat de tweede golf, uitgedrukt in aantal besmettingen, veel erger was dan de eerste.

Gegevensbron: https://ourworldindata.org/coronavirus-source-data.
Gebruikte visualisatiesoftware: Microsoft Power BI.

Sensorgegevens bevatten vaak foutieve of ontbrekende registraties door bijvoorbeeld het tijdelijk falen van een meettoestel of een kortstondige netwerkonderbreking. Deze gebrekkigheid van de data vergt speciale aandacht en dito technieken bij het analyseren en interpreteren van de data om het bekende fenomeen in de computerwetenschappen, namelijk *garbage in, garbage out* te vermijden. Anderzijds is het niet allemaal kommer en kwel. Doordat we over zo veel data beschikken, kunnen we het ons permitteren om slechte data te wissen, of zelfs te corrigeren op basis van de goede data, want bij bigdata-analyses gaat het, anders dan bijvoorbeeld bij het opstellen van facturen of het uitbetalen van lonen, niet om exactheid maar volstaat het om voldoende data over te houden om patronen te zien en op basis daarvan beslissingen te kunnen nemen.

Variety

De nevenproducten van Google, zoals YouTube, Photos of Nest, zorgen ervoor dat ze gegevens uit diverse bronnen en met een grote ***variëteit*** kunnen combineren, wat ons meteen bij de vierde V brengt. Bigdata-analyses komen inderdaad vaak pas volledig tot hun recht indien de data afkomstig zijn uit een gevarieerde bronnenverzameling. Om het energieverbruik voor verwarming in een woning te voorspellen voor de komende dagen, kun je natuurlijk gebruikmaken van historische data over dezelfde woning of vergelijkbare woningen gedurende dezelfde periode in vorige jaren. Het resultaat zal vanzelfsprekend veel accurater zijn indien je deze gegevens combineert met de weersvoorspellingen. Dat is een illustratie van de variëteit van gegevens. Een ander aspect van variëteit heeft te maken met de diverse vormen die gegevens kunnen aannemen. Niet enkel de klassieke cijfers, vaak georganiseerd in rijen en kolommen zoals dagelijkse verkoopgegevens of tempera-

tuurregistraties, maar ook geluidsbestanden, foto's, video's en zelfs handgeschreven teksten kunnen als bron voor bigdata-analyses dienen, zoals we later uitvoerig gaan belichten.

Viscosity

De kracht van big data zit in het potentieel om patronen te herkennen in grote hoeveelheden gegevens. Het volstaat evenwel niet om 'veel' gegevens te hebben, ze moeten ook voldoende uitgebalanceerd zijn over de verschillende combinaties om statistisch relevante uitspraken te kunnen doen. Het is bijvoorbeeld erg moeilijk om te besluiten dat iemand van ouder dan 85 jaar die een zittend beroep had meer kans maakt om een bepaalde ziekte te ontwikkelen als in het gebruikte databestand slechts een handvol personen met die combinatie van kenmerken aanwezig zijn. De spoeling mag dus niet te dun zijn, de data moeten met andere woorden voldoende ***viskeus*** zijn. Dit is de vijfde V.

Virality

Daarnaast hebben data vaak ook een vervaldag. Het is bijvoorbeeld voor Google wellicht niet meer relevant om gegevens over je verplaatsingen, hobby's of beroep, die ze tien jaar geleden over jou verzameld hebben, nog steeds te gebruiken om te beslissen welke advertenties voor jou interessant kunnen zijn. Integendeel, indien je gezins- of werksituatie, en misschien ook wel je financiele status, ondertussen grondig gewijzigd zijn, kunnen deze oude data net leiden tot totaal irrelevante advertenties. Het is duidelijk dat deze vervaldag voor elke toepassing totaal verschillend is. Dit fenomeen noemt men de ***viraliteit* van gegevens**, wat erop duidt dat de relevantie van bepaalde gegevens in het verloop van de tijd aanvankelijk stijgt, een piek bereikt en daarna weer daalt.