

STAtOR

THEMA **BIG DATA**

De Opkomst van de Data Scientist

Big Data en statistiek

BIG DATA? Het is maar wat je gewend bent ... – column

Big Data, het einde van de besliskunde?

Puppy's, injectienaalden en attitudevorming; een eye-tracking studie

Big Data en officiële statistiek. De relevantie van heel veel tweets

Eenvoudige modellen en 'Grote Gegevens' troeven slimme modellen af

Een ode aan Data N. Richmond – column

NS-project AURORA voorspelt voor elke trein het aantal reizigers

Big Data kent geen routineklussen

De dood of de gladiolen – column

Fokwaardeschatting bij dieren

Inferentiële statistiek versus machine learning

Data en geïnformeerde toestemming?

Analyses van genexpressies bij kankeronderzoek

Optimaliseren met Big Data

Wiskunde à la mode – column

STAtOR is een uitgave van de Vereniging voor Statistiek en Operationele Research (VvS+OR). STAtOR wil leden, bedrijven en overige geïnteresseerden op de hoogte houden van ontwikkelingen en nieuws over toepassingen van statistiek en operationele research. Verschijnt 3 keer per jaar.

Redactie

Joaquim Gromicho (hoofredacteur), Ana Isabel Barros, Johan van Leeuwen, Richard Starmans, Gerrit Stermerdink (eindredacteur), Hilde Tobi en Vanessa Torres van Grinsven. Vaste medewerkers: Fred Steutel, Henk Tijms

Kopij en reacties richten aan

Prof. dr. J. A. S. Gromicho (hoofredacteur), Faculteit der Economische Wetenschappen en Bedrijfskunde, afdeling Econometrie, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, telefoon 020 59 86 010, mobiel 06 55 88 67 47, <j.a.dossantos.gromicho@vu.nl>

Bestuur van de VvS+OR

Voorzitter: prof. dr. Jacqueline Meulman <president@vvs-or.nl>
Secretaris: vacature <bestuur@vvs-or.nl>
Penningmeester: dr. Ad Ridder <penningmeester@vvs-or.nl>
Studentlid: Maarten Kampert MSc <student@vvs-or.nl>
Overige bestuursleden: prof. dr. Fred van Eeuwijk (BMS), dr. John Poppelaars (NGB), dr. Eric Cator (SMS), dr. Michel van de Velden (ECS), dr. Jelte Wicherts (SWS)

Leden- en abonnementenadministratie van de VvS+OR

VVS, Postbus 244, 6700 AE Wageningen, telefoon 0317 419 572, <admin@vvs-or.nl>.
Raadpleeg onze website over hoe u lid kunt worden van de VVS of een abonnement kunt nemen op STAtOR of op een van de andere periodieken.

VvS+OR-website

www.vvs-or.nl

Sociale media

Wilt u uw vakgenoten ontmoeten en wilt u discussiëren over actuele thema's, volg dan de VvS+OR en de Young Statisticians via LinkedIn, Facebook, Twitter en Flickr.
Sluit je aan bij de LinkedIn-groep van VvS+OR of Young Statisticians; bekijk foto's op <www.flickr.com/photos/vvs-or/sets>; Like onze Facebook-pagina; volg de President van VvS+OR op <https://twitter.com/#!/dutchstat>.

Advertentieacquisitie

<adverteren.stator@vvs-or.nl>.
STAtOR verschijnt in maart, juni en december.

Ontwerp en opmaak

Pharos | M. van Hootegem

Uitgever

© Vereniging voor Statistiek en Operationele Research
ISSN 1567-3383

BIG DATA

Dit themanummer van STAtOR gaat over BIG DATA. Een boeiend onderwerp met ongelooflijk veel facetten. Men kan overigens tegenwoordig geen tijdschrift op ons vakgebied meer openslaan zonder op de term Big Data te stuiten. De laatste twee jaar is het vrij plotseling opgekomen, het lijkt wel een modeverschijnsel. Toch is dat niet waar, Big Data staan al langer in de belangstelling. Al in november 2003, dus 10 jaar geleden, organiseerde de Sectie Statistische Programmatuur van de VvS+OR haar tweejaarlijks Symposium Statistische Programmatuur met als onderwerp *Large Data Sets*. We citeren uit de Engelstalige aankondiging:

Fifteen years ago, handling of large datasets, let alone analysis on them was a nearly impossible task for researchers. The data were often stored on tape, and even the process of reading the dataset into the memory of a mainframe was slow. Memory was scarce, and so it was difficult to save intermediate results. Such datasets were analyzed using either tailor-made statistical software, or self-written programs using routines from numerical libraries like NAG or IMSL. Maximum-likelihood estimation of non-linear models was non-trivial if not impossible, and researchers often had to be satisfied with one-step improvements over some consistent estimators. Things have changed for the better, from a technical point of view. Huge datasets are routinely available now to researchers in different fields, like finance, marketing, biomedical sciences, particle physics, astronomy, life sciences, and social sciences. Datasets used to be large in the sense of containing many observations on a small number of variables. But nowadays, e.g. in the life sciences we are confronted with datasets with a small number of observations and a huge number of variables. We mogen achteraf best trots zijn op een zo visionaire aandacht van een VvS+OR-sectie.

Als we naar de bijdragen aan dit nummer kijken zien we veel van die aspecten die tien jaar geleden werden genoemd terug.

Op onze algemene oproep in het vorige nummer om artikelen is goed gereageerd, en ook de vele persoonlijke contacten van redactieleden hebben resultaat gehad. Het is dan ook een BIG nummer geworden, misschien moeten we wel zeggen een HUGE nummer. De inhoud is zeer afwisselend en boeiend. Van fundamentele filosofische overwegingen tot zeer concrete toepassingen, er komt erg veel aan de orde.

Vanwege het grote aantal bijdragen zien we in deze inleiding af van het gebruikelijke korte overzicht. We laten het ontdekken ditmaal aan uzelf over.

Met een aan het thema aangepaste formulering van onze traditionele slotzin wensen wij u BIG FUN bij het lezen.

de STAtOR-redactie

INHOUD

- 3 Big Data
- 4 De Opkomst van de Data Scientist
Marc Teerlink & Olav Laudy
- 9 Big Data en statistiek
Kirk Borne
- 12 BIG DATA? Het is maar wat je gewend bent ...
– column
Gerrit Stermerdink
- 14 Big Data, het einde van de besliskunde?
John Poppelaars
- 17 Puppy's, injectienaalden en attitudevorming;
een eye-tracking studie
Roxanne I. van Giesen
- 21 Big Data en officiële statistiek.
De relevantie van heel veel tweets
Piet Daas
- 24 Eenvoudige modellen en 'Grote Gegevens' troeven slimme modellen af
Djoerd Hiemstra
- 27 Over Münchhausen en andere opscheppers.
Een ode aan Data N. Richmond – column
Fred Steutel
- 28 NS-project AURORA voorspelt voor elke trein het aantal reizigers
Jan Hoogenraad, Aart de Vos & Jaap de Vries
- 32 Big Data kent geen routineklussen
Sander Klous
- 35 De dood of de gladiolen – column
Henk Tijms
- 36 Fokwaardeschatting bij dieren. Het schatten van erfelijke aanleg door het oplossen van veel vergelijkingen
Han Mulder
- 40 Inferentiële statistiek versus machine learning:
een aanzet tot verzoening
Richard Starmans
- 43 Bewaren, delen en koppelen van data alleen na geïnformeerde toestemming?
Marcel Verwey
- 46 Analyses van genexpressies bij kankeronderzoek
Kees van Montfort
- 48 Optimaliseren met Big Data: hoe door de schaalbaarheidsmuur heen te breken?
Thierry Benoist, Julien Darlay, Bertrand Estellon, Frédéric Gardi & Romain Megel
- 53 Wiskunde à la mode – column
Johan van Leeuwen
- 54 Young Statisticians
- 55 Back to school 2014
- 55 OPROEP om kandidaten te nomineren voor de VVS+OR Master's Thesis 2013 en de Willem R. van Zwet Award 2013

Mocht u na het lezen van dit nummer nog niet genoeg hebben van Big Data dan moet u zeker de Dag voor Statistiek & OR op 20 maart 2014 bezoeken. Die dag staat in het teken van de 80ste verjaardag van ons erelid Willem van Zwet. Een van de Keynote speakers, Peter Bickel van Berkeley University, zal in zijn lezing aandacht besteden aan Big Data.

In STAtOR 14(2013)2 werd vermeld dat de tekst over Stan van Eeden ontleend was aan de website van de Wiskundemeisjes. Wat wij niet wisten was dat die tekst was geschreven door Willem van Zwet.



DE OPKOMST VAN DE DATA SCIENTIST

MARC TEERLINK & OLAV LAUDY

In de film *Moneyball* wordt Billy Beane manager van de Oakland Athletics, een honkbalteam uit Californië.¹ Vanwege de geldproblemen van de club gebruikt hij een wel heel onorthodoxe manier om resultaten te halen. Hij gebruikt, samen met zijn assistent Peter Brand, een compleet door computers gegenereerde analyse om zijn potentiële nieuwe spelers te vinden en zo een team samen te stellen voor de World Series. Tijdens de Amerikaanse

verkiezingen van 2012 werd Obama's campagne gerund als het politieke equivalent van *Moneyball*. Achter iedere beslissing van de campagne, zelfs het onbeduidendste telefoontje, ging een cijfermatige analyse vooraf door een klein team, geleid door data-wetenschapper Harper Reed (Valk, 2012; Hijink, 2013).

Nate Silver is een statisticus die in Amerika beroemd werd door zijn accurate voorspellingen voor de presidents-

verkiezingen van 2008 en 2012. Silver startte in 2008 zijn datablog FiveThirtyEight <www.fivethirtyeight.com> waarop hij voorspellingen deed die met algoritmes werden onderbouwd. Hij werkte als journalist bij *The New York Times* en stapte onlangs over naar de televisiezender ESPN. Ervaren in *Moneyball*-achtige sportstatistiek voorspelde Silver accuraat de verkiezingsuitkomst van 49 van de 50 staten en 35 van de 2008 'senaat races'. Dit puur op basis van data, de zogenoemde Big Data (Silver, 2012).

Wat zijn Big Data?

Het begrip Big Data staat in eerste instantie voor 'veel'. Door de omvang en de snelheid waarmee de vele data gegenereerd worden, zijn ze niet meer met de traditionele technologieën te verwerken. Data als nieuw productiemiddel, een delfstof die nooit op raakt, waarbij het bedrijfsleven de informatie uit die data gebruikt in de bedrijfsvoering. Om klanten beter te bedienen, om efficiënter te gaan werken en hierdoor meer winst te maken, om de samenleving te verbeteren, of om ons als consumenten een grotere stem te geven (Mayer-Schönberger, 2013). Door het koppelen van data en slim filteren ontstaan tal van nieuwe mogelijkheden. Tim Berners-Lee, de 'uitvinder' van het world wide web, voorspelde jaren geleden al dat door de groei van het internet uit een digitaal proces enorm veel nieuwe mogelijkheden zullen ontspruiten. Silver refereert in zijn blog en boek (Silver, 2012) aan de schatting van IBM dat er dagelijks 2,5 quintillion (inderdaad, zeventien nullen) nieuwe bytes data gecreëerd worden (IBM, 2011). Big Data staan niet alleen voor een grote hoeveelheid gegevens, maar ook voor een veranderlijke snelheid (zoals analyse van *streaming data*), waarheidsgetrouwheid (onzekerheid) en variatie. Verder komen data voor in verschillende vormen en formaten, variërend van het merk tandpasta dat je gisteren kocht tot de locatie vanwaar je naar je kantoor belde. Er zijn ook andere soorten data dan de gebruikelijke lijstjes, formulieren en databases. Denk hierbij aan data die een andere structuur hebben, afkomstig zijn uit sensoren, boeken en documenten, beeld, geluid, gps-locaties of die verkregen zijn door zelf gedane fysieke metingen (hardlopen, bloeddruk en dergelijke).

FEITEN EN CIJFERS

Datastromen veranderen continu, het onderstaande overzicht is slechts een momentopname.

- Mark Zuckerberg van Facebook stelt dat er dagelijks één miljard stukjes *content* worden gedeeld via Facebook (status, vrienden, groepen, likes).
- Eric Schmidt van Google stelt dat we tegenwoordig elke twee dagen net zoveel nieuwe data maken als vanaf het begin van onze beschaving tot 2003.
- Tachtig procent van alle data is ongestructureerd via sensoren, weerpatronen, likes, commentaren, reviews en artikelen.
- De hoeveelheid ongestructureerde data groeit vijftien keer sneller dan gestructureerde data.
- Elke seconde wordt er een uur aan videofilmjes ge-upload naar Youtube.
- Wereldwijd worden er elke 48 uur een miljard tweets verzonden. In maart 2013 was dit het aantal tweets in drie dagen en in oktober 2012 in een week.
- Het record aantal tweets was 456 per seconde, direct na het overlijden van Michael Jackson.
- Zeventig procent van alle data komt van individuele mensen, tachtig procent daarvan wordt door bedrijven opgeslagen en verwerkt.
- Elke maand zijn er meer dan honderd miljard zoekopdrachten op Google, in 2007 was dit nog 2,7 miljard.
- Rond 2020 zal een derde van alle data opgeslagen of door de cloud zijn gegaan en hebben we 35 zetabytes aan data gecreëerd.
- Een onderzoek dat de oorsprong van de big bang probeert te verklaren, verwerkt 14 exabyte aan signalen en sensor-data per dag.
- Meer dan tachtig procent van de 247 miljard verzonden e-mails per dag bestaat uit spam.
- Verkeerde data of data van slechte kwaliteit kosten het Nederlandse bedrijfsleven jaarlijks 39 miljard. Het gaat dan voornamelijk om verkeerde adressen, verkeerde facturen, orders die geaccepteerd worden terwijl het artikel niet op voorraad is en een te optimistische kijk op openstaande bedragen.

Data scientist: de meest sexy job

Toen de *Harvard Business Review* (Davenport & Patil, 2012) in oktober 2012 verklaarde dat de functie van *data scientist* de meest sexy job van de 21ste eeuw is, werd Twitter gevuld met het gejuich van econometristen, operations-researchprofessionals, actuarissen en statistici. Daarnaast vroegen echter de meeste lezers zich in de hierover ontstane *stream* nog af wat zo iemand dan wel doet? Het antwoord op die vraag zal de komende jaren snel duidelijker worden. Steeds meer

bedrijven willen de vruchten van Big Data plukken, zodat deze specifieke taakomschrijving nauwkeuriger wordt ingevuld. De vraag die we vaak krijgen, is of er nog wel een rol is weggelegd voor traditionele beslissingsondersteuning in dit nieuwe tijdperk van *Big Data analytics*.

Eerdere research toonde al aan dat data-analisten een belangrijke en essentiële rol spelen bij de meer succesvolle ondernemingen in verschillende sectoren (Teerlink & Haydock, 2011; Finch, Rodenbeck Reese, Shockley & Balboni, 2013) en verwacht wordt dat deze rol toe zal ne-

men. Wat echter ook duidelijk wordt, is dat – zelfs bij de organisaties die data als belangrijke activa zien – tijdens de besluitvorming niet altijd de boodschap uit die data bruikbaar gemaakt is.

Het einde van de Sample vs Signals & Noise

De dagelijkse quintillions data zijn wat Silver (2012) in zijn blog en in zijn boek *signals* noemt, feiten die we willen hebben en die we nodig hebben, zoals de signalen die ons helpen om sociaal sentiment te herkennen of de benodigde hoeveelheid te bepalen van een specifiek product dat op een bepaalde locatie en tijdstip nodig zal zijn.

Noise of ruis is alle andere, meestal vreemde, informatie welke onze zoektocht naar signalen belemmert of ons daarvan afleidt. Big Data tonen wat er is, zonder dat we de reden daarvoor per se moeten kennen. Big Data houden minder of geen rekening met ‘causaliteit’. Dat laatste is een verandering in denken die dikwijls gemaakt moet worden om de meerwaarde van Big Data in te zien. Als mensen in bedrijven blijven steken in het statistisch analyseren van *samples*, dan biedt een grotere sample zoals in Big Data geen meerwaarde. Juist met Big Data zijn er soms honderden, zo niet duizenden variabelen en een slimme analist kan bijna hier altijd een formule bij creëren welke sommige van deze variabelen zo combineert dat historische resultaten verklaard worden. Veel van die variabelen beïnvloeden echter niet daadwerkelijk het resultaat, dus als je zo’n formule accepteert en je gebruikt vervolgens je middelen om die variabelen te beïnvloeden dan heb je eigenlijk geen invloed op de uitkomst. Hetgeen ons terugbrengt bij een van de belangrijkste discussies tussen *data scientists* en hun businesspartners over het verschil tussen correlatie en causaliteit! De meeste voorspellingen mislukken, vaak gepaard gaand met hoge kosten voor het bedrijf of de maatschappij, omdat de meeste businessmanagers een slecht begrip hebben van wat correlatie, causaliteit, waarschijnlijkheid en onzekerheid eigenlijk zijn. Zowel leken als ervaren gebruikers verwaren vaak de betrouwbaarheid van een voorspelling met de accuraatheid. Ook is overmoed vaak de reden voor Big Data-mislukkingen (Finch, Rodenbeck Reese, Shockley & Balboni, 2013). Als onze kennis van onzekerheid verbetert, kunnen onze voorspellingen beter worden. Dit is de ‘voorspelling paradox’.

Data delen is het nieuwe hebben

Uiteindelijk komt het neer op het volgende: business analisten zijn vaak niet in staat om te bieden wat het besluitvormingsproces het meeste nodig heeft: inzicht en niet alleen cijfers. De data scientists moeten dus nieuwsgierig zijn, lang en goed naar data kijken en trends oppikken. Zoals een geleerde uit de Renaissance moet de data scientist voor alles belangstelling hebben – niet alleen voor de business-aspecten – en een organisatie willen veranderen. Hij moet dienen als schakel tussen business en data. Wat vertellen Big Data ons over onze processen? Hoe kunnen wij beter en sneller functioneren? De data scientist moet meer doen dan alleen maar data analyseren of een model creëren. Het ontsluiten van Big Data is maar een klein onderdeel van zijn taakomschrijving. Hij dient ook de data te visualiseren waardoor de analyse van Big Data inzichtelijk gemaakt kan worden, en daarmee voorspellend wordt in plaats van reactief. Enorme hoeveelheden data bewaren heeft weinig zin als je er niks mee doet. De waarde van Big Data zit hem juist in de analyses die er op los worden gelaten en de wijze waarom er samengewerkt wordt tussen de verschillende bedrijfsonderdelen in het gebruik van de uitkomsten van *Big Data analytics*.

Tips voor de beginnende data scientist

We willen een drietal tips meegeven voor het vertalen van Big Data naar inzicht. Dit inzicht draagt bij aan het realiseren van doelen met meetbare verbeteringen.

1. *Begin met de vraag, niet met de Big Data, en beantwoord de juiste vraag*

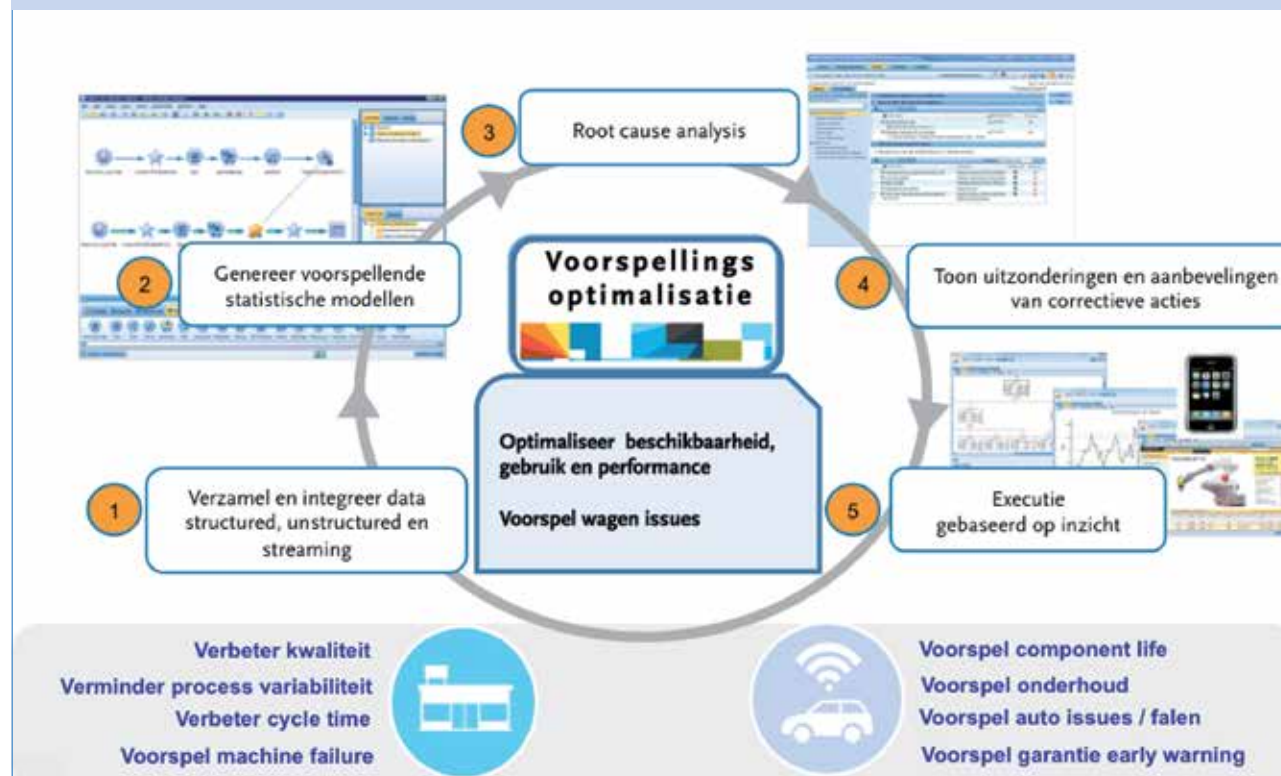
Beantwoord de juiste vragen. Met Big Data kan het verleidelijk zijn om antwoorden op allerlei interessante vragen na te jagen. Maar als na al het analyseren blijkt dat de focus is op vragen die niemand anders in de organisatie stelde, zul je óf delen aan het verhaal moeten toevoegen danwel je moet richten op de vragen die belangrijk zijn voor beslissers. Beter nog, je moet begrijpen dat het je eerste taak is om mensen te helpen betere vragen te formuleren.

2. *Vertel het verhaal; spreek de taal van je bedrijf*

Data scientists, statistici, operations researchers en actuarissen houden van hun data. Ze waarderen robuustheid, nuance en elegantie. Niemand anders geeft meer

BIG DATA BIJ EEN EUROPESE AUTOFABRIKANT: VOORSPELLINGSOPTIMALISATIE

- Produceert motoren in groepen van 400 tegelijk.
- Datametingen beginnen al tijdens het productieproces.
- Door het kleinste foutje tijdens het productieproces zal de motor later worden afgekeurd.
- Volledige inspectie van een motor kan pas plaatsvinden na afkoeling en een kostbare röntgenmeting.
- Afkoelen van de motor neemt veel tijd in beslag.
- Echter door een nauwkeurige voorspelling (monitoring/bewaking) kan een motor al direct na productie afgekeurd en gerecycled worden.
- Motoren die doorgaan naar de eindproductie worden blijvend gevolgd.
- Data van elke auto worden vanaf aflevering gevolgd.
- Externe data worden toegevoegd aan de type-gegevens en de data uit de zelfdiagnose (sensoren), onderhoudsrapporten van dealer, klantcommunicatie etc.
- De autofabrikant is hierdoor in staat om tijdig onvoorzien onderhoud te voorspellen en pro-actief garantie en service te verlenen.



WIND

Veel Big Data-projecten gaan niet over het vinden van onvoorziene verbanden maar over de complexe vraag om een logisch verband te vinden binnen door hoge snelheid en grote hoeveelheid gekenmerkte data. Het Deense Vestas (een producent en plaatser van windmolens) krijgt bijvoorbeeld dagelijks data binnen van meer dan 35.000 meteorologische stations hetgeen leidt tot 24 petabytes aan data (dat is 70.000 jaar HD-tv). Vestas ontdekte dat de plek met de meeste wind niet de meeste elektriciteit oplevert. <<http://bit.ly/1dRTcu2>>

om data dan zij. Maar als je wilt worden gehoord, zul je je bevindingen moeten vertalen in de taal die je partner of beoogd publiek spreekt. Wanneer mensen zelf moeten gaan interpreteren is de kans op invloed al verloren. Begin met het antwoord op de *so what*-vraag, in plaats van met een uitleg van de methodologie. Gebruik dynamische visualisaties om Big Data tot leven te brengen. Wat je ook doet, maak niet slechts één PowerPointslide met een gigantische data tabel...

3. Wees realistisch: wees bescheiden en valideer samen met businesscollega's

De meeste data-masseurs (inbegrepen de auteurs van deze bijdrage) trappen soms in de valkuil van het denken dat onze data de werkelijkheid zijn. We worstelen ons met allerlei variaties en modellen door de data en komen dan met een solide, fascinerende, controversiële conclusie. *So far, so good*. Ons instinct is om dan naar de dichtstbijzijnde beslisser te lopen en te zeggen: 'Kijk!! Je moet het bedrijf veranderen!' Maar dan doe je het niet goed. Het is heel goed mogelijk – misschien zelfs waarschijnlijk – dat we niet het volledige beeld hebben. Een of twee keer vals alarm kan gemakkelijk de geloofwaardigheid van *Big Data analytics* ondermijnen. Dus, in plaats van 'Je moet veranderen!' of 'Je doet dit mis?' is het essentieel om eerst een aantal aannames en conclusies te valideren samen met degenen die een goed inzicht hebben in het bedrijf en historische ervaring. Doe dit om te voorkomen dat er gênante situaties ontstaan waardoor toekomstige analyses voorzichtiger ontvangen zullen gaan worden. Valideer eerst met een of twee businesspartners; laat hen delen in het spel en de erkenning. Luister naar hen. Het is een win-win situatie. Silver (2012) sluit zijn boek dan ook af met de volgende quote '*The more humility we have about our ability to make predictions,*

the more successful we can be in planning for the future.'

Conclusie

Met de huidige technologie is het verzamelen en ontsluiten van Big Data buiten en binnen je bedrijf eenvoudig geworden. Moeilijker wordt het om met echt inzicht te komen. De hoofdprijs komt pas bij het effectief communiceren van deze inzichten.

Hoe meer nederigheid we kunnen opbrengen over ons vermogen om voorspellingen te doen, des te meer succes we zullen hebben in het plannen voor de toekomst.

NOOT

1. *Moneyball* is een Amerikaanse speelfilm uit 2011, geregisseerd door Bennett Miller. De film is gebaseerd op Michael Lewis' non-fictieboek *Moneyball: The Art of Winning an Unfair Game* uit 2003. Zie ook: <www.imdb.com/title/tt1210166/>, <<https://en.wikipedia.org/wiki/Moneyball.3>>, <www.mlbnewsathletics.com/news/nate-silver-brings-moneyball-world-politics>.

LITERATUUR

- Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, 90(10), 70-76.
- Finch, G., Rodenbeck Reese, C., Shockley, R. & Balboni, F. (2013). *Global Study, Analytics: A blueprint for value - Converting big data and analytic insights into results*. IBM Institute of Business Value Executive Report.
- Hijink, M. (2013, 13 mei). Obama's oppernerd ging de wereld redden. *NRC Handelsblad*, pp. 32-33.
- IBM (2011). *Bringing big data to the enterprise: What is big data?* Retrieved from <www-01.ibm.com/software/data/bigdata/>.
- Mayer-Schönberger, V. & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt.
- Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail — but Some Don't*. The Penguin Press.
- Teerlink, M., & Haydock, M. (2011). *Customer analytics pay off: Driving top-line growth by bringing science to the art of marketing*. IBM Institute of Business Value Executive Report.
- Valk, G. (2012, 23 oktober). Campagneteam Obama weet alles van elke kiezer. *NRC Handelsblad*, p. 5.

MARC TEERLINK is Global Strategist & Chief Data Scientist bij IBM.
E-mail <marc.teerlink@nl.ibm.com>

OLAV LAUDY is Worldwide Predictive Analytics Solutions Leader bij IBM.
E-mail <olavlaudy@nl.ibm.com>

BIG DATA EN STATISTIEK

Zullen Big Data ervoor zorgen dat Statistiek achterhaald raakt? Natuurlijk niet! Maar, velen suggereren dat wel.¹ We bespreken hier enkele van de oorzaken en gevolgen van deze manier van denken. Concreet kijken we naar enkele basisprincipes van de elementaire statistiek die bij het analyseren van grote uitgebreide '*full-population*' datasets gemakkelijk worden vergeten of terzijde geschoven.

KIRK BORNE

De methodologie die ons in staat stelt kennis te ontdekken in gegevens is de data-wetenschap. Deze omvat een hele familie van disciplines: een van de belangrijkste daarvan is statistiek. Ook het statistisch denken en redeneren zijn belangrijker dan ooit in het informatietijdperk.² Toch is de verleiding groot beproefde en fundamentele uitgangspunten van een statistische denkwijze naast zich neer te leggen. Een reden voor deze veronachtzaming kan zijn dat Big Data het gemakkelijk maken een handje te lichten met de statistische 'strengheid', want er zijn zoveel gemakkelijke ontdekkingen te doen in grote gegevensverzamelingen dat er blijkbaar geen behoefte meer is om de wiskundige machinerie van de statistiek in te schakelen.

Een korte blik in een standaard leerboek *computational statistics* kan deze manier van denken gemakkelijk doen versterken. Een andere reden voor statistische slordigheid kan zijn het gebrek aan een gedegen statistische opleiding in veel studierichtingen, waardoor het 'statistisch denken' zich niet ontwikkelt.³ Een meer verontrustende uitleg van dit negeren van strengheid is wanneer men van mening is dat de statistiek de discipline van 'kleine data'-analyse is. Immers, als je genoeg data tot

je beschikking hebt om een 1000-voudige cross-validatie of een 1000-variabelenmodel te kunnen maken met daarbij miljoenen replicaties van steekproeven, dan is verdere statistiek toch irrelevant?

Statistische tautologieën

Laten we eens vier fundamentele statistische tautologieën beschouwen die in het tijdperk van Big Data gevaar lopen:

1. Correlatie is geen causaliteit.

Iedereen weet dit, maar velen negeren het. Sommige durven zelfs triomfantelijk te beweren dat dit fundamentele principe van de statistiek niet langer een belangrijk begrip is bij het werken met Big Data. Dit omdat er zulke enorme aantallen correlaties kunnen worden ontdekt in massieve dataverzamelingen, dat er voor sommige van deze correlaties wel een oorzakelijk verband moet zijn. Zo'n redenering is voor hen evident.⁴ In feite, 'Big Data zullen onze obsessie met oorzakelijk verband wel doen verdwijnen' (Mayer-Schönberger & Cukier, 2013). Ik ben zelf deels schuldig aan deze manier van denken;

De Challenger desintegreerde 73 seconden na de lancering van zijn tiende vlucht STS-51-L op 28 januari 1986. Foto: Kennedy Space Center.

ik ben namelijk voorstander van een ongerichte ontdekkingstocht voor vele data-analyse projecten: het zoeken naar patronen, trends, correlaties en verbanden in data zonder vooropgezette ideeën of verwachte gedragsmodellen. Zo'n zoektocht naar de 'onbekende onbekenden' is een van de belangrijkste toepassingen van Big Data: ontdekkingen door datamining (Borne, 2000).

Maar in feite zijn causale relaties de belangrijkste ontdekkingen van big data analytics waar men naar streeft. Het vinden van oorzaken voor waargenomen effecten kan een goudmijn zijn voor ieder die zich met het analyseren van big data bezig houdt, onverschillig of men in een commercieel bedrijf, in de wetenschap, of bij de overheid werkt.

Deze manier van denken is aanvaardbaar als zij is geplaatst in een volledige wetenschappelijke, methodologische context: gevolgtrekking, hypothese-opstelling, experimenteel ontwerp, testen, valideren, hypothese-modificatie. Statistici herontdekken inmiddels ook het belang van causaliteit: een recente publicatie onderzoekt causaliteit bij het modelleren van dynamische systemen vanuit een systeem (mechanistisch) perspectief (Aalen, Roysland & Gran, 2012). Big Data leveren een schat aan informatie voor dynamische systemen van allerlei soorten, dus er is zeker veel reden om aandacht te besteden aan causaliteit, alsmede aan correlatie. Het negeren van statistische tautologie 1 kan leiden tot grappige of verkeerde conclusies. In menig boek of website zijn hiervan fraaie voorbeelden te vinden.

2. Steekproefvariantie gaat niet naar nul, zelfs niet met Big Data.

Dit is een andere makkelijk te vergeten waarheid bij het werken met Big Data. We zijn vertrouwd met het concept van statistische ruis en hoe deze afneemt als de steekproefomvang toeneemt. Maar steekproefvariantie is niet hetzelfde als ruis. De eerste is een fundamentele eigenschap van de populatie, terwijl de laatste is een eigenschap van het meetproces. De fout in onze voorspellende modellen zal waarschijnlijk niet kleiner te maken zijn dan een bepaalde drempel: dit is de intrinsieke steekproefvariantie. Een eenvoudig voorbeeld is regressie: zolang je overfitting van de data voorkomt, zal het regressiemodel zelden de exact juiste waarde van de afhankelijke variabele voorspellen. Voor complexere multivariate modellen geldt: hoe groter de steekproef, des te nauwkeuriger de geschatte variantie in verschillende parameters (variabelen) die de populatie beschrijven.

Dit is een goede zaak, het zou juist 'het ding' kunnen zijn waarnaar je op zoek bent, maar ook een van de fundamentele kenmerken van de populatie. In de astronomie noemt men dit 'kosmische variantie' (Driver & Robotham, 2010). Wij kunnen het gehele heelal bestuderen, met miljarden (zelfs biljoenen) objecten in onze steekproef, maar er zullen nog steeds verschillen in de verdelingen van verschillende parameters zijn bij het kijken in verschillende richtingen aan de hemel. Die kosmische verschillen zullen nooit naar nul gaan. Ook in andere domeinen krijgen we betere schattingen van de fundamentele statistische kenmerken van de populatie, de variantie over de klassen inbegrepen, wanneer we meer gegevens verzamelen. Statistische tautologie 2 is een goede zaak, omdat het voldoet aan een van de grote beloften van Big Data: het verkrijgen van de beste statistische ramingen ooit van de niet-nul parameters die de gegevens van de populatie beschrijven.

Samenvattend: steekproefvertekening kan leiden tot modellen met vertekende resultaten.

3. Steekproefvertekening hoeft niet noodzakelijkerwijs naar nul te gaan, zelfs niet met Big Data.

De neiging om deze eigenschap van de statistiek te negeren ontstaat vooral wanneer we biased methoden voor gegevensverzameling gebruiken, of wanneer onze modellen zijn onderbepaald. Dit laatste is meestal een gevolg is van een slecht modelontwerp en dus onafhankelijk van de hoeveelheid beschikbare gegevens. Simpel gezegd, vertekening kan voortvloeien uit de toepassing van een model dat onvoldoende is gebaseerd op het volledige potentieel van beschikbaar bewijsmateriaal. Zoals Albert Einstein zei: 'een model moet zo eenvoudig mogelijk zijn, maar niet eenvoudiger.' In het tijdperk van Big Data is het nog steeds mogelijk een eenvoudig voorspellend model op te stellen dat geen rekening houdt met relevante patronen en eigenaardigheden in de gegevens.

Een andere situatie, waarin vertekening niet verdwijnt als de steekproef groter wordt, ontstaat wanneer gecorreleerde factoren (of behandelingen) aanwezig zijn in een analyse die ten onrechte een statistische onafhankelijkheid veronderstelt (bijvoorbeeld in een reeks gerandomiseerde A/B trials, een veel voorkomende praktijk in *big data analytics*). In dergelijke gevallen blijft de vertekening bestaan, ongeacht de grootte van de steekproef (Kahan, 2013). Statistische tautologie 3 waarschuwt ons dat het feit dat we veel data hebben niet betekent dat we die gegevens op een juiste wijze toepassen in onze modellen. Ook hier kan steekproef-vertekening leiden tot modellen met vertekende resultaten.

4. Afwezigheid van bewijs is niet hetzelfde als bewijs van afwezigheid.

In het tijdperk van Big Data, vergeten we gemakkelijk dat we nog niet alles hebben gemeten. Zelfs met de alom aanwezigheid van data, hebben we nog steeds niet alle mogelijke gegevens over een bepaald onderwerp verzameld. Daarom moet men zich bij statistische analyses bewust zijn van de noodzaak om voor deze ontbrekende gegevens te corrigeren (gebrek aan bewijs), teneinde scheve conclusies te voorkomen. Omgekeerd kan 'afwezigheid van bewijs' ook zeer waardevolle informatie zijn, als je het kan aantonen. Wetenschappers hebben het belang van deze begrippen onderzocht bij de evaluatie van onderwijsprogramma's die gericht zijn op misbruik van middelen (Foxcroft, 2006). Ze vinden dat, hoewel het onderscheid tussen de twee begrippen ('bewijs van afwezigheid' versus 'gebrek aan bewijs') belangrijk is, sommige beleidsbeslissingen en maatschappelijke reacties op belangrijke problemen toch doorgang moeten vinden.

Dit is een a-typisch geval – meestal zorgt het onderscheid tussen de twee begrippen voor een beïnvloeding van de besluitvorming en de voortgang van het verdere onderzoek. Ik heb bijvoorbeeld een keer aan een directeur van een grote sterrenwacht voorgesteld om een database te maken van zaken die met zijn telescopen waren gezocht, maar nooit gevonden – de EAD: de *Evidence of Absence Database*. Hij vond het een goed idee om een dergelijk 'instrument' in te zetten om daarmee het gebruik te minimaliseren van zijn apparatuur voor zoekacties waarbij we allang weten dat die geen zin hebben. Helaas bood hij niet aan te betalen voor de opzet van zo'n database. Ik geef een voorbeeld van een wetenschappelijk artikel dat aangeeft dit concept wél te hebben begrepen (Berry, Brunner, Popescu & Shukla, 2011). De titel van het paper luidt *Can apparent superluminal neutrino speeds be explained as a quantum weak measurement?* Het volledige abstract bestaat slecht uit één zin: 'Probably not'.

Een meer dramatisch – en rampzalig – voorbeeld van het negeren van dit statistische concept is de ramp met NASA Shuttle Challenger in 1986. Ingenieurs hadden aangenomen dat het ontbreken van bewijs van falen van de O-ringen, tijdens lanceringen bij koud weer, een bewijs was dat O-ringen niet zouden falen tijdens zo'n lancering (Tuft, 1997). In dit geval waren de gevolgen van deze gebrekkige statistische redenering catastrofaal. Dit is natuurlijk een extreem voorbeeld, maar verwaarlozing van statistische tautologie 4 is nog steeds een valkuil die we in het tijdperk van Big Data moeten vermijden.

Waarheen leidt dit ons? Het leidt ons naar een duidelijk voorbeeld van correlatie met causaliteit. Als we het tijdperk van Big Data en analytische toepassingen betreden, zou het gebruik van deze toepassingen kunnen correleren mét en misschien zelfs een ontbreken (of verkeerd toepassen) van statistisch denken veroorzaken. En dat uitgerekend op de thuisplaneet van het Big Data universum: de statistiek!

Dit artikel is, met als titel 'Statistical Truisms in the Age of Big Data', eerder verschenen op 19 juni 2013 op de site <www.statisticsviews.com/details/feature/4911381/Statistical-Truisms-in-the-Age-of-Big-Data.html>. De auteur heeft de redactie toestemming verleend het te vertalen.

NOTEN

1. Zie de discussie hierover bij Larry Wasserman <<http://normaldeviate.wordpress.com/2013/04/13/data-science-the-end-of-statistics/>>.
2. H.G. Wells said, 'Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write!' Retrieved from <www.causeweb.org/cwis/SPT-FullRecord.php?ResourceId=1240>.
3. <www.amstat.org/publications/jse/v11n3/chance.html>.
4. <www.newstatesman.com/sci-tech/sci-tech/2013/05/are-you-ready-era-big-data>.
5. <www.iveybusinessjournal.com/topics/strategy/four-strategies-to-capture-and-create-value-from-big-data>.

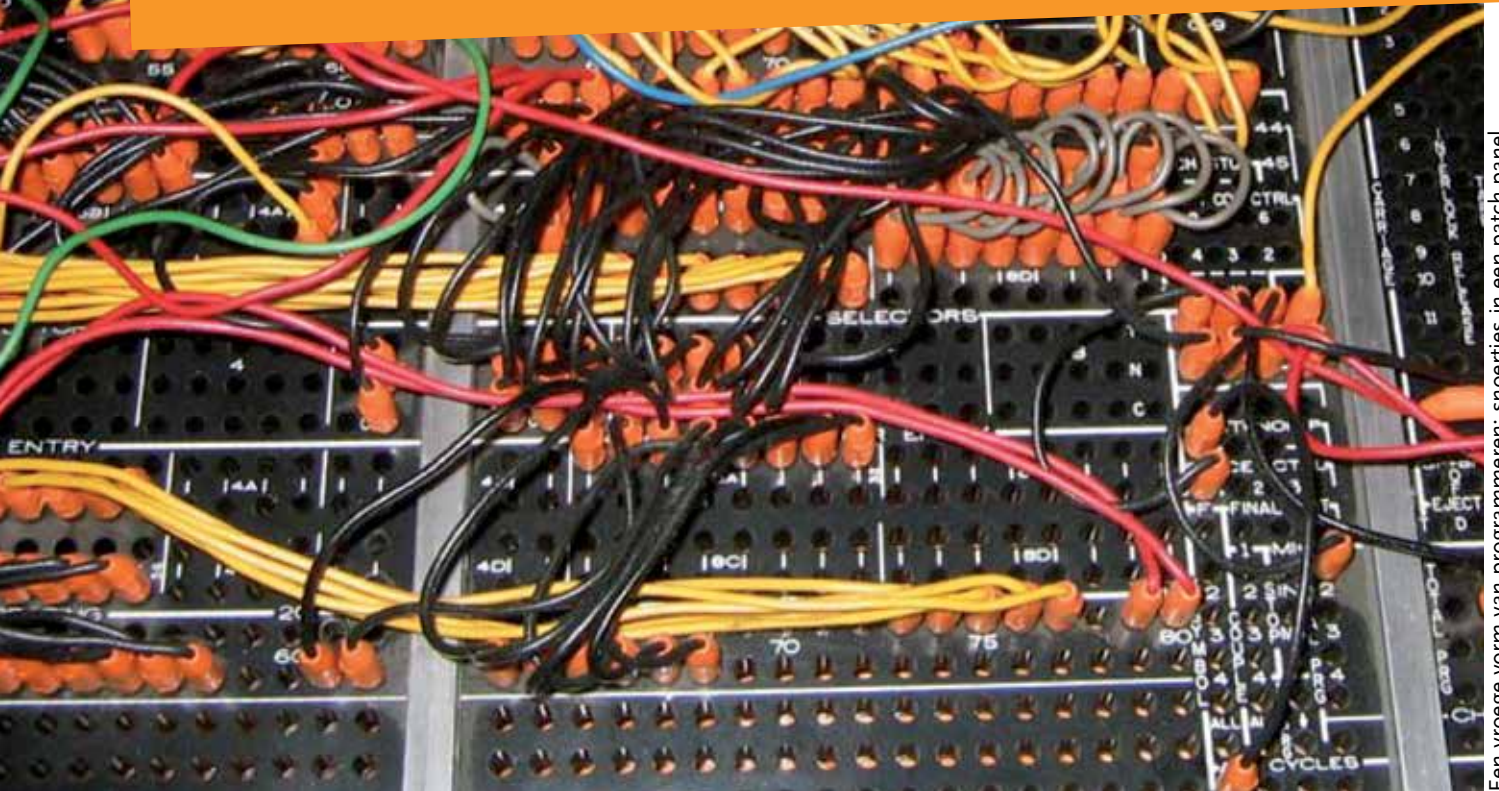
LITERATUUR

- Aalen, O., Roysland, K. & Gran, J. M. (2012). Causality, Mediation and Time: A Dynamic Viewpoint. *Journal of the Royal Statistical Society Series A*, 175(4), 831-861.
- Berry, M. V., Brunner, N., Popescu, S. & Shukla, P. (2011). Can apparent superluminal neutrino speeds be explained as a quantum weak measurement? *Journal of Physics A: Mathematical and Theoretical*, 44(49). doi: 10.1088/1751-8113/44/49/492001
- Borne, K. (2000). Data Mining in Astronomical Databases. Retrieved from <<http://arxiv.org/abs/astro-ph/0010583>>.
- Driver, D. & Robotham, A. (2010). Quantifying Cosmic Variance. *Monthly Notices of the Royal Astronomical Society*, 407(4), 2131-2140.
- Foxcroft, D. (2006). Alcohol Education: Absence of Evidence or Evidence of Absence. *Addiction*, 101(7), 1057-1059. doi: 10.1111/j.1360-0443.2006.01513.x
- Kahan, B. (2013). Bias in Randomised Factorial Trials. *Statistics in Medicine*, 32(26), 4540-4549. doi: 10.1002/sim.5869
- Mayer-Schönberger, V. & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt.
- Tufts, E. R. (1997). *Visual & Statistical Thinking: Displays of Evidence for Decision Making*. Graphics Press.

KIRK BORNE is PhD DataScientist Astrophysicist en hoogleraar Astrophysics and Computational Science bij George Mason University in Fairfax VA.
E-mail: <kborne@gmu.edu>

BIG DATA?

Het is maar wat je gewend bent ...



Een vroege vorm van programmeren: snoertjes in een patch panel.

Ik ben bij toeval in de statistiek beland. Zo hoort het eigenlijk ook bij een vak dat zich met de gevolgen van toeval bezighoudt. Indertijd stond de computerverwerking nog in de kinderschoenen en iets was al snel BIG DATA. Maar dat is gaandeweg veranderd.

Mijn eerste begin was in 1965 bij een marktonderzoeksbureau en daar was alles BIG DATA wat niet in gecodeerde vorm op 1 ponskaart met 80 kolommen paste. De hele verwerking bestond uit het sorteren van die kaarten op de inhoud van één kolom en dan op een formulier noteren wat de tellers van de machine aangaven voor ieder van de mogelijke codes 1 t/m 12 in die kolom. Kruistabellen konden we ook maken, dan sorteerden we eerst op een kolom en vervolgens ieder apart

stapeltje op een andere kolom. Als het onderzoek groot was pasten niet alle vragen op 1 kaart en werden noodgedwongen alle basisgegevens, zoals geslacht, leeftijd, inkomen etc., ook op iedere volgende kaart herhaald. Het was namelijk niet mogelijk om een kruistabel naar leeftijd te maken als die leeftijd niet op dezelfde kaart stond gecodeerd. Op een avond vond ik achterin een stoffige kast een doos snoertjes met stekkers en een handleiding. Zo kwam ik er achter dat we eigenlijk een heel geavanceerde sorteermachine hadden: je kon hem met snoertjes in een groot bord met gaten programmeren! Ineens was BIG DATA een dimensie verder. Nu kon ik de meerdere kaarten die bij één persoon hoorden als één geheel behandelen en hoefde niet meer al die ba-

sisinformatie op alle kaarten aanwezig te zijn. Met een simpele code gaf je aan of een kaart een hoofdkart of een vervolgkaart was, daardoor vielen alle kaarten van een persoon in hetzelfde sorteervak. De dankbaarheid van de directie voor deze ontdekking leverde me een opslag van 10 cent per uur op.

Na dat bureau heb ik heel lang bij een instituut voor onderwijsresearch gewerkt. Daar gebruikten we een échte IBM mainframe computer en hadden we een stel min-of-meer standaard programma's voor rechte tellingen, kruistabellen, variantie-, regressie- en zelfs factoranalyse (SPSS was toen nog niet in Nederland bekend). Zoals alle computers doen, ook nu nog, kunnen ze alleen werken met zaken die in het kerngeheugen aanwezig zijn. De omvang van het probleem wordt dus beperkt door de omvang van dat geheugen. Daarom had het programma voor factoranalyse twee versies: een voor 40 variabelen en een voor 60. Die grote versie kon je alleen maar 's avonds en 's nachts draaien want een job overdag was beperkt tot een maximale geheugengrootte waar een 60x60 matrix niet in paste. Omstreeks 1970 kwam IBM met het revolutionaire concept van het virtuele geheugen. Daardoor kon je met grotere programma's en grotere matrices werken: wat niet in het geheugen paste werd er pas in gestopt op het ogenblik dat het nodig was; een ander op dat ogenblik ongebruikt stuk programma werd overschreven en op zijn beurt weer teruggehaald als dát stuk verderop in de verwerking weer nodig was. Dat zogenaamde swappen was relatief langzaam en het vereiste slim programmeren om dat zoveel mogelijk te beperken. Weer een stap verder in het begrip BIG DATA.

Toen kwamen SPSS en andere programma's als BMDP beschikbaar. De Fortran programmeurs veranderden langzamerhand in dataverwerkers en de wetenschappers waarvoor ze werkten kregen dat ook in de gaten. In hun onschuld dachten die dat nu alles mogelijk was.

Ik herinner me een onderzoek van een onderwijs-socioloog die 1200 leraren een vragenlijst met zo'n 200 vragen had laten invullen. Dat was nog eens BIG DATA! Ik kreeg opdracht om kruistabellen te maken van alle variabelen tegen elkaar. Dat leek me nonsens, ik vond dat hij eerst maar eens hypothesen moest formuleren. Verder had ik uitgerekend dat de totale uitvoer van de grote regeldrukker rond de 5 dozen van ieder 4000 vel papier zou bevatten. Tja, daar had ik een punt, hij zou nooit de tijd vinden om al die tabellen te bestuderen. Maar een dag later kwam hij met een idee: als ik daarnaast ook een 200x200 correlatiematrix zou produceren kon hij zich beperken tot het bekijken van alleen die tabellen waar een hoge correlatie bij hoorde. Mijn volgende bezwaar was dat het merendeel van de variabelen zich niet leende voor een eenvoudige Pearson correlatie. Omdat het een erg slimme socioloog was bedacht hij ook daar een oplossing voor: ik moest SPSS maar opdracht geven het commando Statistics All (kent u dat nog?) te gebruiken bij alle tabellen. Dan zou hij een goedkope student-assistent met een accentueer-stift alle tabellen laten doorlopen en hoge waarden van associatiematen van een kleurtje laten voorzien. Op die manier kon hij zich beperken tot het bestuderen van de ca. 5% tabellen die naar schatting interessant waren. Gelukkig was zijn rekenbudget niet hoog genoeg en dat besliste de zaak.

Nu zijn de tijden geheel veranderd, per nanoseconde produceert een versneller als die bij CERN vele malen meer gegevens dan ik in mijn hele loopbaan bij elkaar heb verwerkt. Dat zijn pas BIG DATA! Maar er zal een tijd komen dat we ook die dimensies weer enkele malen zullen overtreffen. BIG DATA zijn niet te stoppen.

GERRIT STEMERDINK is eindredacteur van STATOR
E-mail: <gjsterdink@hotmail.com>



BIG DATA, het einde van de besliskunde?

JOHN POPPELAARS

Tot een paar jaar geleden was de term Big Data slechts bekend bij een beperkte groep mensen. Vandaag de dag kun je geen website bezoeken, geen artikel meer lezen zonder dat Big Data hun opwachting maken. Een Google query levert met de zoekterm Big Data in minder dan een halve seconde 1,5 miljard hits op; nog geen twee jaar terug lag het op slechts een vijfde van dat aantal. Hits met referenties naar toegenomen bezoekersaantallen van een museum of naar de prestatieverbetering van de *supply chain* van Tesco. Je krijgt de indruk dat Big Data overal zijn.

Big Data worden, al dan niet ingegeven door commerciële motieven, neergezet als het antwoord op alles. Het is het einde 'the end of theory' zoals voormalig *Wired*-redacteur Chris Anderson (2007) ons wil doen geloven. Big Data lijken te impliceren dat in de omvang van de data een soort van magie schuilt. Als de hoeveelheid gegevens maar groot genoeg is, komt vanzelf het juiste antwoord boven alsof Apollo niet langer huist in Delphi maar in grote datasets. Is met Big Data de Deep Thought (Adams, 1979) werkelijkheid geworden en gaan Big Data ons het antwoord geven op alle vragen inclusief het antwoord op de ultieme vraag over het Leven, het Universum en Alles? Zijn Big Data het einde van de besliskunde?

Uitdaging

De term Big Data is aan het einde van de vorige eeuw geïntroduceerd. NASA onderzoekers Michael Cox en David Ellsworth gebruiken in 1997 als eerste de term. Big Data worden door hen gebruikt om gegevenssets aan te duiden die zo groot zijn dat ze niet langer in het computergeheugen of op opslagmedia kunnen worden opgeslagen. Tegenwoordig is er eigenlijk geen opslagprobleem meer, wat overblijft is de analyse-uitdaging. Inmiddels worden Big Data niet alleen door de omvang van de

dataset gedefinieerd maar, zoals door onderzoeks- en consultancybureau Gartner (2011) geïntroduceerd, met behulp van de 3 V's; *Volume*, *Variety* en *Velocity*. *Volume* spreekt voor zich. Door de introductie van IT-systemen zoals Enterprise Resource Planning systemen (ERP) wordt steeds meer over bedrijfsprocessen vastgelegd. Dat volume belangrijk is, blijkt wel uit de hoeveelheid gegevens die we met zijn allen genereren. IBM becijfert dat we elke dag opnieuw 2,5 quintiljoen (10^{30}) bytes aan data genereren en dat groeit elke dag. Zo snel dat 90% van de data waar we nu over beschikken in de afgelopen 2 jaar is ontstaan.

De tweede V is die van *Variety*, ofwel verscheidenheid van formats en bronnen. Naast de gestructureerde data in (ERP) databases kunnen beslissers beschikken over steeds meer ongestructureerde databronnen als sociale media, video, audio, sensor- en meterdata, foto's, financiële transacties, etc. De uitdaging voor beslissers is om uit deze verschillende bronnen, gestructureerd en ongestructureerd, inzichten te halen en daar naar te handelen.

De laatste V is die van *Velocity*, ofwel snelheid. Snelheid waarmee de data ontstaan, maar ook de snelheid waarmee de data moeten worden verwerkt om te kunnen worden geraadpleegd of gebruikt in besluitvorming.

Dat het kunnen analyseren van grote verzamelingen gegevens grote voordelen kan opleveren, blijkt uit de succesverhalen van bedrijven als Google en Amazon. De omzet van Amazon blijft maar groeien, *key driver* van deze omzetgroei is de inzet van op het individu afgestemde aanbiedingen van boeken en andere producten. Amazon beseft als geen ander dat het analyseren van verkoopgegevens en de daaruit voortvloeiende inzichten zeer bruikbaar zijn in het verder opstuwende van verkoopaantallen en daarmee de omzet. Daarbij gaat het Amazon niet om het verklaren van waarom mensen uiteindelijk tot koop van een product overgaan, het feit dat er een correlatie bestaat tussen

twee aankopen is genoeg om het product in het *Customers Who Bought This Item Also Bought*-lijstje op te nemen. Dit principe wordt ook door Google vol benut. Google analyseert gegevens niet om vast te stellen waarom vaak naar een website wordt verwezen, puur het aantal referenties is voldoende om door Google's PageRank algoritme hoog te worden gewaardeerd.

Verwachting

Door deze succesverhalen zijn de verwachtingen van wat met Big Data kan worden bereikt tot grote hoogte gestegen. Dit wordt verder verstrekt door tot de verbeelding sprekende voorbeelden. Hoe kon de supermarktketen Target immers eerder dan de vader weten dat zijn tienerdochter zwanger is (Forbes, 2012)? Door grote hoeveelheden kassabonnetjes in combinatie met klantkaartgegevens te analyseren had Andrew Pole, de statisticus van Target, een patroon kunnen vaststellen in de aankoop van zwangerschapsgerelateerde producten. Tijdens zijn analyse had hij een 25-tal producten geïdentificeerd die, als de verkoopaantallen ervan gezamenlijk werden geanalyseerd, een zwangerschapsvoorspellende waarde hadden (zoals cacaoboterolie, supplementen, een luiertas en een blauw matje). Hij kon zelfs het verwachte geboortetijdspunt schatten. Op basis van dit inzicht kon Target gericht aanbiedingen gaan versturen. Een Big Data succesverhaal of is het een *lucky hit*, door de media opgepikt en uitvergroot? Het is wellicht beangstigend om vast te stellen dat retailers kunnen weten wanneer je zwanger bent op basis van je aankopen (hoewel, vroeger wist de kruidenier op de hoek dit natuurlijk ook meteen). Ligt de privacy op straat? Iets waar menig journalist een sappig verhaal over wil schrijven, zo ook het Target verhaal? We moeten echter oppassen niet het slachtoffer te zijn van de *observation bias*.¹ We weten immers niet hoe vaak Target ten onrechte de aanbieding voor babyartikelen aan een klant heeft gestuurd. Hechten we niet teveel waarde aan de Big Data analyse van Target en is het succes toch meer geluk dan wijsheid?

Een waarschuwing is daarom op zijn plaats. Data zijn immers niet objectief en correlaties impliceren vaker niet dan wel een causaal verband. Data komen tot stand doordat wij ze verzamelen, en krijgen betekenis doordat we ze analyseren. Iedere besliskundige weet dat bij beide stappen vele gevaren op de loer liggen. Een voorbeeld zijn de opvallende resultaten van analyse van Twitter en Foursquare data uit de periode dat de wervelstorm Sandy over New York en omliggende gebieden trok. Uit analyse blijkt dat het leven de nacht na de storm alweer

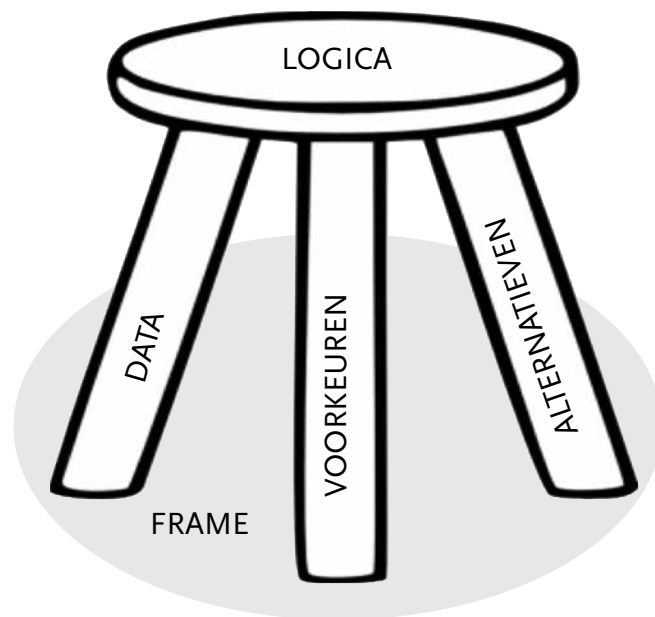
werd opgepakt, kennelijk viel de impact van Sandy mee (Grinberg, 2013). Een onterechte conclusie. Omdat in Manhattan het smart-phone-bezit veel groter is dan in de omliggende gebieden, zoals Coney Island, wekte het grote aantal tweets uit Manhattan de indruk dat daar het zwaartepunt van de ramp lag. Niets was minder waar. In dunbevolkte gebieden zijn minder smart phones aanwezig en als gevolg daarvan ook minder tweets. Na verloop van tijd namen de tweets uit de buitengebieden af, niet omdat er minder te melden was maar doordat de smart phones niet meer konden worden opgeladen vanwege black outs.

Iets vergelijkbaars trad op bij de voorspellingen die Google over de omvang van de griep epidemie in de US begin dit jaar deed (Butler, 2013). Google schatte in dat ongeveer 11% van de bevolking geïnfecteerd zou worden, bijna het dubbele van wat het Centre For Disease Control and Prevention schatte. De oorzaak van de overschatting ligt waarschijnlijk in de uitgebreide media-aandacht voor de griep dat tot veel zoekopdrachten bij Google leidde, de basis voor de trendanalyse van Google. Veel data zijn kennelijk niet altijd alleenzalmakend.

Beslissen

Beslissingen gebaseerd op data alleen kunnen tot voordelen leiden als je klant van Target bent, echter ook nadelen hebben als je op Coney Island woont. Kan een minister verantwoordelijk voor de inkoop van griepvaccinaties, het beste Google Trends in de gaten houden of de voorspellingen van het RIVM? Zou u zich veilig voelen als u wist dat de medicijnen die u neemt gebaseerd zijn op een uit Big Data gedestilleerde correlatie? Kwalitatief hoogwaardige beslissingen zijn niet alleen op data gebaseerd, laat staan dat ze voortkomen uit het lukraak zoeken naar mogelijke verbanden in grote gegevens sets. Data zijn echter wel een hele belangrijke grondstof voor het nemen van beslissingen. Zoals Ron Howard (2012) aangeeft, ligt de basis voor kwalitatief hoogwaardige besluitvorming bij een goede afbakening, *framing*. De beslissing die we uiteindelijk nemen wordt ondersteund door:

1. de mogelijke acties die we kunnen nemen (Alternatieven)
 2. de waarderingscriteria van de mogelijke uitkomsten (Voorkeuren)
 3. de informatie die tot onze beschikking staat (Data)
- Het zijn de drie poten van een kruik waarvan de zitting uit logica bestaat, het wiskundige model. Iedere kwalitatief



goede beslissing bestaat uit ieder van deze elementen. Als er geen alternatieven zijn is beslist immers eenvoudig. Als er geen voorkeuren zijn is er geen onderscheid te maken tussen de alternatieven, ze zijn allemaal even goed. Als je niet over informatie beschikt om de invloed van je beslissing te meten zijn alle mogelijke alternatieven even zinvol. Andersom geldt ook, informatie die niet bijdraagt aan het beoordelen van alternatieven is zinloos om te verzamelen.

Organisaties zouden zich dus niet blind moeten staren op data. Het verschil zal worden gemaakt door die organisaties die data weten te combineren met een gestructureerde aanpak voor het nemen van beslissingen. Beslissingen die de prestaties van de organisatie verbeteren, bijvoorbeeld de betere benutting van productiemiddelen of het verbeteren van de servicegraad. Organisaties doen er verstandig aan niet alleen te investeren in data-verzameling, maar vooral de analytische vaardigheden en beslismodellen van de organisatie te verbeteren. De kracht zit in de combinatie. De resultaten van de jaarlijks terugkerende Franz Edelman competitie van INFORMS laten dat duidelijk zien.² Zo bespaarde TNT Express meer dan €200 miljoen, verbeterde haar service en reduceerde de CO² emissies door de combinatie van data en optimalisatie methoden. Analooeg bespaart Coca Cola jaarlijks \$45 miljoen door de effectiviteitsverbetering van haar distributienetwerken en rationaliseerde HP haar productportfolio en genereerde \$500 miljoen meer omzet.

Als de toegevoegde waarde van alle projecten uit de Franz Edelman competitie sinds 1971 worden opgeteld resulteert het astronomische bedrag van \$180 miljard. Zonder data was dit niet mogelijk geweest, de meerwaarde wordt echter vooral bereikt door de gestructureerde en modelmatige manier waarop de vraagstukken zijn aangepakt.

Om het antwoord op de ultieme vraag te geven rekent Deep Thought, de computer uit de *Hitchhikers Guide to the Galaxy*, 7,5 miljoen jaar om tot de slotsom te

komen dat het antwoord 42 is. Niemand weet wat men aan moet met dit antwoord, aangezien men niet weet wat de ultieme vraag nu eigenlijk is. Om de vraag bij het antwoord 42 te vinden wordt een nieuwe computer gebouwd. Deze computer rekt 10 miljoen jaar, maar wordt vijf minuten voordat de berekeningen zijn afgerond vernietigd. Een besliskundige had hier veel tijd bespaard. Allereerst door het bepalen welke vraag nu precies moet worden beantwoord, vervolgens de relevante data te verzamelen en te valideren, een model te ontwerpen en te ijken om tenslotte met het model en de data het best mogelijke antwoord te bepalen. Ik weet nu al dat het geen 42 is of Big Data zijn, maar de besliskunde.

NOTEN

1. De situatie waarin alleen de successen worden verteld, de mislukkingen worden verzwegen.
2. Zie de website van Informs <www.informs.org/Recognize-Excellence/Franz-Edelman-Award>

LITERATUUR

- Adams, D. (1979). *The Hitchhiker's guide to the Galaxy*. Pan Books.
- Anderson, C. (2007). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, 16(7), <www.wired.com/science/discoveries/magazine/16-07/pb_theory> geraadpleegd 23-6-2013
- Butler, D. (2013). When Google got flu wrong: US outbreak foxes a leading web-based method for tracking seasonal flu. *Nature* 494, 155–156, doi:10.1038/494155a, <www.nature.com/news/when-google-got-flu-wrong-1.12413> geraadpleegd 29-6-2013
- Cox, M. & Ellsworth, D. (1997). *Application-Controlled Demand Paging for Out-of-Core Visualization* (Report NAS-97-010). NASA Ames Research Center, <www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf> geraadpleegd 23-6-2013.
- Forbes (2012). *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did*. <www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/> geraadpleegd op 29-6-2013
- Gartner (2011). *Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data*. Stamford: Gartner, Inc. <www.gartner.com/newsroom/id/1731916> geraadpleegd op 23-6-2013
- Grinberg, N., Naaman, M., Shaw, B. & Lotan, G. (2013). *Extracting Diurnal Patterns of RealWorld Activity from Social Media*. Palo Alto, California: Association for the Advancement of Artificial Intelligence <<http://sm.rutgers.edu/pubs/Grinberg-SMPatterns-ICWSM2013.pdf>> geraadpleegd 29-6-2013
- Howard, R. (2012). *The Foundations of Decision Analysis Revisited*. Los Angeles: USC. <www.usc.edu/dept/create/assets/001/50843.pdf>, geraadpleegd 30-6-2013
- IBM (z.j.). *What is big data*. <www-01.ibm.com/software/data/bigdata/> geraadpleegd op 23-6-2013

JOHN POPPELAARS is voorzitter van het Nederlands Genootschap Besliskunde en Directeur Consulting bij ORTEC Consulting Groep. E-mail: <john.poppelaars@ortec.com>



PUPPY'S, INJECTIENAALDEN EN ATTITUDEVORMING EEN EYE-TRACKING STUDIE

ROXANNE I. VAN GIESEN

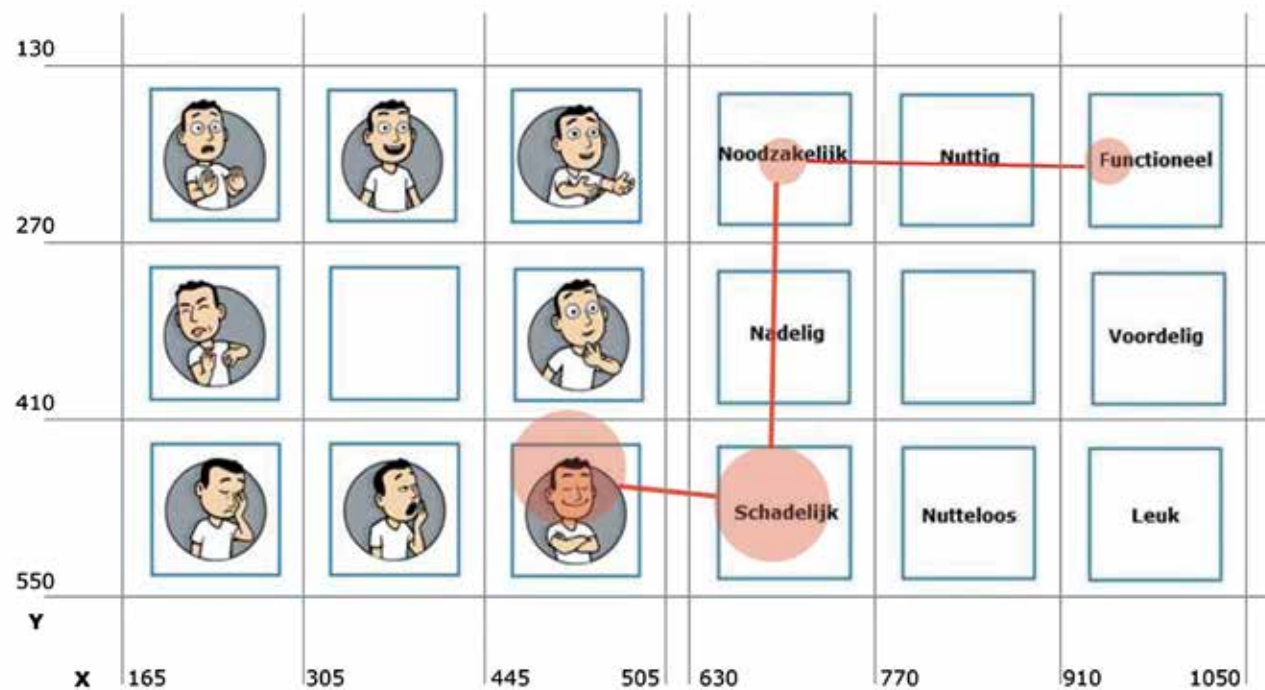
Dagelijks vormen we attitudes (meningen) over allerlei dingen zoals over politieke kwesties, auto's of huisdieren. Attitudes vormen we gedeeltelijk op basis van gevoel (affectief) of door er over na te denken (cognitief). Het vormen van zo'n attitude vraagt mentale inspanning, afhankelijk van het type attitude-object en het soort attitude dat gevormd wordt (affectief of cognitief). Een manier om de mate van mentale verwerking tijdens de attitudevorming te onderzoeken is met behulp van *eye-tracking*. Dit is een psycho-fysiologische meetmethode die het mogelijk maakt om de onverstoordde verwerkingsprocessen te onderzoeken, en de mentale verwerking zelf niet beïnvloedt.

Puppy's versus injectienaalden

Voor sommige attitude-objecten is het gemakkelijk om een attitude te vormen. Er is weinig bewuste inspanning, en dus minder mentale verwerking, voor nodig. Neem

nou de puppy als voorbeeld. Mensen vinden een puppy onmiddellijk 'leuk' of ze voelen zich 'blij' bij het zien van een puppy. Voor eenduidige attitude-objecten, zoals de puppy, worden stabiele attitudes geactiveerd.

Bij meer lastiger attitude-objecten, zoals neutrale en ambivalente attitude-objecten, werkt de mentale verwerking anders. Wanneer het gaat om neutrale attitude-objecten zijn mensen vaak onverschillig. Dat betekent dat als ze gedwongen worden hier een attitude over te vormen ze ook gedwongen worden argumenten te verzamelen. Dit vraagt om meer mentale verwerking. Bij ambivalente attitude-objecten wordt inconsistente informatie geactiveerd, waardoor de evaluatie tegelijkertijd positief en negatief kan zijn. Denk hierbij aan injectienaalden waarbij het zowel over ziekte als genezing gaat. Attitudes ten opzichte van ambivalente attitude-objecten moeten 'on the spot' gecreëerd worden waardoor er diepere mentale verwerking plaatsvindt, want relevante informatie moet onderzocht en geïntegreerd worden tot een eindoordeel.



Figuur 1. Deze schaal werd gebruikt om te bepalen of affect of cognitie doorslaggevend was in de attitudevorming, hierbij mocht een van de 16 hokjes aangeklikt worden. De linkerkant geeft de affectieve schaal weer (gebaseerd op Desmet, 2003), de rechterkant geeft de cognitieve schaal weer (gebaseerd op Crites, Fabrigar, & Petty, 1994). De affectieve, cognitieve, en gecombineerde schalen werden afzonderlijk van elkaar beantwoord.

Affect en cognitie

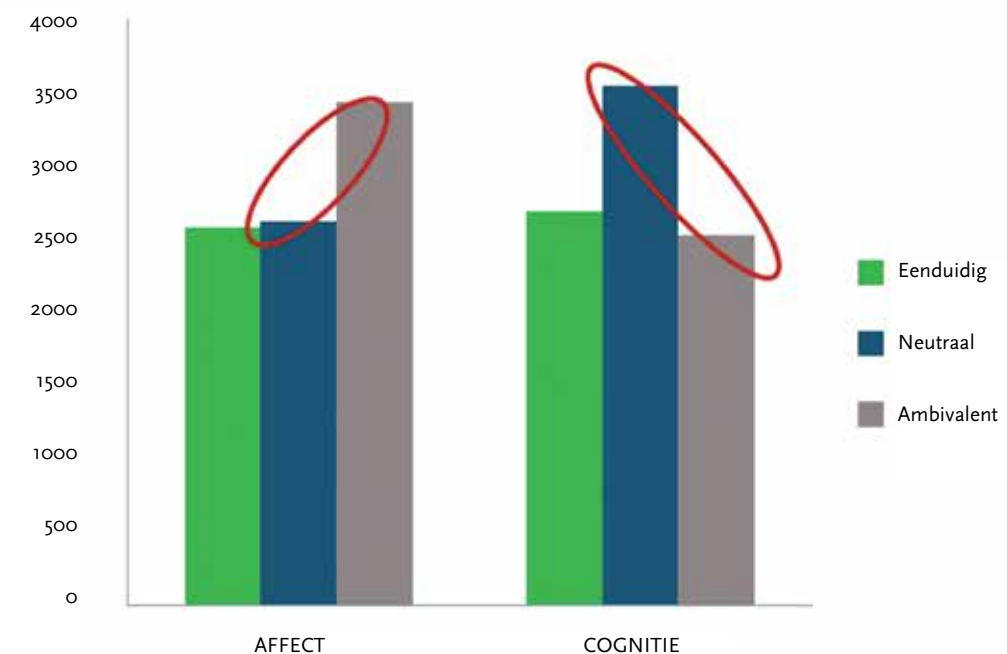
Eenduidige attitude-objecten zullen een vergelijkbare cognitieve en affectieve attitude activeren. Neutrale of ambivalente attitude-objecten zullen niet altijd vergelijkbare cognitieve en affectieve attitudes activeren. Vaak wordt verondersteld dat het minder mentale inspanning vraagt om de affectieve attitude te vormen (Slovic, Finucane, Peters, & MacGregor, 2004). De vraag is daarom of het evenveel mentale moeite kost om de affectieve en cognitieve attitude te vormen voor verschillende attitude-objecten en of mensen zich in hun attitudevorming dan vooral laten leiden door cognitie of affect. Met behulp van eye-tracking data kan inzicht verkregen worden in de onderliggende processen tijdens attitudevorming, en wordt meer informatie verkregen dan via self-report data.

Methode

Met een eye-tracker werd data verzameld bij 21 personen, terwijl ze een affectieve en cognitieve attitude vormden. Ze kregen 15 attitude-objecten te zien, die verdeeld waren in 4 stimulus-categorieën: 3x positief (bv. Puppy),

3x negatief (bv. Ramp), 3x neutraal (bv. Natuur) en 6x ambivalent (bv. Injectienaald). Alle attitude-objecten werden 4 seconden getoond. Per attitude-object werd vervolgens een affectief en cognitief oordeel gevraagd, waarbij respondenten op de antwoordoptie klikten die hun attitude het beste weerspiegelde. Om te bepalen welke attitudecomponent doorslaggevend was voor de eindattitude moesten respondenten één enkele keuze maken uit alle affectieve en cognitieve antwoordopties (zie figuur 1).

Oogbewegingen werden geregistreerd gedurende het hele onderzoek, dus van zowel het bekijken van alle attitude-objecten als het beantwoorden van de vragen. Echter, alleen de oogbewegingen op de antwoordopties waren relevant voor het beantwoorden van de onderzoeksvragen. Deze data dienden daarom 'uitgeknipt' te worden. Per respondent werd er gemiddeld zo'n 12 minuten aan oogbewegingsdata verzameld, waarbij de eye-tracker iedere 50 ms. oogbewegingen registreerde. Er waren tussen de 1200 en 1500 datapunten (fixaties) per respondent (totaal zo'n 28000 fixaties). Door het uitknippen van de relevante data werd het aantal datapunten teruggebracht tot gemiddeld 380 per respondent. Bij 4 personen zijn de oogbewegingen onjuist geregistreerd, waardoor uiteindelijk de oogbewegingsdata van 17 personen



Figuur 2. Totale fixatietijd nodig voor het beantwoorden van de schaal (in ms). De resultaten van het aantal fixaties laten een vergelijkbaar patroon zien.

geanalyseerd is. Soms klikten mensen dubbel op een antwoordschaal, en omdat dat het totaal aantal fixaties beïnvloedt zijn deze er uitgelaten. Uiteindelijk zijn de analyses uitgevoerd op 6451 fixaties. Eye-tracking levert dus een grote hoeveelheid data op, waarbij binnen het huidige onderzoek maar een klein gedeelte van de data verder geanalyseerd is.

In figuur 1 is een voorbeeld weergegeven van een mogelijk oogbewegingspatroon. De rondjes staan voor de fixaties. De grotere rondjes geven een langere fixatieduur aan. Om deze data te analyseren werden de x- en y-coördinaten van de fixaties van de oogbewegingen omgerekend naar de verschillende antwoordopties. Aan de hand hiervan werd uitgerekend hoeveel fixaties er op de schaal waren geweest, per attitude-object, per antwoordschaal. Ook werd de gemiddelde fixatieduur berekend en werd er bekeken hoe vaak men naar een bepaalde antwoordoptie keek. Zo zien we in het voorbeeld in figuur 1 dat er 4 fixaties op 4 verschillende antwoordopties geweest zijn. Meer fixaties (meerdere antwoordopties afwegen) en langere fixatieduur (diepere verwerking) worden gebruikt als indicatoren voor meer mentale verwerking. Bijvoorbeeld, bij puppy's focust men voornamelijk op de blij smiley of de antwoordoptie 'leuk', hier zijn weinig fixaties voor nodig.

Resultaten en conclusie

Een *general linear mixed model* analyse is uitgevoerd om de invloed van de stimuluscategorie op het aantal fixaties en de totale fixatietijd te onderzoeken, waarbij het attitude-object opgenomen was als een herhaalde variabele in het model.

Eenduidige, positieve en negatieve, attitude-objecten zijn gelijk qua benodigde mentale verwerking. Neutrale en ambivalente attitude-objecten vergen ongelijke mentale verwerking. Voor de affectieve schaal werd een verschil tussen eenduidige en ambivalente attitude-objecten gevonden, waarbij ambivalente attitude-objecten om meer mentale verwerking vragen, zoals blijkt uit meer fixaties en een langere totale fixatietijd (zie figuur 2). Voor de cognitieve schaal werd een verschil gevonden tussen eenduidige en neutrale attitude-objecten, waarbij neutrale stimuli om meer mentale verwerking vragen.

Verdere vergelijking tussen ambivalente en neutrale attitude-objecten laat zien dat bij ambivalente attitude-objecten in totaal langer en vaker wordt gefixeerd op de affectieve antwoordschaal, vergeleken met neutrale attitude-objecten. Bij neutrale attitude-objecten wordt daarentegen een hogere totale fixatieduur en meer fixaties gevonden voor de cognitieve antwoordschaal, vergeleken met ambivalente attitude-objecten. Voor ambiva-

CATEGORIE	EINDATTITUDE			
	affectief	cognitief	range percentage affect t.o.v. cognitie gekozen	mode percentage affect t.o.v. cognitie gekozen
Eenduidig	59,30%	40,70%	20-100%	67%
Neutraal	65,11%	34,89%	0-100%	100%
Ambivalent	29,20%	70,80%	0-100%	100%

Figuur 3. Doorslaggevende component voor de eindattitude (affectief/cognitief).

lente attitude-objecten is dus meer mentale verwerking nodig voor de affectieve schaal, terwijl dit voor neutrale attitude-objecten de cognitieve schaal betreft.

De self-report data zijn gebruikt om te onderzoeken welke attitudecomponent bepalend was voor het eindoordeel. Een logistische regressie laat zien dat voor eenduidige attitude-objecten de affectieve attitude meestal de doorslag geeft. De antwoordschaal die de minste verwerking vergt heeft de voorkeur. Zo is het bij ambivalente attitude-objecten mentaal inspannender om een affectief oordeel te vormen en geeft de cognitieve attitude dus meestal de doorslag. Voor neutrale attitude-objecten is het mentaal inspannender om een cognitief oordeel te vormen en geeft de affectieve attitude dus meestal de doorslag. De percentages in figuur 3 illustreren dit.

Ondanks de kleine steekproef worden er systematische verschillen tussen de attitude-categorieën gevonden en wordt aangetoond dat het niet altijd gemakkelijker is en minder mentale inspanning vergt om een affectief oordeel te vormen. Kortom, de eye-trackerdata laten zien dat je over puppy's en rampen minder hoeft na te denken. Injectienaalden en waterdruppels vragen daarentegen om uitgebreidere mentale verwerking, hetzij affectief hetzij cognitief.

Deze vorm van Big Data is veelbelovend en verschaft het sociaalwetenschappelijk onderzoek nieuwe mogelijkheden en uitdagingen. Eye-tracker onderzoek levert een grote hoeveelheid aan data op, waarmee een belangrijke

bijdrage geleverd kan worden aan het verder begrijpen van psychologische processen, zoals hoe intuïtieve en meer doordachte processen zich tot elkaar verhouden. Maar het eye-tracker onderzoek roept ook vragen op over het knippen en analyseren (multiple testen!) van deze data.

Het beschreven onderzoek is samen met Arnout Fischer, Heleen van Dijk en Hans van Trijp uitgevoerd. Het project is gefinancierd door NanoNextNL.

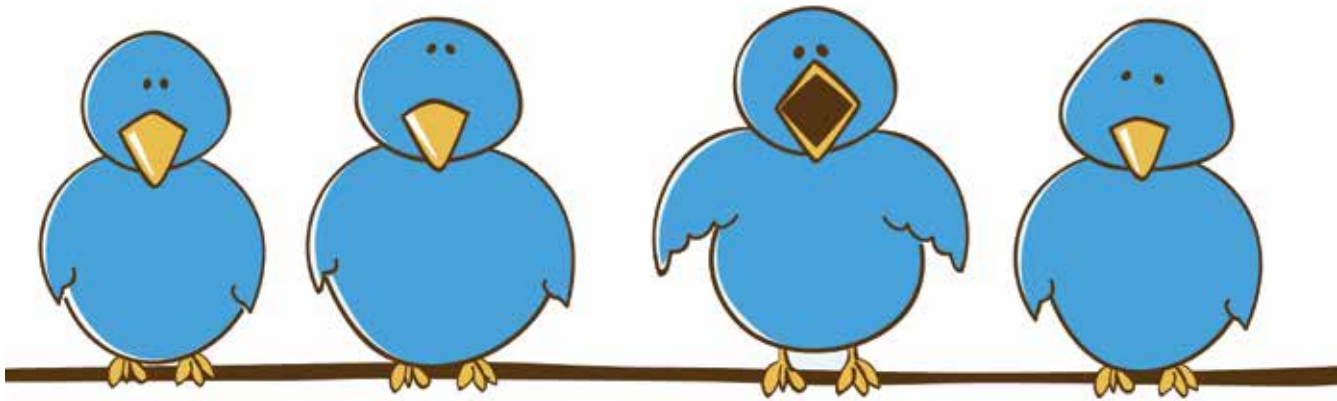
LITERATUUR

Crites, S. L., Fabrigar, L. R., & Petty, R. E. (1994). Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues. *Personality and Social Psychology Bulletin*, 20, 619-634. doi: 10.1177/0146167294206001

Desmet, P. M. A. (2003). Measuring emotion; development and application of an instrument to measure emotional responses to products. In A. F. M. M.A. Blythe, K. Overbeeke, & P.C. Wright (Eds.), *Funology: from Usability to Enjoyment* (pp. 111-123). Dordrecht: Kluwer Academic Publishers.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis*, 24, 311-322. doi: 10.1111/j.0272-4332.2004.00433.x

ROXANNE VAN GIESEN is promovenda bij de leerstoelgroep Marktkunde en Consumentengedrag van de Wageningen Universiteit.
E-mail: <Roxanne.vanGiesen@wur.nl>



BIG DATA EN OFFICIËLE STATISTIEK

De relevantie van heel veel tweets

PIET DAAS

Het is een term die we steeds vaker horen: Big Data. Laat ik maar vanaf het begin duidelijk zijn, ik ben een fan van Big Data en dan vooral van het Big Data-onderzoek dat kijkt naar de bruikbaarheid van data die door allerlei systemen en sensoren om ons heen verzameld wordt. Waarom ik fan ben? Om twee redenen. Ten eerste omdat dergelijke data het gevolg zijn van bepaald gedrag van personen wat erg interessante input is voor allerlei studies. Voorbeelden van Big Data-bronnen zijn call-detail records van mobiele telefoon providers (wanneer belt wie met wie, op welke plaats en hoe lang?), sociale mediaberichten (welke mening ventileert iemand op welk moment en op welke plaats), en gegevens die door sensoren, zoals verkeerslussen, worden verzameld (hoeveel voertuigen bevinden zich op een bepaald tijdstip op een bepaalde locatie). Dergelijke gegevens worden in miljoenen hoeveelheden per dag gegenereerd en bewaard.

Wanneer je toegang tot die data mag hebben, en dat is zowel juridisch als technisch al een hele klus, is het een uitdaging om uit die berg data nuttige informatie te halen. Dit laatste is mijn tweede reden om fan van Big Data te zijn. Om dergelijke enorme hoeveelheden data te analyseren moet je niet alleen handig met computers

zijn maar moet je ook creatief, grondig en vasthoudend te werk gaan. Zeker in het begin heb je dat nodig omdat de hoeveelheid data vaak zo groot is dat het de huidige werkwijze en omgeving van de gebruiker te boven gaat. Big is dan ook een relatief begrip. Zo vindt een klimaatonderzoeker alles boven de 5TB veel, terwijl een andere onderzoeker met 5GB aan data al een probleem heeft. Beide hebben ze een Big Data-probleem, bij de één is het alleen wat groter.

Analysen

Moet je al die data wel analyseren dan? Ehm, bij voorkeur wel. Big Data zijn namelijk niet zomaar op te vatten als data die bijvoorbeeld met behulp van een vragenlijst zijn verkregen. In het laatste geval is van te voren nagedacht over wat gevraagd wordt en hoe dit wordt verzameld. Hierdoor is de informatiedichtheid in met vragenlijst verkregen gegevens ook erg hoog. Big Data bestaan daarentegen meestal uit heel veel records van een zeer beperkte set variabelen, die niet door de onderzoeker zijn gedefinieerd. Zo bestaan de gegevens die verkeerslussen verzamelen uit niet meer dan het aantal

voertuigen per minuut op een bepaalde locatie en op een bepaald tijdstip voor drie verschillende lengtecategorieën. Om uit dat soort bronnen nuttige informatie te halen geldt: hoe meer data, hoe groter de kans op het vinden van een interessant 'signaal'.

Voor de analyse van Big Data heb je bij voorkeur veel rekenkracht nodig. Ik ga hier niet over uitweiden, maar hoe meer rekenkracht je tot je beschikking hebt hoe sneller je de data kunt doorspitten en hoe eerder je iets interessants kunt vinden. Meer is in dit geval altijd beter, dat wil zeggen meer geheugen, meer processoren, en meer computers in je cluster. Overigens kun je met enige creativiteit en een gewone huis, tuin en keuken desktop (met wat extra geheugen) en wat geduld al heel veel doen. Ook is het handig als je de programmeertaal R beheerst. En anders kun je, zolang je geen privacygevoelige data analyseert, altijd nog rekenkracht via internet inkopen; bijvoorbeeld bij die Amerikaans online boekhandel.

In het begin ga je vooral verkennend te werk. Je gaat dan bijvoorbeeld plaatjes maken van de veranderingen in de waarden van variabelen over de tijd. Door de grote hoeveelheid data kan dit soms tot figuren leiden die zoveel datapunten bevatten dat deze ogen als één hele grote zwarte vlek. Hier zijn oplossingen voor zoals heatmaps, die de dichtheid van de meetpunten in verschillende kleuren weergeven, of interactieve figuren waarbij je de transparantie van de punten kunt variëren. Visualisatiemethoden zijn en worden steeds belangrijker bij de analyse van Big Data. Ze zijn in het begin vooral nuttig voor het 'zien' van verbanden en later voor het interpreteren van de bevindingen. Gedurende de verkennende studie zul je ook merken dat er best veel fluctuatie in de data ('ruis') aanwezig is. Mogelijk moet je de ruis (deels) wegfilteren maar deze kan ook door aggregeren al verdwijnen. De verkennende studie moet uiteindelijk leiden tot een aanpak waarbij je een bepaalde set van gegevens op een bepaalde wijze aggregereert. Hierdoor worden Big Data een stuk kleiner (*small data*) en kunnen ze in een standaardomgeving, met de bekende tools en methoden, verder geanalyseerd worden. Dit klinkt allemaal heel makkelijk maar dat is niet altijd zo. Je kunt namelijk makkelijk verdrinken in een zee van data. Ik denk dat een concreet praktijkvoorbeeld alles wat duidelijk maakt.

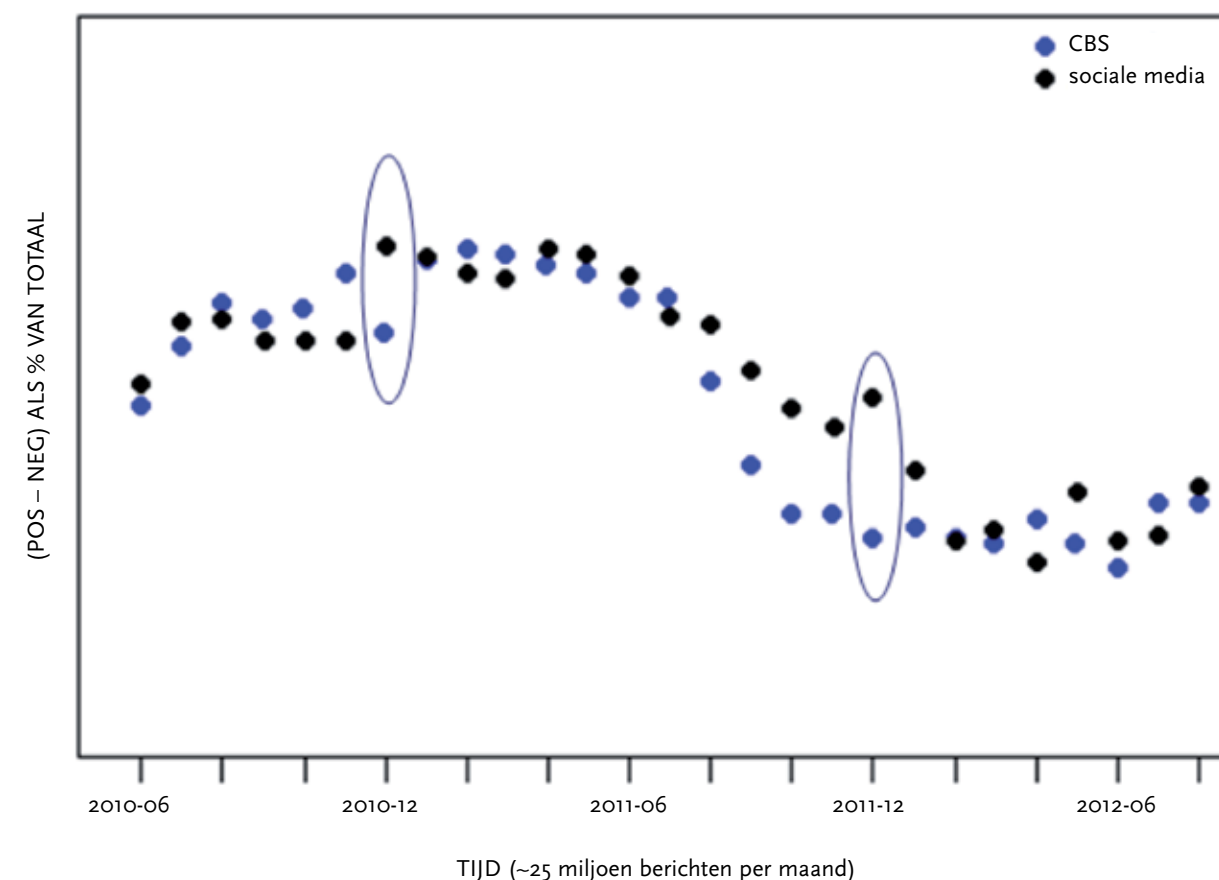
Twitterberichten

We hebben op het CBS redelijk wat tijd gestoken in het onderzoek naar de bruikbaarheid van socialemediagege-

vens voor de officiële statistiek. En, voor alle duidelijkheid, hierbij kijken we uitsluitend naar socialemediaberichten die door gebruikers voor iedereen leesbaar op internet geplaatst worden. Wat dat betreft zijn het dus publiekelijk beschikbare data, iedereen kan eraan komen; mits je enige moeite doet. Als eerste verkennende stap hebben we gekeken welke onderwerpen in dergelijke berichten worden besproken. Hierbij werd een 'selectie' van 12 miljoen Twitterberichten, op dat moment het populairste sociale medium in Nederland, geclassificeerd naar de thema's waarover het CBS publiceert. Dit bleek een behoorlijke bewerkelijke klus. De uiteindelijke resultaten van de tekstuele analyse van deze berichten liet zien dat een kleine 50% van die berichten niet over voor het CBS relevante onderwerpen ging; ze kwamen in de overige categorie terecht. Erger nog, ze gingen vrijwel helemaal nergens over. De andere helft van de berichten was echter wel relevant; deze konden namelijk aan één van de 16 thema's waarover het CBS publiceert worden toegekend.

Dat lijkt gunstig maar het grote probleem is dat door de brede set van woorden die in de niet-relevante berichten gebruikt worden en de beperkte omvang van Twitterberichten (maximaal 140 karakters), het lastig is woorden te vinden die specifiek genoeg zijn om de voor ons interessante berichten te selecteren. Dergelijke woorden zijn of zo specifiek dat erg weinig berichten gevonden worden, en dan heb je nauwelijks iets om te analyseren, of leveren wel redelijk wat berichten op maar dan blijken die aantallen weer heel erg over de tijd te fluctueren. Mede hierdoor zijn we erop gekomen socialemediaberichten slechts zeer beperkt te filteren. Dit heeft als voordeel dat de aantallen berichten gedurende de tijd redelijk constant blijven maar het zijn er dan wel veel; we kwamen uiteindelijk op zo'n miljoen berichten per dag. Die hebben we dan ook niet meer zelf verzameld, er zijn gelukkig firma's die dat voor je doen. Je kunt gewoon toegang tot dergelijke datasets, met uitsluitend publiekelijk beschikbare socialemediaberichten, kopen. Dit bleek een hele goede zet omdat we hierdoor ook een meer geschikte analyseomgeving tot onze beschikking hadden. Dit maakte het mogelijk niet alleen enorme hoeveelheden socialemediagegevens te selecteren en te analyseren, maar we konden ook het sentiment van die berichten bepalen.

Dat laatste leverde iets zeer interessants op. De verhouding positieve en negatieve berichten in sociale media blijkt namelijk over de tijd te veranderen. Wanneer we deze verandering, geaggregeerd op maandniveau (en dan hebben we het over zo'n 30 miljoen berichten per



Figuur 1. Ontwikkeling van het sentiment in sociale media berichten en het consumentenvertrouwen. De omcirkelde gebieden geven de grote positieve afwijking rond de kerstperiode aan door de vele 'fijne feestdagen' en 'beste wensen' berichten.

maand), vergelijken met het consumentenvertrouwen van het CBS (een maandcijfer) bleek dat deze reeksen een vergelijkbare ontwikkeling te vertonen. Dit is in figuur 1 te zien. De correlatiecoëfficiënt van beide reeksen was 0,88. Overigens, ik weet dat een hoge correlatie niet wil zeggen dat er een gemeenschappelijke oorzaak is (hoe vaak ik die opmerking heb gehoord wil je niet weten) maar een dergelijke observatie geeft wel aan dat je mogelijk iets belangwekkends gevonden hebt. En dat is de reden om Big Data te bestuderen, en een hoge correlatie is een beginpunt en beter dan helemaal geen resultaat.

Nadere studies ondersteunen de bevinding dat beide reeksen dezelfde ontwikkeling in de tijd vertonen. Dit suggereert dat de verhouding van positieve en negatieve woorden in socialemediaberichten, die het sentiment van een bericht bepalen, en de antwoorden op de vragenlijst van het CBS over het consumentenvertrouwen iets soortgelijks meten (ik druk me hier heel voorzichtig

uit). Meten we hier het 'onderbuikgevoel' van de Nederlanders? Nauwkeurigere vervolganalyses neigen ernaar dit idee te bevestigen maar ik kan op dit moment – helaas – nog geen definitief antwoord geven. De tijd zal leren wat hier verder uitkomt.

Het voorbeeld laat zien dat je door het analyseren van Big Data op onverwachte zaken kunt stuiten. Het heeft ons in ieder geval doen besluiten om, naast een verdere analyse van sociale mediaberichten, ook te gaan werken aan de ontwikkeling van een sentimentindicator die, gebaseerd op socialemediaberichten, het sentiment per week bepaalt. Omdat je deze op de eerste werkdag van de week kunt bepalen, is dat een erg snel beschikbaar cijfer en mogelijk erg interessant.

PIET J. H. DAAS is als methodoloog en datawetenschapper werkzaam bij het CBS.
E-mail: <pjh.daas@cbs.nl>



Eenvoudige modellen en ‘Grote Gegevens’ troeven slimme modellen af

DJOERD HIEMSTRA

Big Data – of het beter allitererende Grote Gegevens – is een term die sinds het begin van deze eeuw wordt gebruikt om gegevensverzamelingen aan te duiden die moeilijk verwerkt konden worden met behulp van de software van die tijd, verzamelingen met een grootte van vele terabytes of petabytes. Technieken om zulke enorme verzamelingen gegevens te kunnen verwerken en analyseren werden met name ontwikkeld door Google. Het uitgangspunt van Google is: zet heel veel goedkope machines bij elkaar in grote datacentra, en gebruik slimme gereedschappen zodat applicatieontwikkelaars en gegevensanalisten het hele datacentrum kunnen gebruiken voor hun gegevensanalyses. Het datacentrum is de nieuwe computer! De slimme gereedschappen van Google raken veel kernelementen van de informatica: bestandsystemen (Google File System), nieuwe programmeerparadigma's (MapReduce), nieuwe programmeertalen (bijv. Sawzall) en nieuwe aanpakken voor het beheren

van gegevens (BigTable). Allemaal ontwikkeld om grote gegevensverzamelingen gemakkelijk toegankelijk te maken. Deze technieken zijn inmiddels ook beschikbaar in *open source* varianten. De bekendste, Hadoop, werd voor een belangrijk deel ontwikkeld bij Googles concurrent Yahoo. Aan de Universiteit Twente worden de technieken sinds 2009 onderwezen in het masterprogramma *Computer Science*.

De kracht van grote gegevens

Grote Gegevens introduceren nieuwe uitdagingen en nieuwe kansen voor machinaal leren, vooral toegepast op natuurlijke taalverwerking, zoals het zoeken op het web, automatische vertaling en spraakherkenning. In hun artikel *The unreasonable effectiveness of data* beschrijven Google onderzoekers Alon Halevy, Peter Norvig en

Fernando Pereira (2009) de volgende belangrijke uitdaging: *Maak gebruik van grootschalige gegevens die direct beschikbaar zijn, in plaats van te hopen op geannoteerde gegevens die (nog) niet beschikbaar zijn*. Zoekmachines, automatische vertalers en spraakherkenners zijn de laatste jaren enorm verbeterd dankzij de enorme hoeveelheid gegevens die beschikbaar zijn voor het trainen van statistische modellen. Veel gegevens worden routinematig gemaakt, en zijn dus in overvloed aanwezig. Voorbeelden zijn de hyperlinks tussen webpagina's, vertalingen van websites die in meerdere talen beschikbaar zijn, ondertitels voor doven en slechthorenden, de geografische positie bij berichten en foto's als gevolg van GPS in smartphones en camera's, enz.

De kracht van simpele modellen

De statistische technieken waarmee systemen getraind worden zijn de afgelopen jaren minder veranderd dan hun succes doet vermoeden. Nu we in staat zijn om te trainen op grootschalige gegevensverzamelingen doet zich het volgende fenomeen voor: *Eenvoudige modellen getraind met grote gegevens troeven complexe modellen op basis van minder gegevens af*. Michele Banko en Eric Brill (2001), beiden onderzoekers bij Microsoft, waren een van de eersten die dit aantoonde. Ze trainden verschillende methoden voor het desambigueren van woorden met behulp van gegevensverzamelingen van verschillende groottes. Geen van de methoden leek nog asymptotisch gedrag te vertonen bij het trainen op een miljoen woorden, een redelijke hoeveelheid gegevens – zeker voor die tijd. Interessant is dat simpele methoden die het relatief slecht doen op een miljoen woorden, niet onder doen voor complexe modellen als er een miljard woorden beschikbaar is. Eenvoudige methoden die grote gegevens kunnen benutten hebben de voorkeur boven complexere methoden die grote gegevens niet gemakkelijk aankunnen.

Torsten Brants en collega's bij Google (2007) deden een soortgelijk onderzoek naar methoden voor het trainen van statistische taalmodellen voor automatisch vertalen, waarbij ze in staat waren om maar liefst een biljoen woorden te gebruiken. Ze introduceerden daarvoor een nieuwe, eenvoudige methode voor *smoothing* van statistische taalmodellen, *Stupid Backoff* genoemd. Deze methode is goedkoop om te trainen op grote gegevens en benadert de kwaliteit van het complexere en krachtigere *Kneser-Ney smoothing* als de hoeveelheid trainingsgegevens toeneemt. Het werk van Brants en

collega's laat zien dat een complex model gebaseerd op grote gegevens, niet beter zal presteren dan een simpel model gebaseerd op dezelfde hoeveelheid gegevens. De ware kennis zit blijkbaar in de gegevens, niet in de generalisaties die een slim model kan doen.

De kracht van grote gegevens wordt letterlijk geïllustreerd door het werk van James Hays en Alexei Efros (2007) van Carnegie Mellon University, die *afbeeldingsvoltooiing* onderzoeken, dat wil zeggen het invullen of vervangen van een deel van een foto of afbeelding zodanig dat de wijziging niet kan worden gedetecteerd. Voorbeelden zijn het herstellen van een kapotte hoek van een historische foto, of het verwijderen van een ex-echtgenoot van een familiefoto. De methode van Hays en Efros is verbazend simpel. Hun algoritme zoekt in een enorme database naar soortgelijke afbeeldingen en gebruikt daarvan beeldfragmenten om de afbeelding te voltooiën. Ongetwijfeld zou een dergelijke methode niet werken op een kleine database, maar zodra de database groot genoeg is – de auteurs hadden miljoenen foto's – dan overtreft de methode complexere methoden, bijvoorbeeld methoden die de ontbrekende delen van een afbeelding door de analyse van kleuren en textuur proberen te extrapoleren uit de bekende delen.

Een ander mooi voorbeeld van de kracht van grote gegevens is het vraag-en-antwoordsysteem van Susan Dumais en collega's (2002). Vraag-en-antwoordsystemen beantwoorden vragen zoals: 'Wanneer werd Vincent van Gogh geboren?', 'Waar schreef Anne Frank haar dagboek?', of 'Wie was de eerste Nederlandse koning?' Zulke systemen maken veelal gebruik van uitgebreide taalkundige kennis, zoals het ontleden van zinnen, het modelleren van personen, plaatsen en tijden, anaforenresolutie, het gebruik van synoniemen, enz. De aanpak van Dumais en haar collega's beperkt zich tot het herschrijven van de vragen naar simpele beweringen, en het gebruik van een internetzoekmachine waarmee documenten met die exacte bewering wordt gevonden, waarna de tekst volgend op de bewering wordt verzameld. De redenering is als volgt. Als de gegevens groot genoeg zijn, is er altijd wel een bewering te vinden die exact overeen komt met de gestelde vraag. Bijvoorbeeld: 'Vincent van Gogh werd geboren...', 'Anne Frank schreef haar dagboek...' en 'De eerste Nederlandse koning is...'. Op de plaats van de puntjes vindt het systeem vervolgens het antwoord, met minimaal gebruik van taalkundige kennis. Ook deze aanpak zou nooit werken als er geen grootschalige gegevensverzamelingen beschikbaar zouden zijn. Hoe groot moeten de gegevens zijn voordat zo'n aanpak succesvol is? Het werk van Arjen Hoekstra (2006) laat zien dat ook

de hoeveelheid Nederlandse pagina's voldoende groot is voor een 'Grote Gegevens'-aanpak.

Iedereen kan 'Google zijn'

Hoe groot is groot? Brants en collega's van Google trainden hun statistische taalmodellen met behulp van een biljoen woorden. Zelfs als een webpagina gemiddeld 1000 woorden bevat, dan komt dat nog steeds neer op een miljard webpagina's, vele terabytes aan gegevens. Is dat haalbaar voor onderzoekers die niet bij Google in dienst zijn? Het antwoord hierop is: Ja hoor, dat was jaren terug al haalbaar. Eind 2008 kocht de Universiteit Twente, gesponsord door Yahoo, een eerste Hadoop cluster bestaande uit 16 machines, ongeveer 1000 euro per stuk, die gezamenlijk terabytes aan gegevens op kunnen slaan. Bovendien gaf Carnegie Mellon University begin 2009 de ClueWebog webcollectie vrij, een gegevenscollectie bestaande uit een miljard webpagina's. Gecombineerd met de gegevens die op dat moment al aan de Universiteit Twente beschikbaar waren, was dat genoeg voor een biljoen woorden. Medio 2009 had de universiteit dus de kennis, de infrastructuur, en de gegevens in handen om ook een biljoen woorden te analyseren. Wat Google vandaag publiceert, kan men binnen 3 à 4 jaar dus repliceren. Wat vandaag grote gegevens zijn, zijn morgen gewone gegevens. Iedereen kan Google zijn.

Makkelijker zoeken in grote gegevens

De experimenten met ClueWebog laten interessante gevallen zien waarbij simpele modellen net zo goed werken als complexere modellen. De literatuur van het onderzoek naar zoekalgoritmen, in het Engels *information retrieval* genoemd, kent een aantal standaard heuristieken en ordeningsprincipes. De *tf.idf* weging is bijvoorbeeld een bekende heuristiek waarbij het belang van een document voor een zoekvraag wordt berekend uit de *term frequency*, of *tf* waarde van een term: het aantal voorkomens van de term in het document en de *inverse document frequency* of *idf* waarde van een term: het aantal documenten waarin de term voorkomt. De intuïtie is dat termen die in weinig documenten voorkomen (met een hoge *idf*) belangrijker zijn dan termen die in alle documenten voorkomen, en dat documenten met veel voorkomens van de termen (met een hoge *tf*) belangrijker zijn dan documenten met weinig voorkomens van de termen. De *tf.idf* heuristiek kan ook verklaard worden door het gebruik *smoothing*

van statistische taalmodellen.

Uit de ClueWebog experimenten blijkt dat *smoothing*, of het gebruik van de *idf* component in *tf.idf*, niet langer noodzakelijk is voor het met hoge precisie vinden van documenten op het web. Simpele modellen – taalmodellen zonder *smoothing*, of een weging met enkel de *tf* component – werken net zo goed als de gegevens maar groot genoeg zijn (Hiemstra & Hauff, 2011). Zoeken wordt dus gemakkelijker als de gegevens waarin gezocht wordt groter zijn.

Conclusie

In de wereld van Grote Gegevens troeven eenvoudige modellen de complexe modellen af. Toch zijn veel onderzoekers nog altijd in de weer om betere, complexere modellen te bedenken, om dan met weinig gegevens aan te tonen dat die modellen een verbetering opleveren. De in dit artikel beschreven ervaringen met grote gegevens suggereren het volgende advies aan onderzoekers: Misschien is het goed om even pas op de plaats te maken, en de komende tijd te besteden aan het verzamelen van *grote gegevens* in plaats van aan het ontwikkelen van nieuwe, complexere methoden.

LITERATUUR

- Banko, M. & Brill, E. (2001). Scaling to very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Brants, T., Popat, A., Xu, P., Och, F. & Dean, J. (2007). Large Language Models in Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Dumais, S., Banko, M., Brill, E., Lin, J. & Ng, A. (2002). Web Question Answering: Is more always better? In *Proceedings of the 25th ACM Conference on Research and Development in Information Retrieval (SIGIR)*.
- Halevy, A., Norvig, P. & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2).
- Hays, J. & Efros, A. (2007). Scene Completion Using Millions of Photographs. *ACM Transactions on Graphics (SIGGRAPH)* 26(3).
- Hiemstra, D. & Hauff, C. (2011). MapReduce for Experimental Search. In *The Nineteenth Text Retrieval Conference (TREC 2010)*. NIST Special Publications.
- Hoekstra, A., Hiemstra, H., Van der Vet, P. & Huibers, T. (2006). Question Answering for Dutch: Simple does it. In *Proceedings of the 18th BeNeLux Conference on Artificial Intelligence (BNAIC'06)*.

DJOERD HIEMSTRA is universitair hoofddocent Database Technology aan de Universiteit Twente.
E-mail: <hiemstra@cs.utwente.nl>

Over Münchhausen en andere opscheppers Een ode aan Data N. Richmond

Van de in de titel genoemde Richmond zult u wel nooit gehoord hebben. Dat kan ook niet, want het gaat niet om een persoon, maar om een verschijnsel. Als u het niet toch al begrepen had, begrijpt u het wel als u de in de titel genoemde naam zorgvuldig uitspreekt. Inderdaad: *data enrichment* oftewel gegevensverrijking.

Statistici hebben vaak minder waarnemingen tot hun beschikking dan ze zouden wensen. Gelukkig zijn er allerlei methoden om een verzameling gegevens uit te breiden. De meest voor de hand liggende is: meer waarnemingen doen. Maar dat kan niet altijd, en als het wel kan, is het tijdrovend en kostbaar; bovendien leidt het niet altijd tot een gewenst resultaat. Een tweede methode heet *resampling*, een procedure waaraan volgens sommigen ook een luchtje zit. De namen die deze methoden dragen duiden ook niet op grote betrouwbaarheid. Eén van die namen is 'bootstrap methode'. De naam is ontleend aan het verhaal dat Baron von Münchhausen zich aan zijn *bootstraps* (laarzenlussen) uit het moeras getrokken zou hebben. Het verhaal is apocrief: in het oorspronkelijke verhaal trekt hij zich niet aan zijn laarzen omhoog, maar aan zijn haar. De naam *bootstrap* is dus om twee redenen verdacht: Von Münchhausen liegt natuurlijk toch al, en er is ook nog een fout in het verhaal geslopen. De naam voor sterk verwante 'resampling methode' is Jackknife. Ook tamelijk louche. *Jackknife* betekent knipmes, zo'n ding waarmee Mackie Messer onder anderen Schmul Meier en Jenny Towler vermoordt. U weet wel, in het Duits en Engels:

*Und der Haifisch der hat Zähne,
und die trägt er im Gesicht,
und Macheath, der hat ein Messer,
doch das Messer sieht man nicht.*

*Oh the shark has pretty teeth, dear,
and he shows them pearly white.
Just a jackknife has MacHeath, dear,
and he keeps it out of sight.*

De Jackknife deugt dus eigenlijk ook niet. De derde vorm van gegevensverrijking is de laatste jaren regelmatig in het slechte nieuws: verzonnen, aangevulde of gemanipuleerde data. Van de opvallendste verrijkers zijn de namen bekend; ik zal ze niet nog eens noemen. Tot mijn genoegen is het probleem niet beperkt tot Brabant, waar behalve Tilburg ook Eindhoven – lang geleden – een schandaaltje had. Soortgelijke problemen deden zich recentelijk voor aan het Erasmus MC, de VU en Nijmegen; zelfs helemaal in Korea. Maar daartoe zal het niet beperkt zijn; er is overal wel wat, want: waarnemingen heeft men nooit genoeg!

FRED STEUTEL is emeritus hoogleraar kansrekening aan de TU Eindhoven.
E-mail: <f.w.steutel@tue.nl>



NS-PROJECT AURORA VOORSPELT VOOR ELKE TREIN HET AANTAL REIZIGERS

JAN HOOGENRAAD, AART DE VOS & JAAP DE VRIES

De NS organiseerde in 2009 een wedstrijd tussen drie wiskundige teams. De opdracht was om in twee maanden zo goed mogelijk in te schatten hoeveel reizigers er per traject en tijdstip in de treinen hebben gezeten. Het ging om treinen die in 2008 in de Kop van Noord-Holland hebben gereden. De wedstrijd is gewonnen door econometristen van de Vrije Universiteit Amsterdam (VU) maar vormde ook de start van het project AURORA. In *STATOR* werd verslag gedaan van de wedstrijd.¹

De wedstrijd heeft een ruim vervolg gekregen: AURORA verschaft al jaren voorspellingen voor iedere trein en ieder traject. Twee miljoen voorspellingen per twee maanden op basis van honderden miljoenen chipkaartreizen. Sinds 2011 zijn deze prognoses de basis voor de planning van de capaciteit van treinen. Dat laatste gaat nu met een geheelallig lineair programmeringsprobleem, waardoor de planning meermalen per jaar kan worden bijgesteld. De noodzakelijke handmatige bijstel-

ling van prognoses en plannen is veel kleiner geworden. De treinen zijn gemiddeld korter geworden, vooral midden op de dag, en het aantal problemen (te volle treinen) dat een gevolg is van foute prognoses is afgenomen (maar nog niet nul). Kortom: AURORA is een succes.

Het project begon in 2009 als een voorspelwedstrijd. Aan de hand van 'Meten in de Trein' (MidT) gegevens. In iedere wagon worden vier maal per jaar de kaartsoorten geteld ten behoeve van de verdeling van opbrengsten tussen vervoerders. Sinds 2007 was de dienstregeling nauwelijks veranderd, dus waren er 200.000 datapunten beschikbaar voor een model.

De opzet was eenvoudig: 10% van de data werd geheimgehouden en wie bij het voorspellen de laagste Root Mean Square Error (RMSE) had won. Als voorbeeld was er een referentiemodel, gebaseerd op een trainings-sample van 10% en allerlei ideeën om zo goed mogelijk hierop te scoren. Op de geheime set had dat dan een wat

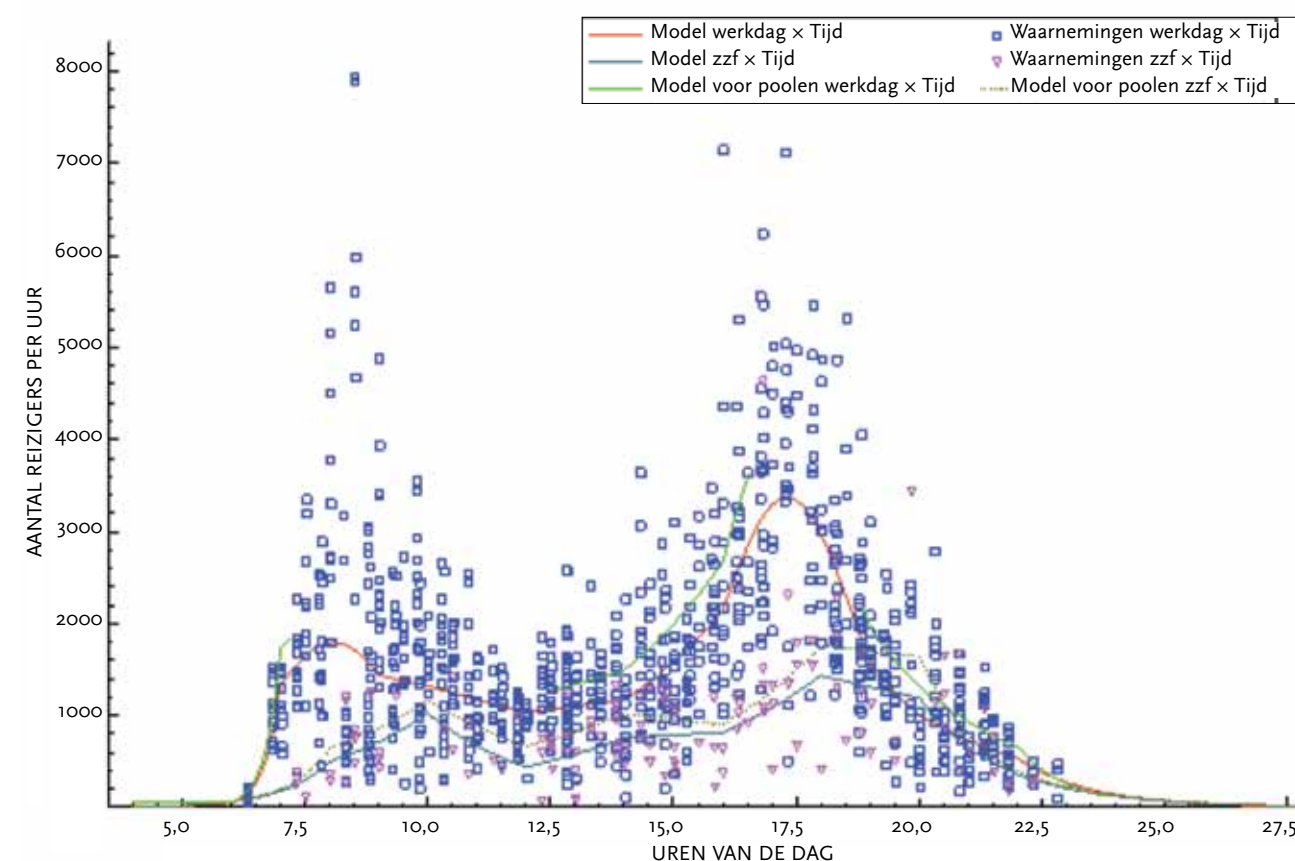
lagere score door overfitten, maar het werkte best aardig. Twee teams – CQM in Eindhoven (Consultants in Quantitative Methods) en de VU – kwamen bij de score van het referentiemodel in de buurt met multiplicatieve modellen voor tijd(splines)- en trajecteffecten. Het derde team, LIME van de Technische Universiteit Eindhoven, poogde reizigersgedrag te modelleren, maar dat bleek met de beperkte tijd en de schaarse data niet goed haalbaar.²

Leerzaam proces

Er volgde AURORA-10 waarbij het hele land in kaart werd gebracht. Het team van de VU modelleerde op het niveau van treinen, maar een trein die het halve land doorgaat heeft in de loop van zijn traject te maken met veel rolwisselingen. Het CQM-team knipte het land in stukjes, dat ging beter. Duidelijk werd dat de codering van data cruciaal was. Treinnummers veranderen soms, er komen nieuwe stations, en er zijn allerlei verschillen in administratieve details in de loop der jaren. Op basis van de geleverde prognoses werd in 2010 voor het eerst

een aparte materieelomloop voor de zomervakantie gemaakt.

Op 1 juli 2010 was het eindelijk zo ver dat AURORA-11 (voorspellen voor 2011) van start ging. Een project om op de basis van MidT-gegevens te komen tot prognoses. Het werd een bijzonder project. Een botsing tussen methodologieën. Aan de ene kant de methode van het referentiemodel, waarbij razendsnelle programma's (in PERL) spreadsheets met in bakjes opgesplitste uitslagen produceerden en creatief werden gemanipuleerd om een zo laag mogelijke score in de RMSE te behalen. Aan de andere het team op de VU dat gecompliceerde statistische trucs toepaste, op logaritmen, en met een computertaal Ox, een objectgeoriënteerde taal die alles in matrices doet maar veel aanpassingen vergde om met grote databestanden vol strings en uitzonderingen te werken. En daartussenin CQM die met het pakket STATA zeer snel nieuwe aspecten van de data tevoorschijn toverde. Het wekelijks overleg kwam voor sommige aanwezige NS'ers over als één grote ruzie maar voor de betrokkenen als een leerzaam proces. Samengevat met de volgende limerick



Figuur 1. Reizigersstroom op een traject als functie van de vertrektijd, apart voor werkdagen en voor zaterdagen/zondagen/feestdagen (zzf). Dit plaatje is uit 2011, met alleen de data van Meten in de Trein.

De VU levert schatters met dakjes
En Jan stopt alles in bakjes
En wat zit daar in?
De RMS BIN
Misschien begrijpen we 't strakjes

De MidT-tellingen leverden per traject plaatjes als in figuur 1 (een log-lineaire spline met spitsparabolen). Er doken allerlei problemen op. Soms was de melding 'treinstel te vol' of 'slechts een deel geteld', passagiers zijn vaak ongelijk verdeeld over de trein (het 'voor, midden, achter effect'), er was niet bijgehouden welk deel eerste-klasseplaatsen (overdag vaak leeg) betrof, enz. En daarbovenop waren er allerlei effecten van verstoorde treinenloop, popfestivals, studenten die allemaal in dezelfde trein zitten, en dergelijke. En weinig tellingen in weekends en 's morgens vroeg en 's avonds laat. Daar kan je op allerlei manieren mee omgaan. Waarbij niet alleen de statistische fit maar ook het doel van de exercitie (capaciteit van de treinen plannen) telt. Op de 10% van de data die niet gebruikt is voor de modellen wordt de betrouwbaarheid van de voorspellingen geëvalueerd. Daaruit wordt berekend welke capaciteit nodig is om bijvoorbeeld met 90% kans voldoende plaatsen te bieden.

Resultaat van het werken met meer modellen was de 'mix', een gewogen gemiddelde voorspelling op basis van de beste performance. Ook werd een systeem opgezet waarin je kunt zien wanneer de modellen sterk verschilden bij een voorspelling. Met die inzichten konden de modellen verbeterd worden, en een berede- neerde keuze gemaakt welke prognoses mee te nemen in de mix voor de planning. Bij het uiteindelijk mondel- ing overleg met mensen van NS was dit systeem zeer belangrijk.

Een doorbraak binnen de NS organisatie volgde toen 1200 volledige tellingen werden uitgevoerd en bleek dat de voorspellingen een factor twee nauwkeu- riger waren dan de gebruikelijke op basis van conduc- teurstellingen. Ook kwam de werkelijke nauwkeurig- heid overeen met de binnen AURORA (met de 10% toetsdata) bepaalde.

Eindeloos waren de discussies over transformaties (de logaritme of misschien toch de worteltransfor- matie?) en de relevante beoordelingscriteria. Waarbij vooral opviel dat de rangorde van de scores weinig ge- voelig is voor het criterium dat men kiest. Uiteindelijk werd gekozen voor de RMSratio, een transformatie van de bekende R^2 . Zonder logaritmen, die leefden slechts voort in de modellen en in de volgende limerick.

Ln(OV)

Er was een geleerde uit Amsterdam
Die steeds de logaritme nam
En op de vraag
Is dat niet traag?
Zei hij: veel sneller dan de **Tram**

Eens bij een reis naar Nieuwegein
Sprak hij: kijk dat is fijn
Een passagiersaantal van nul
Geeft logaritmisch flauwekul
En hij stapte in een lege **Trein**

Hij legde uit: zo zit het dus
De log maakt van een maal een plus
Zo dat men zonder hinder
Werkt met percentages meer of minder
En het klopt altijd als een **Bus**

Alles bij elkaar was AURORA-11 een groot succes. Het bleek mogelijk om ondanks de schaarste aan data rei- zigersaantallen goed te voorspellen en daarmee beter de nodige capaciteit van treinen te plannen. Binnen NS kwam hiermee ook een proces op gang om sneller en nauwkeuriger te plannen. In dit proces vervulde de projectleider AURORA ook de rol van data scientist: het begeleiden in het gebruik van de data.

Puzzels

AURORA-12 had de ambitie om de OV-chipkaartgegevens te verkennen. Maar de OV-chipkaart kwam maar langzaam op gang; gemiddeld 12% van de reizigers reisde in 2011 er mee. Hiermee omvatte het project pas echt Big Data: we analyseren nu 13 miljoen reizen per maand. Uit de gegevens van stations en tijdstippen waar in- en uitgecheckt wordt is zeer goed af te leiden hoeveel OV-chipkaartreizigers zich in iedere trein bevinden. Hiertoe is een algoritme ROCKT ontwikkeld. De hoeveelheid data is enorm, maar na twee dagen rekenen levert het programma de gewenste output. De implementatie is in handen van NS, het AURORA-team hielp bij de ontwikkeling en onderzocht de voorspellende kwaliteit. Ook werden weeggegevens toegevoegd: een aan- tal treinen is uitgerust met weegapparatuur, waardoor we nog meer verschillende gegevens krijgen.

AURORA-13 volgde. Er waren nog vele puzzels op te lossen, en antwoorden te vinden op vragen die vroeg of laat gesteld worden. Hoogtepunt van dit project was de ontwikkeling van een algoritme om het effect van

AURORA: MODELLEN EN DEELNEMERS

In de AURORA-projecten werkten enkele afdelin- gen van NS samen. Voor het maken van model- len werden externe organisaties ingehuurd. De volgende personen werkten daaraan:

Model-C, gemaakt door CQM, Consultants in Quantitative Methods:

- Monique van den Broek (wiskundige, senior consultant bij CQM)
- Marcel van Vuuren (wiskundige, senior consul- tant bij CQM)
- Marnix Zoutenbier (principal consultant bij CQM)

Model-R (Referentie), voorbereiding, evaluatie en mixen:

- Jan Hoogenraad (natuurkundige, ervaring als programma manager, test- en kwaliteitsmana- ger en change manager)

Model-V (in eerste instantie ontwikkeld op de VU)

- Aart de Vos (UHD econometrie, nu gepensio- neerd).
- Kai Ming Lee – tot en met AURORA-10 (econo- metrist).
- Jaap de Vries – vanaf AURORA-11 (econo- trist met veel kennis van openbaar vervoer)
- Zes studenten econometrie – in AURORA-11

NS-afdelingen:

- Commercie (Edwin Eshuis, Bert de Vries, Ad Ligtenberg, Henri Kroesen)
- Finance Center (Victor van den Broek, Kees Jong, Karen Slijkhuis)
- Logistiek (Hans van Uden, Herman Poelen, Ruud Drabbe)
- Proceskwaliteit en Innovatie (Ellis Leijte, Dirk Jekkers, Ramon Lentink)

Vanaf de start van AURORA-13 (juli 2012) werken Jan Hoogenraad, Aart de Vos en Jaap de Vries samen in een nieuwe bv 'Spoorglaren'. Dit is het team dat sinds het voorjaar van 2013 is overge- bleven voor AURORA.

grote veranderingen in de dienstregeling te voorspel- len, op basis van OV-chipkaartgegevens. Eind 2012 is de Hanzelijn in gebruik genomen, en is op veel plaatsen in het land de dienstregeling ingrijpend gewijzigd. Om de gewijzigde reispatronen te kunnen voorspellen was een uitbreiding van de AURORA-methodiek nodig.

Die kon worden ontwikkeld en getest op een wijzi- ging een jaar eerder: in 2012 werden treinen van Amster- dam naar Rotterdam via Haarlem geleid in plaats van Schiphol. Hoe voorspel je wat mensen gaan doen? Je neemt de chipkaartreizen van 2011, neemt aan dat men- sen rond dezelfde tijd hun reis willen maken, voegt een preferentiefunctie in met onder meer voorkeur voor wei- nig overstappen, en schat dat om zo goed mogelijk 2012 te voorspellen. Dit gebruikten we om de gevolgen van de nieuwe Hanzelijn te voorspellen. Ook lukte het te la- ten zien hoeveel effect verstoringen hebben op aantallen passagiers in treinen.

Conclusie

Hoeveel mensen zitten er in de trein? Een eenvoudige vraag. Maar het is een fantastische uitdaging om met al- lerlei bronnen van informatie er iets over te zeggen. Een uitdaging die alles bevat wat een modellenbouwer zich kan wensen. Tijd en plaats en alles wat er mis kan gaan. En geen theorie, maar harde en zeer voorstelbare werke- lijkheden. Een fout maken kan betekenen dat mensen niet in de trein kunnen. Goede planning spaart elektriciteit en geld (verbetering van de bezettingsgraad leverde een aan- zienlijk bijdrage aan de energiebesparing van NS van de afgelopen twee jaar). En bestaande data slim gebruiken is vele malen goedkoper dan het verzamelen van nieuwe.

NOTEN

1. Hoogenraad, J., Lentink, R. & De Vries, B. (2010). Hoeveel reizigers zitten er in de trein? Wedstrijd NS groot succes! *STATOR*, 11(1), 8-10.
2. LIME werd onderdeel van Sioux, en deed in het vervolg niet meer mee.

JAN HOOGENRAAD is directeur van Spoorglaren, een bureau voor dienstverlening aan de spoorsector.
E-mail: <jan.hoogenraad@spoorglaren.nl>

AART DE VOS was universitair hoofddocent aan de VU en heeft een adviesbureau Ecstasystems.
E-mail: <ecstasystems@gmail.com>

JAAP DE VRIES werkt als consultant bij Spoorglaren en bereidt een proefschrift voor aan de VU.
E-mail: <jaap.devries@spoorglaren.nl>



SANDER KLOUS

Data zijn het nieuwe goud. Dat blijkt bijvoorbeeld weer nu Twitter heeft aangekondigd een beursgang voor te bereiden. De schattingen over hoeveel geld het sociaal medium precies waard is variëren onder analisten nogal, maar de waarde zal in elk geval in de miljarden lopen. Die waarde schuilt natuurlijk in de data die het bedrijf heeft over de miljoenen tweets wereldwijd: op basis van hun tweets weet men onder andere waar deze mensen zich bevinden, wat hun interesses zijn en hoe hun netwerken eruit zien. Twitter mag dan tot nog toe geen winst hebben gemaakt, vrijwel iedereen is het erover eens dat de kennis over de gebruikers uiteindelijk tot waarde zal leiden.

De slag om de data is nog maar net begonnen en tal van grote bedrijven zullen zich de komende jaren nadrukkelijk mengen in deze slag. Technologiebedrijven als Google, Apple en Samsung willen dolgraag dat u hun apparaten – fotocamera's, telefoons, televisies – gaat gebruiken. Google zet bijvoorbeeld hoog in op Google Glass; Apple op Apple TV en Samsung op het slimme horloge (Galaxy Gear). Niet (alleen) omdat ze met de verkoop van die hardware geld verdienen maar vooral omdat ze daarmee weer een 'sensor' op u kunnen plakken. Met die sensor krijgen ze meer inzicht in uw gedrag en daar ligt waarschijnlijk hun toekomstig bestaansrecht.

Het commerciële voordeel van die data is evident: deze bedrijven hoeven niet meer met hagel te schieten, maar kunnen een precisiebombardement van aanbiedingen afvuren. Zeker als ze gebruik kunnen maken van data uit verschillende bronnen. Er is niet veel verbeeldingskracht voor nodig om te bedenken wat er dan mogelijk is. McDonald's weet dat u al een paar uur in de auto zit en doet u in de buurt van een van haar vestigingen een kortingsvoorstel voor uw favoriete menu. Technisch gezien is dat geen enkel probleem, mits McDonald's toegang heeft tot de real-time data van bijvoorbeeld een mobiele provider.

Maatschappelijke winst

Zo zijn er tal van nieuwe manieren van marktwerking te bedenken. Toch is dit maar één kant van het verhaal over Big Data. De echte uitdaging voor bedrijven zit erin om het concept van Shared Value Creation (Porter & Kramer, 2011) van de tekentafel naar de praktijk te brengen. In dat concept gaat een onderneming niet alleen maar voor bedrijfseconomische winst, maar vooral ook voor waardecreatie voor *people* en *planet*. Vertaald naar Big Data toepassingen: het verwerken van grote hoeveelheden gegevens moet leiden tot gemak, tot nieuwe diensten waar de consument echt wat aan heeft, of tot efficiëntere processen (minder rompslomp of lagere kosten). Alleen als de maatschappij en de klant het nut inzien van de verwerking van (hun) gegevens, zijn ze bereid te accepteren dat er ook financieel gewin in zit voor de uitvoerende organisatie. Zuiver commerciële toepassingen stuiten begrijpelijkerwijs op steeds meer weerstand uit de privacy hoek.

In het eerder genoemde voorbeeld van McDonald's is weinig maatschappelijke winst te ontdekken. Maar er zijn tal van voorbeelden denkbaar – die deels al in de praktijk bestaan – waarin dat wel het geval is.

Voorbeeld 1. De koppeling van de data van een mobiele provider aan de data van een creditcardmaatschappij maakt het mogelijk een creditcarduitgave te blokkeren als die creditcard zich op meer dan 100 meter afstand bevindt van uw telefoon. Commerciële winst: lagere fraudekosten. Maatschappelijke winst: veiliger betalen.

Voorbeeld 2. Een verzekeringsmaatschappij voorziet klanten van kleine kastjes met een GPS-systeem in de auto die het rijgedrag te monitoren. Hoe veiliger het rijgedrag – waarover directe feedback in de auto wordt gegeven - hoe lager de premie voor de klant. Commerciële winst: een scherpere premie. Maatschappelijke winst: hogere verkeersveiligheid.

Deze wereld kan veel beter. Sander Klous, bij KPMG verantwoordelijk voor een Data Analytics team met een rijke ervaring bij de Zwitserse deeltjesversneller CERN, meent dat toepassing van Big Data analyses daarvoor kan gaan zorgen. Big Data kunnen processen in de zorg verbeteren, voorkomen dat burgers in financiële problemen komen, de veiligheid in de maatschappij verhogen en ondertussen ook nog commerciële kansen ontsluiten voor het bedrijfsleven. Hij waarschuwt tegelijkertijd dat we Big Data niet moeten zien als eenvoudig toe te passen Haarlemmerolie voor elk probleem. Want wie de kansen van Big Data wil verzilveren krijgt in de praktijk vaak te maken met tegenslagen en moet heel kritisch zijn bij het interpreteren van de resultaten. Simpson's paradoxen zijn aan de orde van de dag in deze wereld.

Voorbeeld 3. Een aannemer die een snelwegproject uitvoert kan met de data van de mobiele provider precies zien waar het verkeer zich ophoopt en daar gericht actie op ondernemen. Commerciële winst: in veel gevallen wordt deze aannemer contractueel financieel geprikkeld om die congestie zoveel mogelijk te beperken, dus het gaat om wezenlijke voordelen. Maatschappelijke winst: minder congestie op de weg.

Data analyse – geen routine

We staan nog maar aan het begin van een uitrol van Big Data toepassingen en de toekomst ziet er zonder meer veelbelovend uit. Technisch is er veel meer mogelijk dan een jaar of tien geleden en de maatschappij verwacht ook dat ze beter wordt bediend. Grote hoeveelheden rekenkracht zijn zo gemakkelijk te mobiliseren dat analysemogelijkheden van gestructureerde en ongestructureerde data binnen bereik komen die tot voor kort onhaalbaar waren. Die rekenkracht is ook hard nodig, want de opgave is niet gering. Volgens schattingen produceren we dagelijks 2,5 quintillion bytes (of 2.5 exabytes: 2.5×10^{18}) aan gegevens, hebben we in de afgelopen twee jaar 90% van alle data ooit geproduceerd met exponentiële groei en is 80% van de beschikbare data ongestructureerd.

Juist de technologische mogelijkheden vormen in zekere zin echter ook een bedreiging voor een succesvolle opkomst van Big Data toepassingen. Het feit dat we nu relatief eenvoudig complexe analyses kunnen loslaten op enorme bakken data is namelijk bepaald nog geen garantie voor succes. Opgejut door de hype ontstaat soms een beeld van Big Data als een machine waar je naar believe grote hoeveelheden data in giet en er vervolgens aan de andere kant kant en klare oplossingen uithaalt. De werkelijkheid is een stuk weerbarstiger. Het realiseren van succesvolle toepassingen op het gebied van Big Data is vaak een proces vol tegenslagen. Data scientists mogen niet te snel juichen als ze bepaalde verbanden aantreffen maar heel kritisch zijn op wat de analyses opleveren. Ze moeten zich verplaatsen in de wereld waar de data vandaan komt om te ontdekken waarom bepaalde relaties in data ontstaan. En dat is ook precies waarom het vak van data scientist zo uitdagend en leuk is.

Het managen van verwachtingen is een belangrijke taak voor data scientists. We moeten blijven uitleggen dat Big Data veel meer is dan *number crunching*. Van groot belang is dat het verschil tussen correlatie en causaal verband goed wordt begrepen bij de analyse van data. Daarom leggen we ook aan het management uit

dat Simpsons' paradox heel relevant is. Dat kan aan de hand van eenvoudige en treffende voorbeelden. Een daarvan is dat uit statistieken blijkt dat zeelieden die overboord zijn geslagen zonder reddingsvest vaker werden gered door de hulpdiensten dan de zeelieden die wel een reddingsvest droegen. Dat druist in tegen elke intuïtie maar blijkt bij nadere analyse heel goed verklaarbaar. De zeelieden bleken namelijk vooral in slechte weersomstandigheden ervoor te kiezen wel een reddingsvest te dragen. Het is typerend voor de wereld waarin een data scientist opereert: als je gegevens op een handige manier combineert, komen daar soms onwaarschijnlijke resultaten uit. Een goede data scientist moet de resultaten kunnen begrijpen, mogelijk met interessante inzichten voor de opdrachtgever tot gevolg. Een foute analyse kan levensgevaarlijk zijn. Het zal de eerste keer niet zijn dat op basis van een dergelijke data analyse wordt besloten om het gebruik van zwemvesten dan maar te staken.

In een omgeving met grote hoeveelheden data – die complex, ongestructureerd en multi-interpretabel kunnen zijn – is het vaak veel minder eenvoudig om relaties tussen variabelen te begrijpen dan in de casus van de reddingsvesten. De ervaringen die we hebben opgedaan bij de deeltjesversneller CERN in de buurt van Genève zijn daarbij zeer nuttig. In deze omgeving – een ring van 27 kilometer met reusachtige detectoren zo groot als de Dom in Utrecht – klapten de afgelopen twee jaar talloze protonen met bijna de snelheid van het licht tegen elkaar aan. Bij deze botsingen, waarbij protonen uiteen spatten in allerlei subdeeltjes, werd gezocht naar sporen van het tot vorig jaar nog onvindbare Higgs-deeltje. We stonden daar elke keer weer voor de taak elkaars resultaten uit te dagen – we zagen onder andere de stand van de maan en het treinschema van de TGV terug in de analyses – en doen dat nu ook in Big Data toepassingen. Data scientists zijn dus veel meer dan slimme jongetjes en meisjes die handig zijn met statistiek. Ze moeten in staat zijn om als een ware Sherlock Holmes te werk te gaan. Ze moeten ook tegenslagen aankunnen bij het zoeken naar en begrijpen van significante resultaten. En ze moeten zich als geen ander kunnen inleven in de wereld waar de data vandaan komt. Dat alles maakt dat Big Data projecten nooit op routine gaan. En dat is precies waarom data scientists elke keer weer met groot enthousiasme op een nieuwe uitdaging duiken.

SANDER KLOUS is Data & Analytics lead bij KPMG Management Consulting.
E-mail: <Klous.Sander@kpmg.nl>

De dood of de gladiolen

Honderd gevangenen worden door een bewaker bijeengeroepen en krijgen meegedeeld dat ze zullen worden overgeplaatst. De overplaatsing is naar een berucht strafkamp waar hen een vrijwel zekere dood wacht. De gevangenisdirecteur heeft echter besloten dat de gevangenen een kans krijgen op vrijlating in plaats van overplaatsing. Op een lange tafel staan 100 afgesloten dozen en de naam van elke gevangene zit in precies één doos in een volledig willekeurige volgorde. De gevangenen mogen één voor één hooguit 50 dozen openen om hun naam te vinden. De gevangenen mogen tijdens de zoektocht geen onderling contact hebben en ze mogen ook geen dozen verplaatsen of namen in de dozen verwisselen. Als iedere gevangene zijn naam vindt, dan worden de gevangenen vrijgelaten. De gevangenen krijgen een half uur om het voorstel onderling te bespreken en daarna de bewaker te laten weten of ze erop ingaan.

Dit lijkt een voorstel dat onmogelijk tot succes kan leiden. Als iedere gevangene willekeurig 50 dozen onderzoekt, dan is de kans op vrijlating van de groep gelijk aan $(\frac{1}{2})^{100}$ en dus onvoorstelbaar klein. Gelukkig heeft één van de gevangenen econometrie gestudeerd voordat hij als investeringsbankier aan lager wal geraakte en in het gevang belandde. Deze wiskundig onderlegde gevangene komt met de mededeling dat met een kans van ongeveer 30% de vrijheid lonkt. Zijn medegevangenen kunnen dit nauwelijks geloven. De oplossing is echter simpel en eenvoudig te begrijpen. Elke doos wordt aan één gevangene toegewezen en iedereen noteert op een papiertje welke doos bij welke gevangene hoort. Elke gevangene gaat eerst naar de hem toegewezen doos. Treft hij daar een andere naam dan de zijne, dan gaat hij vervolgens naar de doos die aan de andere naam toege-

wezen is. Dit gaat voort totdat de gevangene zijn eigen naam vindt of vruchteloos 50 dozen heeft geïnspecteerd.

Bij deze strategie is de kans 31,2% dat alle gevangenen hun naam vinden. Dit is in te zien met het begrip cykel van een permutatie. Dit begrip lichten toe aan de hand van de permutatie (3, 5, 1, 4, 7, 2, 6) van (1, 2, 3, 4, 5, 6, 7). Deze permutaties bestaat uit de drie disjuncte cyclen $1 \rightarrow 3 \rightarrow 1$, $2 \rightarrow 6 \rightarrow 7 \rightarrow 5 \rightarrow 2$ en $4 \rightarrow 4$ van de lengtes 2, 4 en 1. De kans dat elke gevangene zijn naam vindt, is niets anders dan de kans dat een random permutatie van de getallen 1,2,...,100 geen cykel bevat met een lengte groter dan 50. Wat is deze kans? Beschouw daartoe een random permutatie van de getallen 1,2,...,2n. Op grond van het feit dat een permutatie van 2n getallen hooguit één cykel met lengte groter dan n kan bevatten, is de kans dat de random permutatie een cykel met lengte groter dan n bevat gelijk aan

$$\sum_{k=n+1}^{2n} \frac{\binom{2n}{k} (k-1)! (2n-k)!}{(2n)!} = \sum_{k=n+1}^{2n} \frac{1}{k}.$$

Deze kans wordt voor n voldoende groot benaderd door $\ln(2n) - \ln(n) = \ln(2)$, oftewel de kans op vrijlating is ongeveer gelijk aan $1 - \ln(2) \approx 0,307$ bij 2n gevangenen, ongeacht hoe groot n is. Voor het geval van 100 gevangenen is de precieze waarde van de kans op vrijlating gelijk aan 0,312. Zo zie je maar, wiskunde komt altijd van pas.

HENK TIJMS is emeritus hoogleraar operations research aan de Vrije Universiteit en auteur van diverse leerboeken over operations research en kansrekening.
E-mail: <tijms@quicknet.nl>



Foto: Pieter Bosch.

FOKWAARDESCHATTING BIJ DIEREN

Het schatten van erfelijke aanleg door het oplossen van veel vergelijkingen

Fokkerij is het verbeteren van de erfelijke aanleg van de dieren in de volgende generatie door de beste dieren uit de huidige generatie te gebruiken als ouderdieren. Het schatten van de erfelijke aanleg van dieren, oftewel fokwaardeschatting, is daarbij een belangrijk instrument. Fokwaardeschatting van dieren wordt uitgevoerd met behulp van zogenaamde mixed model vergelijkingen. Door het grote aantal waarnemingen en dieren, moeten miljoenen mixed model vergelijkingen opgelost worden. Door DNA informatie kan al op jonge leeftijd een redelijk betrouwbare schatting gedaan worden van de erfelijke aanleg van een dier. Gebruik van DNA-informatie zorgt voor een verdere verhoging van benodigde computercapaciteit.

HAN MULDER

Fokkerij van dieren is gericht op het verbeteren van eigenschappen van dieren door gerichte selectie van ouderdieren. In feite is fokkerij een vervolg op domesticatie, het aanpassen van dieren voor het beoogde doel waarvoor dieren gehouden worden, zoals melk of vlees. Rond 1950 ontstonden voor verschillende landbouwhuisdieren specifieke fokprogramma's om door gerichte selectie bepaalde kenmerken van dieren te verbeteren. Om fokprogramma's op te zetten zijn een aantal stappen nodig: 1. Het fokdoel: wat wil je verbeteren?; 2. Het verzamelen van data; 3. Het schatten van de erfelijke aanleg van dieren door middel van statistische methodes oftewel fokwaardeschatting; en 4. Selectie en paring van de ouderdieren. Fokwaardeschatting is daarbij een belangrijke stap om de genetische beste dieren als ouders van de volgende generatie te kunnen selecteren.

In dit artikel geef ik een beeld hoe fokprogramma's georganiseerd zijn met daarbij vooral aandacht voor hoe fokwaardeschatting bij dieren werkt en hoe DNA-informatie, naast afstammingsgegevens, daaraan een bijdrage kan leveren.

Het fokdoel: wat wil je verbeteren?

Voordat een fokprogramma wordt opgezet, moet worden vastgesteld welke kenmerken genetisch verbeterd moeten worden. Dit wordt het fokdoel genoemd. In het fokdoel geven we aan welke kenmerken verbeterd moeten worden en in welke richting, bijvoorbeeld het verhogen van melkproductie bij melkvee of het verhogen van de vruchtbaarheid bij varkens. Vaak zijn deze fokdoelen gebaseerd op economie. Het totale fokdoel bestaat uit een aantal kenmerken: sommige direct op het eindproduct gericht, bijvoorbeeld melk, en sommigen op andere aspecten zoals gezondheid, levensduur en reductie van de ecologische voetafdruk.

Verzamelen van data

Wanneer het fokdoel vastgesteld is, moet worden nagedacht hoe die kenmerken gemeten kunnen worden,

aan welke dieren, hoe vaak en wanneer. Hierbij loop je soms tegen dilemma's aan. Sommige kenmerken, zoals melkproductie, kunnen alleen gemeten worden aan vrouwelijke dieren, terwijl je ook mannelijke dieren hierop zou willen selecteren. Andere kenmerken, zoals vleeseigenschappen, kunnen alleen gemeten worden als het dier geslacht wordt. De consequentie is natuurlijk wel dat het dier dan niet meer gebruikt kan worden voor fokkerij. Een oplossing voor dit soort problemen is de kenmerken meten aan familieleden van het dier dat mogelijk geselecteerd wordt. Een ander soort dilemma is dat het nauwkeurig meten van sommige kenmerken alleen mogelijk is met een dure meting. Als er geen goedkoper alternatief bestaat betekent dit dat maar van een zeer beperkt aantal dieren waarnemingen beschikbaar zullen zijn.

De hoeveelheid data die wordt verzameld en geanalyseerd voor fokwaardeschatting is voor veel landbouwhuisdieren erg groot. Bijvoorbeeld bij melkvee wordt in Nederland per jaar van ongeveer 670.000 koeien melkproductiegegevens geregistreerd, waarbij per koe 10 tot 12 metingen gedaan worden. Omdat ook data uit het verleden gebruikt wordt, zijn er in totaal ~170 miljoen dagproducties van 9 miljoen koeien beschikbaar voor fokwaardeschatting. Naast de dagproducties worden van deze koeien ook andere kenmerken zoals vruchtbaarheid vastgelegd.

Fokwaardeschatting

Als de data verzameld zijn, kan de erfelijke aanleg van een dier geschat worden met zogenoemde fokwaardes. Een fokwaarde geeft aan hoeveel beter de genetische aanleg van het dier is ten opzichte van het populatiegemiddelde. Tussen 1950 en 1980 werden regressiemethodes gebruikt om fokwaardes te schatten. Als ouderdieren gemiddeld beter zijn dan generatiegenoten, dan verwacht je dat de nakomelingen dat ook zijn. Maar omdat niet alle variatie die je ziet wordt veroorzaakt door genetische verschillen, maar ook door milieu invloeden, zoals voerkwaliteit, zullen de nakomelingen het gemiddeld minder goed doen

dan hun ouders. Deze relatie tussen de fenotypes van nakomelingen en ouders is door een lineaire regressie te bepalen.

- De regressiemethode had twee belangrijke nadelen:
1. het was niet mogelijk om rekening te houden met systematische effecten, zoals voerkwaliteit, waardoor er bias op kon treden;
 2. niet alle beschikbare informatie van familieleden werd meegenomen omdat dit erg bewerkelijk was.
- Het gevolg was dat de fokwaardeschattingen minder nauwkeurig waren dan gewenst.

In de jaren 70 en 80 van de vorige eeuw is daarom de zogenaamde BLUP-methode (best linear unbiased prediction) geïntroduceerd in de fokwaardeschatting, waarbij mixed model vergelijkingen worden gebruikt om simultaan fixed effecten en fokwaardes te schatten (Henderson, 1975). Fixed effecten zijn systematische milieueffecten zoals voerkwaliteit, bedrijf, jaar, seizoen of leeftijd. De BLUP methode zorgt ervoor dat alle afstammingsrelaties kunnen worden meegenomen in een optimale weging van alle informatie die beschikbaar is. Dit wordt ook wel een diemodel genoemd. Het BLUP-model kan worden weergegeven als:

$$y = Xb + Zu + e \quad (1)$$

waarbij **y** de vector is met fenotypes, **X** is de incidentie matrix om fenotypes aan fixed effecten te relateren, **b** zijn de fixed effecten, bijvoorbeeld bedrijf, **Z** is de matrix om fenotypes te relateren aan fokwaardes, **u** is de vector met random fokwaardes en **e** is de vector met residuen. We kunnen schattingen van fixed effecten en fokwaardes krijgen door mixed model vergelijkingen op te lossen:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \quad (2)$$

waarbij λ de ratio is tussen de residuele variantie en de genetische variantie en A^{-1} de inverse van de matrix met alle zogenaamde additief genetische relaties tussen dieren. De relatie tussen een ouder en een nakomeling is bijvoorbeeld 0,5, omdat de nakomeling de helft van de genen van zijn ouder krijgt. De relatie tussen twee volle broertjes of zusjes is ook 0,5, omdat gemiddeld genomen de helft van het genetische materiaal hetzelfde is. Het aantal op te lossen vergelijkingen in een BLUP-model kan heel groot zijn, bijvoorbeeld in melkvee moeten meer dan 300 miljoen

TYPE INFORMATIE	NAUWKEURIGHEID
eigen fenotype	0,55
fenotypes ouders	0,39
5 volle broers/zussen	0,48
100 halfzussen	0,47
100 nakomelingen	0,94
DNA 2000 dieren*	0,74

* Daetwyler et al., (2008); 2000 dieren met fenotypes en genotypes.
Tabel 1. Nauwkeurigheid van de geschatte fokwaarde als 30% van de fenotypische variantie door erfelijke aanleg verklaard wordt.

vergelijkingen worden opgelost voor melkproductie. Door de grote hoeveelheid data worden mixed model vergelijkingen vaak iteratief opgelost, omdat het berekenen van een directe inverse van de *left-hand side* vaak niet mogelijk is voor dergelijke grote systemen van vergelijkingen.

De nauwkeurigheid van de fokwaardeschatting hangt af van de hoeveelheid en type informatie die gebruikt wordt. Die nauwkeurigheid is gelijk aan de correlatie tussen de geschatte en echte fokwaarde. Deze nauwkeurigheid kan voorspeld worden voor een aantal situaties. Tabel 1 geeft aan wat de nauwkeurigheid is bij een aantal standaard situaties. Figuur 1 geeft weer hoe de nauwkeurigheid van een fokwaarde verandert als het aantal nakomelingen of (half)broers en (half)zussen met een fenotype toeneemt. Informatie van nakomelingen geeft de hoogste nauwkeurigheid, maar betekent dat lang gewacht moet worden met selectie, immers de nakomelingen moeten eerst het fenotype kunnen laten zien zoals melkproductie bij koeien.

Gebruik van DNA informatie

De vergelijkingen 1 en 2 gaan ervan uit dat er geen infor-

matie over de precieze DNA samenstelling van het dier is. In de laatste 5 jaar is het echter mogelijk geworden om voor elk dier op een vast aantal (vaak > 60.000) punten op het DNA het genotype te bepalen. Dit betekent een enorme hoeveelheid data, bijvoorbeeld 2000 dieren maal 60000 is al 120 miljoen datapunten! Deze DNA informatie kan gebruikt worden om fokwaardes nauwkeuriger te schatten en ook al voor dieren die net geboren zijn, dus voordat ze een eigen waarneming hebben aan het kenmerk. Als er van 2000 dieren genotypes en fenotypes beschikbaar zijn, is de nauwkeurigheid van de fokwaarde ongeveer 0,74 voor een kenmerk zoals melkproductie.

Selectie en paring

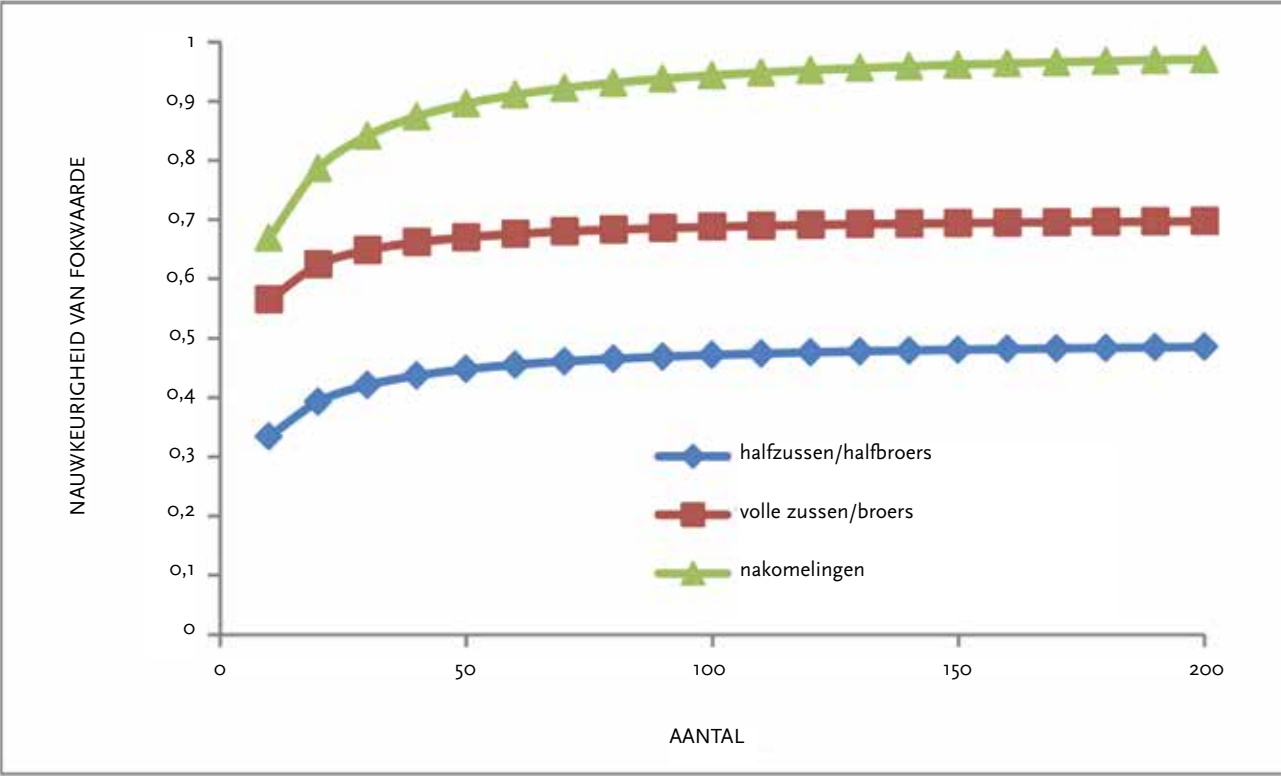
Wanneer de fokwaardes berekend zijn, kunnen de genetische beste dieren geselecteerd worden door de dieren op rangorde van fokwaarde te zetten. Vervolgens worden de beste mannelijke dieren en de beste vrouwelijke dieren met elkaar gepaard. De verwachting is dan dat de volgende generatie dieren beter presteert dan de huidige generatie dieren, bijvoorbeeld een hogere melkproductie.

Conclusie

Fokwaardeschatting is een belangrijk instrument om de erfelijk beste dieren te kunnen selecteren. Mixed model vergelijkingen worden gebruikt om van veel dieren fokwaardes te schatten. Het aantal op te lossen mixed model vergelijkingen loopt daarbij in de miljoenen. Door DNA informatie kan al van een dier op jonge leeftijd een redelijk betrouwbare schatting gedaan worden van zijn erfelijke aanleg. Gebruik van DNA informatie zorgt voor een verder verhoging van benodigde capaciteit.

LITERATUUR
Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* 3(10): e3395. doi:10.1371/journal.pone.0003395
Henderson, C. R. (1975). Best linear unbiased estimation and the prediction under a selection model. *Biometrics* 31(2), 423-447.

HAN MULDER is universitair docent Animal Breeding and Genomics Centre bij de Wageningen University.
E-mail: <han.mulder@wur.nl>



Figuur 1. De nauwkeurigheid van de geschatte fokwaarde als functie van het aantal halfzussen/broers, volle zussen/broers of nakomelingen als 30% van de fenotypische variantie door erfelijke aanleg verklaard wordt.



INFERENTIËLE STATISTIEK VERSUS MACHINE LEARNING een aanzet tot verzoening

RICHARD STARMANS & MARK J. VAN DER LAAN

Filosofen als Thomas Kuhn en Michel Foucault betoogden dat de ideeëngeschiedenis steevast cruciale gebeurtenissen of perioden van radicale omwentelingen kent, die de mens dwingen zijn eigen fundamentele aard, positie in de kosmos, verantwoordelijkheden en zelfbeeld opnieuw te bezien. Zij zagen het als hun taak deze cesuren aan te wijzen en te benoemen. In hun recente boek *Big Data: A Revolution that will Transform how we Live, Work and Think* (2013) scharen Viktor Mayer-Schönberger en Kenneth Cukier zich nadrukkelijk in deze traditie. De these dat de snelle opkomst van informatica en ICT zo'n cesuur markeert kent een in rap tempo toenemende schare van pleitbezorgers. Onmiskenbaar behoren beide Big-Data-apologeten hiertoe. De Italiaanse filosoof Luciano Floridi spreekt in dit verband van een Vierde Revolutie, in gang gezet door de Britse wiskundige Alan Turing. Eerst maakte de Copernicaanse Revolutie ons duidelijk dat wij niet onbeweeglijk in het centrum van het universum vertoeven. Daarna veroorzaakte de Darwiniaanse Revolutie een radicale breuk met onze vertrouwde opvattingen over de menselijke soort en afkomst. Vervolgens leerde de Freudiaanse Revolutie ons dat wij allerminst volledig transparant voor onszelf zijn. De ICT, aldus Floridi, doet ons beseffen dat we geen op zichzelf staande entiteiten zijn, maar zogenaamde informationele organismen (inforgs), die met andere (al dan niet artificiële) actoren een gemeenschappelijke leefwereld delen, welke uiteindelijk gemaakt is van informatie, de 'InfoSphere'.

Floridi's informationele metafysica mag voor menigeen een brug te ver zijn, het door de fysicus John Wheeler verwoorde adagium dat informatie fundamenteeler is dan materie (*it-from-bit*), weerklinkt ook in Mayer-Schönbergers boek. Met Big Data nadert de voornoemde revolutie zijn voltooiing en wordt de belofte die in Turings werk besloten lag eindelijk volledig ingelost.

Volledige kennis

Het fenomeen Big Data mag dan wetenschap, overheid en bedrijfsleven in zijn greep houden, conceptueel lijkt het verschijnsel noch goed gedefinieerd, noch erg nieuw. De keuze voor het onbepaalde hoofdtelwoord in 'veel data' impliceert een dynamisch aan te passen en verschuivende betekenis, afhankelijk van omstandigheden als persoon, voortschrijdende techniek, beschikbare opslagcapaciteit, rekenkracht en andere contingenties. Daarnaast zijn Big Data niets anders dan een uiting van het aloude menselijke streven alles vast te leggen, op te slaan, te analyseren en te interpreteren, teneinde kennis te verwerven, de natuur te beheersen, de Schepper naar de kroon te steken, etc. Dit ideaal van volledige kennis is zo oud als de uitvinding van het schrift en het numerieke stelsel, maar filosofisch gezien uiteraard problematisch, zo niet utopisch of incoherent. Kennis lijkt immers intrinsiek onvolledig in die zin dat het meeste van wat

het geval is c.q. waar is, ons niet bekend is en alle inspanningen ten spijt ook niet of nauwelijks te verkrijgen is en wellicht zelfs niet kenbaar is. We kunnen niet alles waarnemen, begrijpen, onthouden, afleiden, opsporen, aan de vergetelheid onttrekken, opslaan, integreren, etc. Het ideaal van volledige kennis negeert eindigheid en tijdelijkheid, de grenzen van de menselijke cognitie en conditie, alsmede het belang van *computational tractability*. Het is daarmee vanuit kennistheoretisch oogpunt ongrijpbaar, zo niet illusoir.

Door Big Data lijkt de gepostuleerde incoherentie van het hierboven geschetste kennisideaal echter te worden uitgedaagd. Enorme hoeveelheden zeer heterogene data komen continu en met grote snelheid beschikbaar. Door betere opslag, retrieval, rekenkracht, intelligente algoritmen en statistische technieken verschuiven de grenzen van het haalbare voortdurend. Daarbij lijkt het erop dat Big Data ook de aard van de kennis zelf transformeren: de wijze van kennisverwerving, onderzoeksmethodologie, aard en status van modellen en theorieën. In de beschouwingen over dit alles tekent zich dikwijls een tegenstelling af tussen enerzijds de wiskundige, inferentiële statistiek en anderzijds de computationele, op veel data en machine learning algoritmen gebaseerde data-analyse. Soms is deze wat grimmig of onverzoenlijk, leidend tot een annexatie van Big Data door een van beide disciplines. Ook de klaarblijkelijk weinig statistisch onderlegde Mayer-Schönberger maakt zich hieraan enigszins schuldig door een 'gedataficeerde' wereld te proclameren en te vereenzelvigen met een doorzoekbare wereld.

Statistiek versus machine learning

Gelet op het grote belang van data-analytische technieken lijkt zo'n controversie en bijbehorend debat over het (vermeende) primaat van één van beide onwenselijk; elke poging tot verzoening is dan de moeite waard. De vraag of de waarheid zichzelf openbaart of onthult in aanwezigheid van veel data, dan wel verborgen is in een achterliggende, slechts via sophisticated wiskundige inferentie toegankelijke werkelijkheid, is bovendien niet nieuw. Zij heeft de geschiedenis van de statistiek steevast parten gespeeld, lang voordat de computer haar intrede deed. Zij gaat zelfs terug tot de periode voordat Jakob Bernoulli in zijn precies 300 jaar geleden postuum verschenen *Ars Conjectandi* (1713) de empirische wet van de grote getallen formuleerde. In dit boek verschaftte hij als eerste enig inzicht in de vraag hoe de waarheid zich prijsgeeft in

aanwezigheid van grote aantallen gegevens, maar gaf hij tevens een opmaat tot de inferentiële statistiek! De meest heldere manifestatie van de tegenstelling treedt evenwel op eind 19e en begin 20e eeuw. Het was Karl Pearson die weliswaar de kansverdeling als vervanging voor de werkelijkheid introduceerde, maar deze plaatste in de verschijnselen; zij was waarneembaar, kenbaar met behulp van de door Francis Galton verzamelde Big Data. De 'bezielde' rekenkracht werd geleverd door zogenaamde compilatores, door Pearson betaalde vrouwen, die zijn beroemde vier parameters in een soort atelier moesten berekenen. Ronald Fisher daarentegen beschouwde de statistiek als random en veronderstelde dat deze ook een verdeling bezat. De data waren een realisatie van het mechanisme dat de data produceerde en een statistisch model moest de bijbehorende verdeling bevatten. De anti-causalistische, empiristische biometricus Pearson stond tegenover de causalistische Mendeliaan Fisher. Pearsoniaanse *goodness-of-fit* versus Fisheriaanse p-value's en Maximum Likelihood Estimation.

In deze aanzet tot verzoening beperken we ons hier tot een drietal aspecten, die typerend zijn voor de huidige praktijk van data-analyse en tevens geworteld zijn in de oude Pearson-Fisher controverse: status en betekenis van het modelbegrip, de noodzaak van een notie van causaliteit en de rehabilitatie van parameterschatting in de data-analyse. Vervolgens schetsen we kort hoe zo'n aanzet tot verzoening bewerkstelligd kan worden vanuit de in Van der Laan (2011) voorgestelde methodiek van Targeted Maximum Likelihood Estimation, gecombineerd met Super Learning algoritmen (TMLE/SL), hier kort aangeduid als Targeted Learning.

Targeted Learning

De huidige praktijk van statistische data-analyse leunt doorgaans zwaar op parametrische modellen en MLE als schattingsmethode. De zuiverheid van MLE wordt uiteraard bepaald door de correcte specificatie van het model. Een belangrijke assumptie hierbij is dat de kansverdeling die de data genereerde bekend is tot op een eindig aantal parameters. Schending van deze aanname en misspecificatie van het model kan leiden tot extreem onzuivere en moeilijk te interpreteren schatters, vaak coëfficiënten in een (logistisch) regressiemodel. Dit is niet te repareren door een grotere *sample size* of *Big Data*. Indachtig George Box' fameuze dictum dat '*Essentially, all models are wrong, but some are useful*', tekent zich echter ook in de statistiek een erosie van het modelbegrip af

en lijkt een klassiek waarheidbegrip obsoleet te worden. Modellen bevatten vaak aantoonbaar niet de (benadering van de) ware datagenererende verdeling en negeren de beschikbare realistische achtergrondkennis. De modellen moeten dus 'groter' worden, maar dat maakt MLE problematisch. Essentieel is hierbij het gegeven dat voornoemde maximum likelihood schatters doorgaans non-targeted zijn, terwijl voor het beantwoorden van nagenoeg elke denkbare onderzoeksvraag slechts een laagdimensionale target-parameter van de verdeling hoeft te worden geschat. Omdat bij een non-targeted benadering een evaluatiecriterium wordt gehanteerd dat gericht is op de fit van de gehele verdeling, wordt ook de error verspreid over de gehele verdeling. De MLE van de target parameter wordt dan onnodig onzuiver, vooral in hoog dimensionale datasets (zoals *micro-array data*) en/of data met duizenden potentiële covariaten of interactietermen. Hoe groter het statistische model, des te problematischer de non-targeted benadering.

Targeted Learning start met de specificatie van een non-parametrisch of semi-parametrisch model, dat alleen de realistische achtergrondkennis bevat en richt zich op de parameter of interest, die als een eigenschap van de vooralsnog onbekende, ware, datagenererende verdeling wordt beschouwd. Methodologisch heeft dit een duidelijke imperatief: model en parameter of interest moeten vooraf worden gespecificeerd. De (empirische) onderzoeksvraag moet worden vertaald in termen van de parameter of interest. Daarnaast wordt een rehabilitatie van het modelbegrip bewerkstelligd. Vervolgens behelst Targeted Learning een schattingsprocedure die op een data-adaptieve, flexibele manier in twee stappen verloopt. Eerst wordt een initiële schatting gezocht op basis van het relevante deel van de ware verdeling dat nodig is om de target parameter te evalueren. Deze initiële schatter wordt gevonden met behulp van het SL-algoritme. Kort gezegd is dit gebaseerd op een bibliotheek van vele uiteenlopende analytische technieken variërend van logistische regressie tot ensemble technieken, random forest en support vector machines. Omdat de keuze voor een van deze doorgaans subjectief is en de variatie in de resultaten van de diverse technieken meestal aanzienlijk is, wordt een soort gewogen som van de waarden berekend met behulp van cross-validation.

Op basis van deze initiële schatter kan vervolgens de tweede fase van de schattingsprocedure worden gestart, waarbij de initiële fit wordt geupdated met als doel een optimale bias-variance trade-off voor de parameter of interest. Dit wordt bewerkstelligd met een targeted maximum likelihood schatter van de fluctuatie parameter van

een speciaal geselecteerd parametrisch submodel door de initiële schatter. De statistische inferentie wordt dan voltooid door standard errors te berekenen op basis van *influence-curve theory* of resampling technieken. Daarmee behoudt parameter schatting een cruciale plaats in de data-analyse. Wil men recht doen aan variatie en verandering in de verschijnselen, dan kan men niet om Fishers onwrikbare inzicht heen dat randomness intrinsiek is en impliceert dat de schatter van de parameter of interest zelf een verdeling heeft. Big Data, ja zelfs census-onderzoek of andere pogingen de gehele werkelijkheid in de dataset te verdisconteren of te coderen, heffen dit niet op.

Tot slot kort aandacht voor de notie van causaliteit, die altijd een moeizame plaats innam in de statistiek, maar ironisch genoeg door Mayer-Schönberger hiermee nu wordt geassocieerd en uiteraard in aanwezigheid van Big Data als overbodig wordt beschouwd. (Correlaties zijn voldoende!) Het ervaren van oorzaak-gevolg relaties in de werkelijkheid is inherent aan de menselijke conditie en de meeste data-analyses behelzen effectstudies of hebben anderszins causale connotaties. De TMLE-parameter kan in het statistisch model statistisch worden geïnterpreteerd. Door toevoeging van niet testbare causale aannamen kan de parameter echter ook causaal worden geïnterpreteerd en worden verbonden met de schattingstheorie.

In dit korte exposé konden uiteraard slechts enkele facetten van de problematiek worden beschreven. We hopen te hebben beargumenteerd dat statistiek en machine learning in een gedatificeerde wereld veeleer twee zijden van dezelfde medaille vormen, dan dat zij met elkaar in tegenspraak zijn.

LITERATUUR

Van der Laan, M. J., Rose, S. (2011). *Targeted Learning, Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. New York, NY: Springer Publishing Company.

Mayer-Schönberger, V. & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt.

RICHARD STARMANS is verbonden aan de Faculteit Bètawetenschappen (Department of Information and Computing Sciences) van de Universiteit Utrecht. Hij doet onderzoek op het snijvlak van filosofie, statistiek en informatica.
E-mail: <starmans@cs.uu.nl>

MARK J. VAN DER LAAN is hoogleraar statistiek en biostatistiek aan de University of California at Berkeley.
E-mail: <laan@berkeley.edu>



Bewaren, delen en koppelen van data alleen na geïnformeerde toestemming?

MARCEL VERWEIJ

De moderne biomedische wetenschap en gezondheidszorg produceren enorme hoeveelheden data en biologisch materiaal. Onderzoeksprojecten, screeningsprogramma's en de gewone dagelijkse klinische praktijk vergen continu anamneses, vragenlijsten, testen, scans, bepalingen in bloed, urine of ontlasting. Die gegevens worden idealiter allemaal digitaal bewaard, al dan niet gekoppeld aan bijbehorende lichaamsmaterialen in het lab of in de koel- of vrieskist.

Die databestanden en biobanken kunnen op zich weer gebruikt worden voor nader onderzoek, nieuwe vraagstellingen of meta-analyses. Door veel instanties wordt dan ook benadrukt dat dergelijke bestanden zo veel mogelijk gedeeld worden in de onderzoekspraktijk, en natuurlijk ook dat verschillende databestanden aan elkaar gekoppeld worden. In Nederland is er bijvoorbeeld het Parelsnoer waarin de universitair medische

centra klinische data en biomaterialen bundelen om zo nieuw medisch en farmaceutisch onderzoek mogelijk te maken (Van 't Riet et al., 2012). Ook sponsors van wetenschappelijk onderzoek zoals de Wellcome Trust, onderschrijven het belang van het delen van data (Wellcome Trust, 2011).

De doelstelling is lovenswaardig, maar in de praktijk doen zich allerlei vragen voor over de vergelijkbaarheid van data en de haalbaarheid om bestanden echt breder voor onderzoek beschikbaar te maken. Veel kwesties en mogelijke obstakels hebben bovendien een ethische dimensie: Van wie zijn de data en wie moeten er toegang toe hebben? Wie zouden van nieuw onderzoek moeten profiteren? Hebben wetenschappers een professionele plicht om hun data breed toegankelijk te maken? En vooral ook: wat betekent het voor de individuen en groepen om wier gegevens of lichaamsmateriaal het gaat?

Privacygevoelig

Zeker die laatste vraag is van groot belang voor een verantwoorde praktijk. Medische gegevens beschouwen we nu eenmaal als ‘gevoelig’ – behorend tot de privacy van het individu. Zorgverleners behoren ze als vertrouwelijk te behandelen. Niemand wil dat zijn medisch dossier op straat komt te liggen. Misbruik zou de patiënt of de respondent in een onderzoek kunnen schaden. Hoe meer databestanden met persoonlijke informatie opengesteld worden voor nader onderzoek, en gedeeld worden met andere partijen, hoe groter de kans is dat, bedoeld of onbedoeld, data gebruikt worden op een manier die mogelijk ingaat tegen de belangen van individuen. In een complexe praktijk waarin veel verschillende mensen, soms vanuit verschillende organisaties, betrokken zijn, gaat onherroepelijk wel eens wat mis. De vraag is dan: hoe verhoudt bescherming van individuen zich tot bevorderen van onderzoek dat mogelijk belangrijk wetenschappelijk en maatschappelijk belang kan dienen?

Dit is een centrale vraag in de ethiek van medisch-wetenschappelijk onderzoek met mensen. De geschiedenis kent tal van uitwassen op dit terrein. Bekende voorbeelden zijn de medische en toxicologisch experimenten die in de Tweede Wereldoorlog op gevangenen in concentratiekampen werden uitgevoerd. Aangezien er in die tijd geen helder juridisch kader voor verantwoord medisch onderzoek bestond, formuleerden de rechters in de processen van Neurenberg regels voor medische experimenten met mensen. Door te eisen dat onderzoek alleen op basis van vrije toestemming van het onderzoekssubject mocht plaatsvinden, werd gehoopt dat individuen beter beschermd zouden zijn tegen schadelijke experimenten.

Maar desondanks werden ook daarna zowel in ontwikkelde als in ontwikkelingslanden talloze medische studies verricht, die wellicht vaak goedbedoeld waren, maar niettemin weinig respect toonden voor de proefpersonen (Beecher, 1966). In de laatste decennia is de praktijk van medische experimenten sterk gereguleerd, onder meer op basis van de Verklaring van Helsinki, waarbij de aandacht meer verschoven is van voorkomen van misbruik tot het opleggen en handhaven van criteria voor respectabel onderzoek. Vaak wordt daarbij gesteld dat het belang van de proefpersoon altijd zwaarder weegt dan dat van het wetenschappelijk of maatschappelijk belang. Toestemming van de proefpersoon of diens vertegenwoordiger, op basis van goede en volledige informatie over het onderzoek en de risico's, wordt als een welhaast absolute eis gezien. Goede informatie

en toestemming – soms in de vorm van gezamenlijke besluitvorming – is laatste decennia sowieso steeds meer gemeengoed geworden in de gezondheidszorg. Niet alleen ter voorkoming van misbruik, maar vooral ook uit respect voor de autonomie van de patiënt. En natuurlijk als bescherming (c.q. vrijwaring van aantijgingen) van de onderzoeker. Het onderzoeksvoorstel en de informatie- en toestemmingsprocedure moet bovendien vooraf beoordeeld worden door een medisch-ethische toetsingscommissie.

Het opslaan en later opnieuw gebruiken van data en biologisch materiaal, en het koppelen van databestanden, is echter van een totaal andere orde dan experimenteel onderzoek op individuen. Hoe deze steeds verder groeiende praktijk goed te reguleren is onderwerp van discussie in de ethiek van biomedisch onderzoek. Kan en moet het toestemmingsbeginsel net zo'n cruciale rol spelen als bij experimenteel onderzoek met patiënten of gezonde vrijwilligers? Patiënten hoeven van het beschikbaar maken en hergebruiken van data en materialen immers weinig te merken. Maar natuurlijk kan er van alles misgaan. Gegevens kunnen onbedoeld terecht komen bij de verkeerde mensen of instanties – al zal de eis van geïnformeerde toestemming daar op zich zelf niet tegen helpen. Er kunnen ook redenen zijn voor onderzoekers om naar de ‘deelnemers’ terug te gaan. Nieuwe bevindingen, die nader onderzoek of wellicht therapeutische interventies vergen. Mogelijk hadden die deelnemers daar liever niet mee geconfronteerd willen worden – of hadden ze daar in ieder geval daarover vooraf geïnformeerd willen worden. De toestemmingseis geeft hen dan meer controle over wat er met de gegevens zal gebeuren.

Biobanken

In de praktijk lijkt het me echter vrijwel onmogelijk om ook hier *geïnformeerde* toestemming als de centrale norm te zien. Het idee van die eis is dat individuen een vrije keuze kunnen maken op basis van goed inzicht in de consequenties van hun keuzen – in dit geval: goed inzicht in de aard van het onderzoek met databestanden en materiaal – en de mogelijke consequenties die dat ook voor hen als individu kan hebben. Het interessante aan het toegankelijk maken en koppelen van databestanden en biobanken is echter dat het nieuwe onderzoekswegen doet ontstaan die op het moment dat een individu materiaal ‘afstaat’ nog helemaal niet overzien kunnen worden. Op bloedmonsters kunnen nu bepalingen worden gedaan die een decennium geleden nog ondenkbaar waren. Hetzelfde

geldt voor de onderzoeksmogelijkheden die ontstaan als verschillende databestanden gekoppeld worden.

In de discussies over biobanken wordt daarom door velen gesteld dat een brede, globale toestemming kan volstaan. Bijvoorbeeld: toestemming dat persoonsgegevens en materialen alleen op een gecodeerde (en dus niet rechtstreeks herleidbare) wijze, maar wel voor een breed, niet nader gespecificeerd palet aan medisch-wetenschappelijk onderzoek gebruikt kunnen worden. Het probleem is dat dit moeilijk kan worden beschouwd als een geïnformeerde keuze die recht doet aan de autonomie van de patiënt of burger; het is als het ware een instemming met een *carte blanche*.

Aan de andere kant, als wél zoveel mogelijk – zij het in abstracto – wordt uitgelegd wat voor onderzoek wordt toegelaten en wat voor consequenties er in theorie voor individuen zouden zijn, worden zowel het begripsvermogen van de gemiddelde burger, alsook de communicatiemogelijkheden van de onderzoeker, al gauw overvraagd. De kans dat de informatie dan toch wordt genegeerd, is groot. Denk ook aan het gemak waarmee vrijwel iedereen akkoord gaat met de licentievoorwaarden van software op computer of smartphone. Geïnformeerde toestemming wordt zo al snel een wassen neus en daarom kan het niet de centrale spil zijn voor de regulering van dit onderzoek.

Als *echte* geïnformeerde toestemming een *conditio sine qua non* zou zijn voor de omgang met gegevens in gekoppelde databestanden en biobanken, zou die procedure en de positie van het individu flink versterkt moeten worden. Ik vraag me af of dat echter een juiste afweging biedt van de spanning tussen enerzijds het bevorderen van wetenschappelijk onderzoek en anderzijds het beschermen van de belangen van het individu. Over het algemeen zal de impact van onderzoek met data op het welzijn en de privacy van individuen nihil zijn, terwijl het wetenschappelijk belang heel groot kan zijn.

Er valt dan ook veel voor te zeggen om het toestemmingsaspect minder centraal te stellen in de ethiek van het bewaren en gebruiken van gegevens voor onderzoek. Dat betekent niet dat de burger maar gewoon de wetenschappelijke praktijk moet vertrouwen. De wetenschapper zit ook niet op blind vertrouwen te wachten.

Toetsing

Cruciaal voor een zorgvuldige onderzoekspraktijk is een robuust systeem voor de toetsing van onderzoek dat van gekoppelde databestanden gebruik wil maken. Die

medisch-ethische toetsing mag bijzonder strenge criteria volgen; gegeven dat individuen maar in beperkte mate vat kunnen hebben op ‘hun’ data moeten alleen projecten toegelaten worden waarvan het wetenschappelijk en maatschappelijk belang bijzonder groot is. Idealiter zijn niet alleen wetenschappers en juristen in die toetsing betrokken, maar vooral ook burgers die met enige distantie een redelijke afweging kunnen maken bij mogelijke conflicten tussen maatschappelijk belang en privacy van individuen.

Daarnaast is transparantie van groot belang: op welke wijze wordt getoetst wie met welk onderzoek toegang tot (naar personen herleidbare) data krijgt. Dat betekent vooral: heldere criteria voor nader gebruik van data, en openbare discussie over gevallen waar de criteria zich niet eenvoudig laten toepassen, of wanneer zich dilemma's voordoen. Maak bovendien naar de samenleving duidelijk voor wat voor projecten de data worden gebruikt en wat de belangrijkste opbrengsten zijn voor de medische praktijk.

In zo'n systeem heeft individuele toestemming – en dus de mogelijkheid van individuen om hun toestemming later in te trekken – nog steeds een plaats. Die toestemming is onvermijdelijk globaal: instemming voor het gebruik van persoonsgegevens en materialen voor uiteenlopende onderzoeksdoeleinden. Toestemming is dan geen zaak van ‘respect voor autonomie’ of een instrument om de burger te beschermen – die functies kan brede toestemming niet vervullen. Wel kan het een basis bieden voor vertrouwen in de onderzoekspraktijk. Een vertrouwen dat die praktijk vervolgens met goed onderzoek, kritische toetsing, en openbare discussie over lastige kwesties voortdurend moet waarmaken. De toestemming kan immers worden ingetrokken.

LITERATUUR

- Van 't Riet, E., Schram, M. T., Abbink, E. J., Admiraal, W. M., Dijk-Schaap, M. W., Holleman, F., Dekker, J. M. (2012). *The Diabetes Pearl: Diabetes biobanking in The Netherlands*. BMC Public Health 12:949, <www.biomedcentral.com/1471-2458/12/949>, doi:10.1186/1471-2458-12-949 Wellcome Trust (2011). *Sharing research data to improve public health: joint statement of purpose*. <www.wellcome.ac.uk/publichealthdata>.
- Beecher, H. K. (1966). Ethics and Clinical Research. *New England Journal of Medicine*, 274(24), 1354-1360.

MARCEL VERWEIJ is hoogleraar Filosofie, Wageningen University. Hij onderzoekt ethische en filosofische vragen inzake de relatie tussen maatschappelijke gezondheidszorg, voeding, milieu en landbouw.
E-mail: <marcel.verweij@wur.nl>

ANALYSES VAN GENEXPRESSIES BIJ KANKERONDERZOEK

KEES VAN MONTFORT

Biostatistiek speelt een belangrijke rol bij de analyse van grote gegevensbestanden die worden verkregen door nieuwe technologische ontwikkelingen. Beeldvormende en moleculaire technieken genereren grote hoeveelheden gegevens. Met behulp van fluorescentie, analyses van snp's, GWA scans en analyses van genexpressies kan inzicht worden verkregen in de ketens van onze genen en de ontwikkeling van ziekten. Dit kan de mogelijkheid geven om te voorspellen welke behandeling voor welke patiënt het meest geschikt is op basis van individuele karakteristieken van de patiënt en de ziekte.

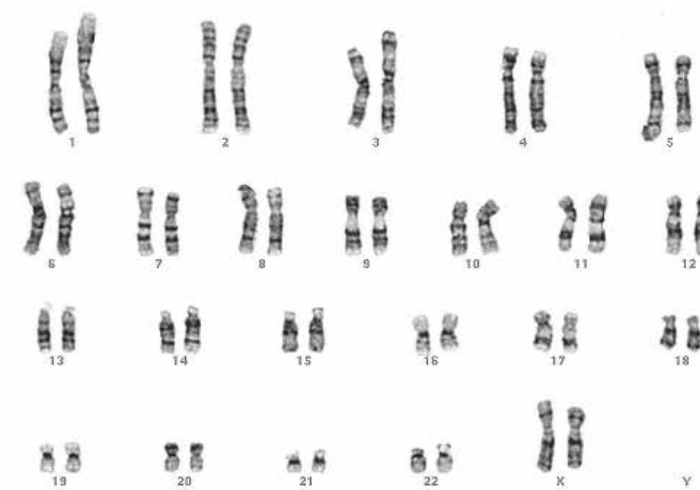
De gegenereerde gen-data leveren statistische problemen op. Het aantal patiënten is doorgaans extreem veel kleiner dan de hoeveelheid gemeten variabelen. 100 patiënten met 10.000 genetische of beeldvormende markers zijn statistisch moeilijk te analyseren. Deze markers kunnen bovendien complexe afhankelijkheden vertonen.

Figuur 1 toont de 23 chromosomenparen van een mens. Een chromosoom is opgebouwd uit allerlei verschillende genen en eiwitten. Naar schatting zijn er ongeveer 26.000 genen mogelijk. Met behulp van de statistiek kan worden gekeken of bepaalde genen met een bepaald ziektebeeld corresponderen. Bij dit soort onderzoek worden steeds grotere databestanden aangelegd. Bij zeer grote databestanden is het kleinste verschil echter al statistisch significant. Dit bemoeilijkt de interpretatie van de statistische analyses. De vraag is natuurlijk of de statistische significantie ook klinisch relevant is. Ten gevolge van het feit dat bij zeer grote databestanden het kleinste verschil al statistisch significant wordt, zijn er inmiddels talloze wetenschappelijke publicaties met praktisch onbruikbare conclusies en aanbevelingen verschenen.

Afgelopen decennia is er veel tijd, energie en geld in het verzamelen en analyseren van genexpressies gestoken. Genexpressies geven informatie over de aard en het stadium van kankertumoren. Dit kan helpen bij het kiezen van een behandeling of het voorspellen van overlevingskansen. Binnen het kankeronderzoek heeft de 'grote doorbraak' echter nog niet plaatsgevonden.

De huidige opzet van het kankeronderzoek spoort vooral effecten van individuele genen op. De statistische/methodologische uitdaging ligt in een accurate beschrijving van de onderliggende relatie tussen de gen-expressie en het 'klinische eindpunt', zoals de prognose of response met betrekking tot een specifieke behandeling van een individuele patiënt. Afhankelijk van het klinische eindpunt kan de relatie worden beschreven met behulp van logistische regressie (bij respons-eindpunt) of Cox-regressie (bij survival-eindpunt). De genexpressie data bevatten doorgaans echter een groot aantal variabelen (gene probes) voor een relatief klein aantal patiënten. Het probleem bij zulke *high-dimensional* data is het gevaar van overfitting. Zoals hierboven reeds is vermeld zullen vanwege het grote aantal *explanatory* variabelen, diverse *explanatory* variabelen statistisch significant zijn gerelateerd aan het klinische eindpunt, zelfs als er feitelijk geen relatie bestaat.

Bovengenoemde univariate (oftewel: gen-na-gen) regressie aanpak wordt vooral toegepast wanneer de onderzoekers zijn geïnteresseerd in het ontdekken van individuele genexpressies met afwijkende responsen of prognoses van patiënten. Het kan eveneens relevant zijn om het gezamenlijke effect van verschillende genen op het klinische eindpunt te detecteren. Een voorbeeld hiervan is de evaluatie van de prognoses van patiënten gebruikmakend van de volledige (of een deelset van de)



Figuur 1. De 23 chromosomenparen van een mens.

genexpressie profiles. Dit vereist de simultane analyse van een groot aantal genexpressie profiles, welke niet kunnen worden geanalyseerd met de standaard regressiemethoden. Een mogelijke aanpak voor dit probleem is een verkleining van de dimensie van de data. Zo kan een deelset van de genen worden geselecteerd met behulp van de univariate gen-na-gen aanpak (Riaz et al., 2012), of met behulp van de toepassing van principal-component-analyse (Kuiper et al., 2012) of cluster-analyse (Alizadeh et al., 2000; Sieuwerts et al., 2011). Andere aanpakken zijn gebaseerd op de analyse van de volledige set van genen met behulp van *penalised Cox-regression* (Pawitan et al., 2004; Van Houwelingen et al., 2006).

Cluster-analyse algoritmen groeperen objecten in clusters, zodat de *within-cluster similarity* zo groot mogelijk en de *between-cluster similarity* zo klein mogelijk wordt. Gegeven een tweedimensionale genexpressie matrix met m rijen en n kolommen, waarbij de n kolommen de samples bevatten, en elke sample genexpressie data voor m probes bevat, zal een cluster analyse methode of de rijen of de kolommen clusteren. Het heeft echter de voorkeur om de rijen en kolommen van de data matrix simultaan (in plaats van gescheiden) te groeperen. Deze aanpak heet biclustering (Yang et al., 2013). Het gebruik van biclustering kan nuttig zijn bij de analyse van genexpressies, die alleen zijn gecorreleerd in een deelset van de samples. Dergelijke cluster zijn medisch-biologisch interessant aangezien zij ons in staat stellen om eveneens genen te identificeren die zich onder alle condities hetzelfde gedragen.

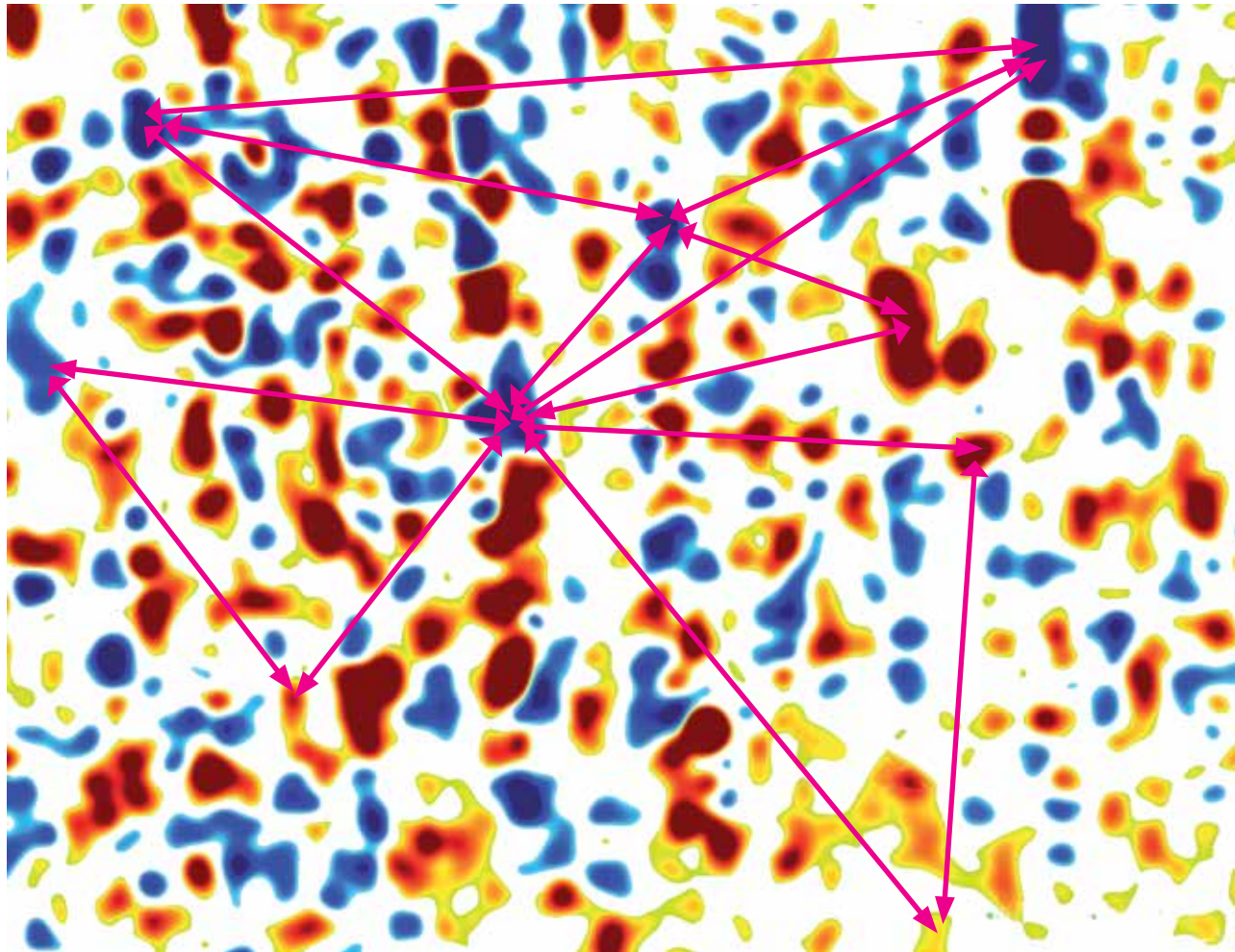
Tot nu toe zijn genexpressie-data in het kankeronderzoek nog niet geanalyseerd met behulp van biclustering. Bij het Erasmus Medisch Centrum te Rotterdam zal daar de komende jaren energie in worden gestoken. De mees-

te biclustering problemen zijn echter exponentieel in de rijen en kolommen van de genexpressie data matrix. In feite is er nog geen consensus over het antwoord op de vraag welke biclustering aanpak het meest veelbelovend is. De volgende artikelen geven een beeld van de verschillende biclustering aanpakken: Cheng & Church, 2000; Eren et al., 2012; An et al., 2012; Yang et al., 2013. Toekomstig onderzoek moet meer duidelijkheid geven over welke biclusteringmethoden de beste resultaten geven.

LITERATUUR

- Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., ... Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503-511.
- An, J., Liew, A. W. C., Nelson, C. C. (2012). Seed based biclustering of gene expression data. *PLoS One* 7(8).
- Cheng, Y., Church, G. M. (2000). Biclustering of expression data. *Proceedings of the International Conference of Intelligent Systems for Molecular Biology*, 8, 93-103.
- Eren, K., Deveci, M., Kucuktunc, O., Catalyurek, U. V. (2013). A comparative analysis of biclustering algorithms for gene expression data. *Brief Bioinform*, 14(3), 279-292. doi: 10.1093/bib/bbs032. Epub 6 juli 2012.
- Kuiper, R., Broyl, A., De Knecht, Y., Van Vliet, M. H., Van Beers, E. H., Van der Holt, B., ... Sonneveld, P. (2012). A gene expression signature for high risk multiple myeloma. *Leukemia* 26, 2406-2413.
- Pawitan, Y., Bjohle, J., Wedren, S., Humphreys, K., Skoog, L., Huang, F., ... Bergh, J. (2004). Gene expression profiling for prognosis using Cox regression. *Statistics in Medicine* 23(11), 1767-1780.
- Riaz, M., Sieuwerts, A. M., Look, M. P., Timmermans, M. A., Smid, M., Foekens, J. A., Martens, J. W. (2012). High TWIST1 mRNA expression is associated with poor prognosis in lymph node-negative and estrogen receptor-positive human breast cancer and is co-expressed with stromal as well as ECM related genes. *Breast Cancer Research* 14(5).
- Sieuwerts, A. M., Mostert, B., Bolt-deVries, J., et al. (2011). mRNA and microRNA expression profiles in circulating tumor cells and primary tumors of metastatic breast cancer patients. *Clinical Cancer Research* 17(11), 3600-3618.
- Van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., Van't Veet, L. J., Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine* 18, 3201-3216.
- Yang, J., Wang, H., Wang, W., Yu, P. Enhanced biclustering on expression data. 2013 to be published.
- Zhang, S., Wang, K., Chen, B., Huang, X. (2011). A new framework for co-clustering of gene expression data. In M. Loog et al. (Eds), *Pattern recognition in bioinformatics*, Heidelberg: Springer Verlag.

KEES VAN MONTFORT is hoogleraar Biostatistiek bij het Erasmus Medisch Centrum te Rotterdam.
E-mail: <c.vanmontfort@erasmusmc.nl>



OPTIMALISEREN MET BIG DATA

hoe door de scalability wall heen te breken?

Het integreren van local search technieken in een mathematische programmerings-solver

THIERRY BENOIST, JULIEN DARLAY, BERTRAND ESTELLON, FRÉDÉRIC GARDI & ROMAIN MEGEL

Er bestaat geen twijfel over dat een van de meest krachtige gereedschappen van de Operationele Research (OR) het gemengde geheeltallige lineaire programmeren oftewel Mixed Integer Linear Programming (MIP), is. Het eenvoudige gebruik ervan zorgt ervoor dat OR-specialisten hier vaak een beroep op doen: de gebruiker modelleert het probleem als een geheeltallig lineair model en het MIP-programma lost dit op door middel van enu-

meratie en snijvlakgeneratie technieken. Deze 'modelleer en voer uit' benadering kan de inspanningen voor de ontwikkeling en het onderhoud van optimalisatie-software aanzienlijk verminderen. Andere zoekboomgerelateerde technologieën zoals Constraint Programming (CP) volgen dezelfde weg.

Echter, ondanks de opmerkelijke progressie in de afgelopen vijftien jaar (Bixby, 2012), lukt het MIP-solvers

niet om veel grootschalige combinatorische problemen uit de praktijk van de OR aan te kunnen. Het *car sequencing* probleem (Estellon et al., 2008), een klassiek probleem in de OR-literatuur, is een goed voorbeeld. Bij dit probleem lukt het MIP-solvers tot op heden niet om toegelaten oplossingen van problemen te vinden, zelfs al hebben ze slechts duizenden binaire beslissingen. Bovendien, ook al zouden ze toegelaten oplossingen hebben gevonden dan is de lineaire relaxatie toch niet goed genoeg om de exponentiële *branch-and-bound tree* op een efficiënte manier te verkleinen. De zoektocht levert ook na uren van berekenen geen toegelaten oplossing. Dit is wat wij de *scalability wall* noemen.

Als we grootschalige problemen willen aanpakken, is men strikt gesproken zelden geïnteresseerd in het vinden van de optimale oplossing. Ook zal men beseffen dat het bewijzen van het optimum een utopie is. Bovendien zijn kwalitatief hoogstaande oplossingen over het algemeen gesproken voldoende, vooral wanneer de wiskundige modellen slechts een benadering geven van de alledaagse situatie. Ten slotte, wat wij observeren als OR-beoefenaren, is dat vandaag de dag eindgebruikers steeds meer interactie willen met de beslissingsondersteunende systemen: ze willen snel kwalitatief goede oplossingen.

Rest de vraag wat OR-professionals doen als MIP- of CP-solvers ineffectief zijn? Normaal gesproken implementeren ze een specifieke *local search* heuristiek. In tegenstelling tot bovengenoemde zoekboom technieken, bestaat Local Search (LS) uit het iteratief toepassen van veranderingen aan een oplossing (stappen genoemd) teneinde de doelfunctie te verbeteren. Hoewel deze techniek incompleet is, wordt zij alom gewaardeerd omdat het de gebruikers toestaat om kwalitatief goede oplossingen te verkrijgen in korte rekentijden (in de orde van minuten voor zeer grootschalige instanties. Het ontwikkelen en implementeren van local search algoritmes is echter niet eenvoudig, zelfs niet als er al enkele raamwerken zijn ontwikkeld om de programmeur te helpen in deze taak. De algoritmische laag is vooral moeilijk te ontwikkelen omdat het een expertise in algoritmes én een behendigheid in programmeren vereist (zie Aarts en Lenstra (1997) voor een overzicht van het LS paradigma en zijn applicaties).

Deze situatie heeft geleid tot de ontwikkeling van LocalSolver¹, een mathematische programmeringssolver gebaseerd op local search. Het doel van het project, dat in 2007 is gestart, was het combineren van de eenvoud van het gebruik van een 'modelleer-en-voer uit' solver en de kracht van local search technieken

voor combinatorische optimalisering. We zullen laten zien hoe het OR-beoefenaren toestaat zich te focussen op de modellering van het probleem, terwijl er gebruik wordt gemaakt van een eenvoudig formalisme, en daarna de actuele vóór-oplossing over te dragen aan een solver die gebaseerd is op efficiënte en betrouwbare local search technieken.

Modellering

LocalSolver bevat een innovatieve wiskundig gemodelleerde scripttaal voor snelle prototypen, namelijk LSP (Local Search Programming). We zullen deze taal gebruiken in onze voorbeelden hoewel sommige object-georiënteerde API's ook beschikbaar zijn voor volledige integratie (C++, Java, .NET). LocalSolvers modellering-formalisme lijkt op klassieke wiskundige programmeringsformalismes maar is daarnaast verrijkt met algemene wiskundige operatoren. Deze staan een natuurlijke modellering van reële problemen toe en maken het daardoor eenvoudiger voor OR-deskundigen. Onderstaande regels definiëren bijvoorbeeld het model van het rugzakprobleem (*knapsack problem*) met n objecten waarvan het gewicht en de waarde is gegeven.

```
function model() {
  for [i in 1..n] x[i] <- bool();
  knapsackWeight <- sum[i in 1..n](weights[i] * x[i]);
  constraint knapsackWeight <= knapsackBound;

  knapsackValue <- sum[i in 1..n](values[i] * x[i]);
  maximize knapsackValue; }
```

In dit basisvoorbeeld zijn binaire beslissingsvariabelen $x[i]$ geïntroduceerd met het `bool()`-statement. Vervolgens is het gewicht in de *knapsack* geïntroduceerd als een som-expressie en is er een voorwaarde gesteld aan deze waarde. Tenslotte is de waarde in de *knapsack* gedefinieerd als een maximalisatie-doel. Merk op dat verschillende doelfuncties toegevoegd kunnen worden, welke kunnen worden geïnterpreteerd als een lexicografische doelfunctie. Het cruciale punt hier is dat er verder niets meer hoeft te worden gedefiniëerd en vooral dat er geen oplossingsomgevingen hoeven worden meegegeven. Alleen dit model wordt meegegeven aan de solver, verder niets. Het is de verantwoordelijkheid van de solver om een search algoritme toe te passen dat kan werken aan de abstracte combinatorische structuur welke is geïnduceerd door het model van de gebruiker. In eerste

ARITHMETIC	LOGICAL	RELATIONAL
sum, min, max, prod, div, mod, log, exp, pow, sin, cos, tan, abs, dist, sqrt, floor, ceil, round	and, or, not, xor, if, at	==, !=, <=, >=, <, >

Tabel 1. Mathematische operatoren beschikbaar in LocalSolver.

instantie voor het vinden van een toegelaten oplossing en vervolgens voor het iteratief verbeteren van deze oplossing. De belangrijkste principes van deze stappen worden gegeven in de volgende sectie.

Hoewel dit eenvoudige *knapsack*-voorbeeld alleen gebruik maakt van lineaire expressies, staan de onderliggende oplostechieken het gebruik toe van nonlineaire operatoren, inclusief conditionele expressies (*als A, dan B of C*, geschreven als $A \rightarrow B : C$) of zelfs *array indexing* (de expressie $A[N]$ staat voor het *N-de element in array A*) toe. Tabel 1 geeft een lijst van beschikbare operatoren. Het introduceren van logische, rekenkundige of relationele operatoren heeft twee belangrijke voordelen in een local search verband: expressiviteit en doeltreffendheid. Met dergelijke *low-level* operatoren is modeleren gemakkelijker dan met de basis MIP syntax, terwijl het behouden van de basis syntax zorgt voor een snellere acceptatie door beginners (vooral voor degenen die niet vertrouwd zijn met het computerprogrammeren). Bovendien kunnen de invarianten die geïnduceerd zijn door deze operatoren benut worden door de interne algoritmes van de LS solver om daarmee de local search te versnellen.

Beschouw bijvoorbeeld het *P*-mediaan probleem (Beasley, 1985), dat bestaat uit het selecteren van een deelverzameling S van P steden uit N (voor bijvoorbeeld het vestigen van openbare voorzieningen), zo dat de som van de afstanden van elke stad naar elke dichtstbijzijnde stad in S geminimaliseerd wordt. In onderstaand model laat het gebruik van conditionele en min-operatoren toe dat het model bijna letterlijk geformuleerd kan worden als hierboven. Deze eenvoud levert een model op dat alleen gefocust is op de relevante beslissingsvariabelen, namelijk het selecteren of niet selecteren van elke stad $x[i]$. Na het gelijkstellen van de som van deze variabelen aan P , wordt de $\text{minDistance}[i]$ van elke stad i naar de dichtstbijzijnde stad in S geschreven als het minimum van de afstanden naar de andere steden. De afstand naar steden buiten S worden beschouwd als

oneindig (InfiniteDistance kan worden gesteld als maximumwaarde in de afstandsmatrix). De doelfunctie zal de som zijn van deze $\text{minDistance}[i]$.

Model *P*-median probleem

```
function model() {
  x[1..N] <- bool();
  constraint sum[i in 1..N](x[i]) == P;

  minDistance[i in 1..N] <- min[j in 1..N]
    (x[j] ? distance[i][j] : InfiniteDistance);
  minimize sum[i in 1..N] (minDistance[i]); }
```

Alles wat de solver moet doen is in enkele seconden een kwalitatief goede oplossing vinden van deze probleemstelling (een gemiddeld verschil van 0,6% van de 40 instanties van de OR-Library met een tijdslimiet van één minuut). In de volgende sectie zullen we het interne mechanisme beschrijven dat dit mogelijk maakt.

Oplossen

Onze benadering van een autonoom Local Search werd bepaald door het volgende fundamentele principe: *de LS solver moet werken zoals een LS beoefenaar werkt*. Dit impliceert dat LocalSolver gestructureerde stappen uitvoert om te zorgen dat de haalbaarheid van de oplossingen bij elke iteratie gehandhaafd blijft, waarbij de evaluatie wordt versneld door het gebruik van invarianten die feitelijk zijn geïnduceerd door de structuur van het model. De sleutelcomponenten voor de efficiëntie van het algoritme zijn: een incrementele machine die in staat is om zeer snel de impact van een transformatie van de oplossing te evalueren; meervoudige autonome stappen die de combinatorische structuur van het model gebruiken om *toegelaten* omgevingen van een oplossing te ontdekken; een globale adaptieve zoekstrategie die de zoekopdracht leidt naar kwalitatief goede oplossingen.

De incrementele machine is gebaseerd op een representatie van het model als een directe a-cyclische graaf (DAG), waarvan de wortels de beslissingen zijn en de bladeren de beperkingen en de doelstellingen. De interne knopen van deze DAG zijn de operatoren uit tabel 1. Het toevoegen van stappen aan een bestaande oplossing bestaat uit het wijzigen van de waarden van de beslissingen (wortels) en het evalueren van de beperkingen en doelstellingen (bladeren) door het verspreiden van deze wijzigingen in de DAG. Door een bijzonder geoptimaliseerd algoritme kunnen er miljoenen stappen per minuut worden berekend in *real-life* modellen.

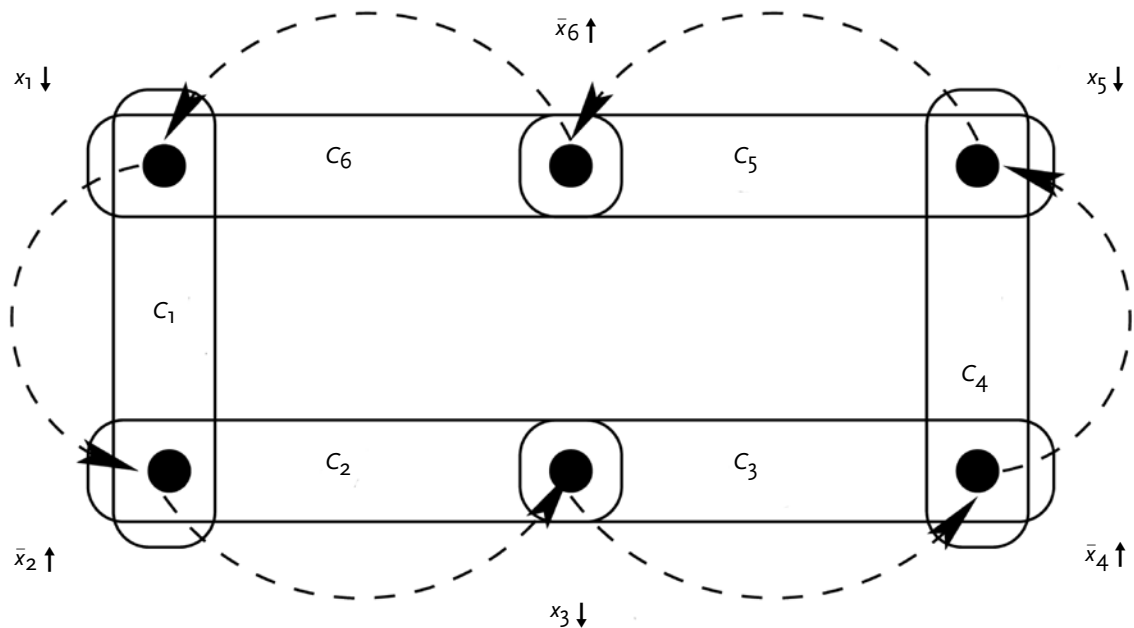
De meest eenvoudige generieke stap is de verandering van de waarde van een beslissing, zoals bijvoorbeeld het veranderen van de waarde van een $x[i]$ beslissing in ons *P*-mediaan voorbeeld. Echter, als we van één toegelaten oplossing naar een andere willen gaan, zoals in de meeste local search algoritmes, zijn er meer gestructureerde stappen nodig. De solver construeert dergelijke stappen op grond van het analyseren van de structuur van het model. Bijvoorbeeld, de kardinaliteits-voorwaarde in het *P*-mediaan probleem wordt gebruikt voor het ontwerpen van stappen die het aantal geselecteerde steden constant houdt en daarmee de toelaatbaarheid van de oplossing waarborgt. Meer algemeen, het volgen van *ejection chains* of *ejection cycles* in de *constraint hypergraph* (figuur 1) levert

zeer krachtige stappen op. Naast deze kleine omgevingen, kunnen grotere omgevingen worden verkend door het combineren van kleinere omgevingen binnen een verwerp & repareer mechanisme. Tenslotte kan de identificatie van speciale combinatorische structuren de activatie van specifieke omgevingen op gang brengen. Merk daarbij op dat deze omgevingen op verschillende manieren kunnen worden verkend: vrijwel willekeurig of door zich te richten op stappen met een grotere kans op succes.

De verkenning van de zoekruimte is verdeeld langs verschillende lijnen die periodiek onderling worden gesynchroniseerd. De globale verandering van de zoekopdracht wordt gewaarborgd door simulated annealing met *reheating* en herstart mechanismen. Statistieken over de prestaties van de stappen worden benut om de algehele prestaties van de zoekopdracht te verbeteren.

Testen en conclusies

Ondanks hun schijnbare conceptuele eenvoud leveren de uitgangspunten van de voorgaande sectie opmerkelijk goede resultaten in de praktijk, vooral voor grootschalige combinatorische optimaliseringsproblemen die buiten het gebied van de state-of-the-art wiskundige programmerings solvers vallen. Bij een aantal van de moeilijkste



Figuur 1. Een *ejection cycle* met zes boolean variabelen x_1, x_3, x_5 (waarvan de huidige waarde 1 is) en $\bar{x}_2, \bar{x}_4, \bar{x}_6$ (waarvan de huidige waarde 0 is), en zes beperkende sommaties $C_1, \dots, C_6$. Elke variabele behoort tot twee sommaties (bijvoorbeeld, x_1 behoort tot C_1 en C_6). Nu geldt dat x_1, x_3, x_5 zijn afgenomen ($\downarrow$), terwijl $\bar{x}_2, \bar{x}_4, \bar{x}_6$ zijn toegenomen ($\uparrow$). Deze stap bewaart de waarden van de sommaties, en dus de toegelatenheid van de voorwaarden.

MIPLIB-voorbeelden², overtrof LocalSolver onlangs de beste MIP-solvers. Een voorbeeld is het *car sequency* probleem³, dat bestaat uit het plannen van auto's naar verf- en assemblagelijnen rekening houdend met sequency voorwaarden. De oplossing van LocalSolver was na 10 seconden beter dan die van de beste MIP-solvers in één dag rekentijd.

Het vermogen om grootschalige combinatorische problemen aan te pakken op deze model-and-run wijze werd geïllustreerd tijdens de Roade/Euro Challenge 2012⁴, waar de opdracht was om de processen op Google-servers te herschikken, rekening houdend met verschillende bronnen en afhankelijkheidsvoorwaarden. Als 25ste gerangschikt van de 82 deelnemende teams was LocalSolver de enige model-and-run, general-purpose solver die zich kwalificeerde voor de finaleronde. Het team van LocalSolver paste een honderd regels bevattend model toe dat binnen één dag tijd was geschreven. Meer resultaten op zowel academische als industriële problemen kunnen worden gevonden in Benoist et al. (2011) of op de website van LocalSolver.⁵

In 2012 was LocalSolver gereed om van een onderzoeksproject over te gaan naar een commercieel product. Heden te dage wordt het in verschillende industrieën over de hele wereld gebruikt, van het maximaliseren van televisiereclame-inkomsten in Frankrijk tot het optimaliseren van de toelevering van een bakkerij in Japan (een niet-lineair probleem met drie miljoen 0-1 beslissingen!). LocalSolver ontwikkelt zich in de richting van een volledige, alles-in-een programmerings solver met alle optimalisatietechnieken (Mixed-integer en niet-lineaire programmeringstechnieken, constraint programming en satisfiability technieken, local en direct search technieken). De volgende versie 4.0, die gepland staat voor het eind van 2013, zal de eerste stap zijn in de richting van deze nieuwe generatie wiskunde programmerings solvers voor grootschalige mixed-variable niet-convexe optimalisatie. Deze nieuwe versie zal verscheidene nieuwe belangrijke functies bieden vanuit functioneel en technisch oogpunt: kleine omgeving stappen om te optimaliseren over continue of gemengde beslissingen; verkenning van grote, exponentieel grote omgevingen bij 0-1 of gemengde beslissingen gebruikmakend van enkele zoekboomtechnieken (bijvoorbeeld afrondingsheuristieken op basis van lineaire relaxatie); verkenning van grote omgevingen bij continue beslissingen door het herzien van opeenvolgende lineaire programmeertechnieken voor niet-lineair programmeren (gebaseerd op een simplexalgoritme); berekening van ondergrenzen gecombineerd met voorwaarden propagatie en duale lineaire relaxatie.

NOTEN

1. www.localsolver.com
2. www.localsolver.com/misc/LocalSolver_ECCO_2013.pdf
3. www.challenge.roade.org/2005/en
4. <http://localsolver.com/new.html?id=3>
5. www.localsolver.com

LITERATUUR

- Aarts, E., Lenstra, J. (1997). *Local search in combinatorial optimization*. John Wiley & Sons.
- Bixby, R. (2012). A brief history of linear and mixed-integer programming computation. In M. Grötschel (Ed.), *Optimization Stories, 21st International Symposium on Mathematical Programming, Berlin, August 19-24, 2012*, (pp. 107-121). Bielefeld: Documenta Mathematica.
- Beasley, J. (1985). A note on solving large p-median problems. *European Journal of Operational Research*, 21(2), 270-273.
- Benoist, T., Estellon, B., Gardi, F., Megel, R., Nouioua, K. (2011). Localsolver 1.x: a black-box local-search solver for 0-1 programming. *4OR-Quarterly Journal of Operations Research* 9(3), 299-316.
- Estellon, B., Gardi, F., Nouioua, K. (2008). *Two local search approaches for solving real-life car sequencing problems*. *European Journal of Operational Research*, 191(3), 928-944

THIERRY BENOIST, JULIEN DARLAY, BERTRAND ESTELLON, FRÉDÉRIC GARDI & ROMAIN MEGEL zijn werkzaam bij LocalSolver te Parijs.
E-mail: <contact@localsolver.com>



The best of all optimization technologies in one math programming solver

- Solve highly nonlinear models
- High-quality solutions in seconds
- Scale up to millions of variables
- Unique neighborhood search techniques
- Innovative math modeling language
- Lightweight APIs for C++, Java, .NET
- Simple and competitive pricing
- Free licenses for academics

<http://www.localsolver.com> • contact@localsolver.com
+33 1 44 20 11 06 • 24 avenue Hoche, 75008 Paris, France

JOHAN VAN LEEUWAARDEN

column



Wiskunde is à la mode

In 2011 sprak ik bij de finale van de Nederlandse Wiskunde Olympiade. Van de ruim 5.000 leerlingen die meededen aan de voorronde kwamen op die dag de beste 150 naar de finaleronde om zich te kwalificeren voor het Nederlands team. Tijdens de wedstrijd gaf ik mijn presentatie. Niet voor de finalisten, die waren driftig in de weer met sommen, maar voor hun ouders.

De meeste ouders keken aanvankelijk wat angstig, maar raakten gaandeweg geboeid, zeker toen ik na wat formules aangaf dat wiskunde ook nuttig kan zijn, en dat de maatschappij wiskundigen hard nodig heeft. Ik begriep de ouders ook wel. Je zult maar een kind hebben dat wiskunde leuk vindt, en zelfs overweegt het te gaan studeren. Wat zeg je tegen zo'n kind, als ouder? Zijn er geen andere studies, waar je later wat meer mee kunt? Ik ben geneigd om nee te zeggen.

Wiskunde is niet stoer in Nederland. Nog steeds laten velen op feestjes geen kans onbenut om te zeggen hoe slecht ze wel niet in wiskunde waren. Vreemd, want met onze sportprestaties doen we in de regel het tegenovergestelde. Het tij lijkt echter wel te keren. De laatste jaren gaan meer scholieren wiskunde studeren. In 2006 nog lag de jaarlijkse instroom in Nederland onder de 200, maar sindsdien is er een duidelijke opwaartse trend zichtbaar. De instroom is sinds 2006 grofweg verdrievoudigd. Wiskunde wordt steeds populairder.

Deze ontwikkeling laat zich moeilijk verklaren. Ligt het aan de toegenomen media-aandacht, of betere voorlichting op de middelbare school? Dat zal vast helpen. Maar ik denk dat het bovenal een natuurlijke ontwikkeling is. Een kwestie van vraag en aanbod. Uit een onderzoek van *Elsevier* van deze zomer bleek dat wiskundigen vrijwel meteen een vaste baan vinden, in de huidige arbeidsmarkt vrij zeldzaam. De maatschappij – bedrijfsleven, onderwijs en wetenschap – ontvangt je met open armen.

Goed om te weten, en ook wel stoer, zou je zeggen. Maar zie hier de spagaat van de wiskundige. Enerzijds is er het verlangen naar die abstracte wereld, met de uitdagende som en de prachtige formule, en anderzijds is er de roep uit de echte wereld van de toepassing. En toch kan dat samengaan. Als wij voorlichting geven aan middelbare scholieren, of beter nog, aan hun ouders, dan kunnen we eerst een prachtige stelling bewijzen, dan een maatschappelijk relevante toepassing laten zien, en op de laatste slide de harde carrièrecijfers tonen. Daarmee zeggen we niets te veel, en kiest een scholier voor een wiskundeopleiding, dan kunnen we de drie elementen van de presentatie ook waarmaken.

JOHAN VAN LEEUWAARDEN is hoogleraar wiskunde aan de TU Eindhoven.
E-mail: <j.s.h.v.leeuwaarden@TUE.nl>

Young Statisticians

Hi STATOR readers!

The Young Statisticians have a new addition to the board:

Iris is 27 years young, has a masters degree in clinical psychology and methodology and statistics, and is currently doing a PhD at the VU.

During our members meeting September 30th, we have discussed some great new ideas. Would you like to become an active member of the young statisticians? Please send an email to <young@vvs-or.nl> and join us!

Furthermore, the young statisticians have had some fun and interesting meetings.

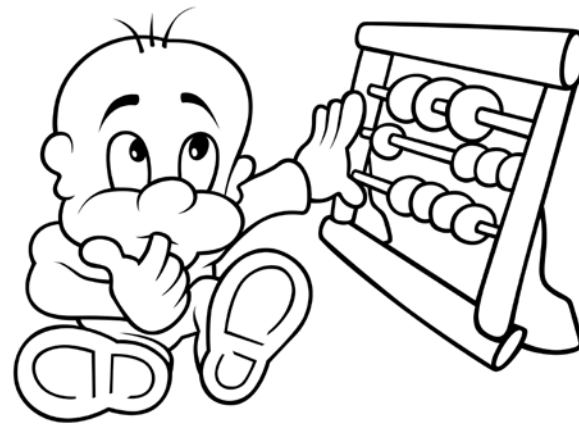
MEET & GREET, AND BOOK MARKET

On a very stormy September evening, the young statisticians gathered in Utrecht for a statistical book market and 'meet & greet'. The evening started with three small talks from young statisticians amongst us. We got a short introduction to misclassification in survival analysis, timetabling for high schools and the grading of Dutch as a secondary language. This was not only very interesting, it also showed the diversity of fields a young statistician might end up working. After these short pitches, the young statisticians met and greeted each other, and everyone browsed through the books to see if there was something of their interest. It was nice to see how calculus books were popular in a bar for once!

We would like to thank all statisticians that donated books and made this event possible.

WORKSHOP

Our next event will be a workshop at Unilever, on November 12th. During the afternoon we will learn how statistics is applied at this company. More information will follow in our newsletter.



STATISTICAL PUB QUIZ

Shortly before the summer, Young Statisticians met in Leiden for a pub quiz. We were assigned to teams and discussed the correct answers in several rounds of trivia questions. Who knew John von Neumann named his dog Inverse? We also got to know some young statisticians from a different perspective. It was very interesting to see which team member recognized 'Take a chance on me' by ABBA within the first beat and who was able to name all 150 Pokemon...

Thanks Stéphanie and Suzanne for organizing this great event.

PR

Our PR committee helps with the promotion of the Young Statisticians and the activities that are organised. It makes posters and keeps the social media updated. So join our Young Statisticians facebook to receive updates!

DATABASE

In order to know where our members study and work, and to get an idea of the fields in which statisticians end up, a database with information about the whereabouts of our members and other statisticians is made. When it is finished, you can look up companies in which (our) statisticians work.

Do you have any ideas for new activities? Or do you want to help the Young Statisticians by making posters, organizing the science cafe, helping to find sponsors or even joining the board?
Please let us know by emailing to <young@vvs-or.nl>

Back to school 2014

Lunteren 16 januari 2014

Alexander Rinnooy Kan was er duidelijk over in zijn lezing op de OR2013 conferentie in Rotterdam. Ondanks dat Nederlandse Operations Research-ers zeer succesvol zijn – kijk maar naar de recente Franz Edelman-successen (www.informs.org/Recognize-Excellence/Franz-Edelman-Award) – is het op de hoogte blijven van recente ontwikkelingen in ons vakgebied van groot belang. Vorig jaar heeft het NGB samen met het LNMB voor het eerste een Back to School Seminar georganiseerd met exact dat voor ogen. Het seminar leverde zoveel positieve reacties op dat we ook in 2014 een Back to School organiseren. Centraal staan de thema's Constraint Programming en Multi Objective Optimisation.

In de afgelopen jaren heeft Constraint Programming (CP) in de OR-wereld veel aandacht gekregen vanwege het potentieel om lastige praktijkproblemen op te lossen. CP heeft een stevig theoretisch fundament en heeft veel commerciële aandacht gekregen, vooral als het gaat om heterogene optimalisatie- of satisfactie-problemen. Ondanks al deze aandacht is CP nog een relatief onbekend en onbegrepen modelleertechniek. Vele curricula in Operations Research bevatten niet eens een inleiding CP. Voor ons een reden om CP op de agenda te zetten van Back to School 2014.

Multi objective optimisation kent vele toepassingen in bijvoorbeeld *data mining*, *machine learning*, financiële vraagstukken, *management science*, speltheorie, engineering en medische beslissingsondersteuning. Een traditioneel optimalisatievraagstuk kent slechts een doelstelling, in de praktijk komt het echter vaak voor dat er sprake is van meerdere elkaar tegenwerkende doelstellingen. Denk maar aan de risico-rendementsafwegingen in de financiële wereld. Vaak worden deze vraagstukken omgevormd tot een enkelvoudige doelstelling waarbij de overige doelstellingen als een beperking in het op te lossen model worden opgenomen. Recente ontwikkelingen in de theorie maken het echter beter mogelijk met meerdere doelstellingen om te gaan.

Het seminar ziet er als volgt uit. Per hoofdonderwerp worden eerst de achtergrond en theorie toegelicht. Vervolgens wordt het gebruik van de technieken in de praktijk besproken. Afgesloten wordt met een praktijkvoorbeeld. Binnenkort volgt meer informatie over het programma en de sprekers. Houdt de websites <www.lnmb.nl>, <www.ngb-online.nl> en linkedIn-groepen in de gaten. Voor meer informatie zie <www.lnmb.nl/conferences/2014/announcementngblnmbseminar/>.

John Poppelaars,
voorzitter van het Nederlands Genootschap Besliskunde

UITMUNTENDE MASTER'S OF PH.D. THESES

OPROEP

OM KANDIDATEN TE NOMINEREN VOOR DE

VvS+OR MASTER'S THESIS 2013

&
WILLEM R. VAN ZWET AWARD 2013

Ter bekroning van een uitzonderlijke afstudeerprestatie aan een Nederlandse instelling voor wetenschappelijk onderwijs of hoger beroepsonderwijs looft de VvS+OR al sinds 1989 een scriptieprijs uit: de VvS+OR Master's Thesis Award. Sinds vorig jaar is daar een prijs bijgekomen voor dissertaties: de Willem R. van Zwet Award.

De VvS+OR roept op tot nominaties voor deze prijzen. Beide prijzen bestaan uit een oorkonde en een geldbedrag van 1000 Euro.

Genomineerd kunnen worden personen die tussen september 2011 en september 2012 zijn afgestudeerd respectievelijk gepromoveerd en die nog niet eerder zijn genomineerd.

Hierbij worden supervisors (begeleiders) opgeroepen om een uitmuntende afstudeerscriptie (Master) of dissertatie (Ph.D.) te nomineren voor de VVS+OR Master's Thesis Award cq. Van Zwet Award 2013.

De indiening van een nominatie dient vergezeld te gaan van een aanbevelingsbrief van de supervisor van de genomineerde. De precieze procedure voor beide prijzen, alsmede de reglementen en het nominatieformulier zijn te downloaden op de website van de VvS+OR <<http://www.vvs-or.nl>>. De nominatie dient binnen te zijn voor 14 december 2013.

Namens de VvS+OR,
Prof. dr. Eric Cator (juryvoorzitter Master's Thesis Award 2013), prof. dr. Peter Grünwald (juryvoorzitter Van Zwet Award 2013) en dr. Mark van der Loo (secretaris Jury Master's Thesis Award 2013)

FELICITATIE

De redactie van STATOR feliciteert haar collega Johan van Leeuwen met zijn benoeming tot gewoon hoogleraar wiskunde aan de Technische Universiteit Eindhoven. Johan zal zich daar wijden aan de studie van Stochastische Netwerken. Hij heeft zijn boeiende oratie uitgesproken op 20 september j.l. Vanaf eind december te downloaden via <www.tue.nl/oraties>

