

‘Ik stel grote vragen aan kleine datasets’

Hoe meer data je hebt, hoe meer je te weten komt, zo wordt vaak gedacht. Maar wat als die ‘big data’ er niet zijn en je toch een antwoord wil op je vragen? Dit is de uitdaging van statisticus Rens van de Schoot (37). ‘Ik ga altijd naast de onderzoeker zitten.’

Maar het antwoord vinden op deze onderzoeks-vraag is niet gemakkelijk. Want hoewel een grote groep jongeren deze gruwelijke aanslag moest meemaken, is de dataset relatief klein. Om betrouwbare statistische conclusies te kunnen trekken hebben onderzoekers door-gaans duizenden slachtoffers nodig. En die zijn er niet. ‘En wachten op de volgende “shooting” doen we niet,’ zegt Van de Schoot. Daarom kijkt hij hoe hij met de bestaande beperkte informatie toch betrouwbare antwoorden kan krijgen.

Wereldwijd wordt u gevraagd te helpen bij het onderzoeken van kleine datasets. Waarom heeft men juist u nodig? ‘Vaak wordt niet de juiste vraag gesteld, maar de vraag die leidt naar het meest wenselijke antwoord. Het is ook voor onderzoekers lastig om concreet te worden. Ik probeer uit te zoeken wat onderzoekers precies willen weten. Neem de website van het FD. Waarom wordt dat ene artikel het meest gelezen? Is dat omdat de redactie goed aanvoelt wat de lezers willen, door dat ene artikel bovenaan op de homepage te plaatsen? Of is dat omdat het artikel de hele ochtend al boven aan de webpagina staat, en lezers dit artikel het eerste zien en er daarom juist op klikken? Ik vraag in zo’n geval ook hoe lang het artikel boven aan de pagina heeft gestaan en vergelijk dat met de artikelen die minder prominent zijn gebracht. Dan weet je of de keuze die de redactie heeft gemaakt, de juiste is. Ik voorkom de cirkelredenering.’

Dus uw werk bestaat uit vragen stellen? ‘In het voorbeeld van het FD zou de redactie misschien heel graag willen dat ze de juiste artikelen hebben gekozen voor op de voorpagina. De redactie zal dan net zo lang zoeken in de data om steun te krijgen voor deze gedachte. Ik en alle methodologen met mij staan hier veel neutraler in. Ongeacht de uitkomst willen wij een goede statistische analyse die de vraag beantwoordt. Ik ga altijd naast de onderzoeker zitten. Als ik dan zie dat hij in zijn onderzoek de verkeerde vragen stelt, dan wijs ik hem daarop.’

Ik ben een soort datafluisteraar: dat is denk ik mijn kracht. Dat is waarom mensen mij naar Oslo laten vliegen, of zoals binnenkort voor een dag naar Canada laten komen. Anderhalve maand geleden was ik in Denemarken. Dat gaat dan



Als ik zie dat een onderzoeker de verkeerde vragen aan het stellen is, dan wijs ik hem daarop



Rens van de Schoot: ‘Ik help onderzoekers de juiste vragen te stellen aan data’.
FOTO: PETER BOER
VOOR HET FD

om de militairen die een aantal jaren op een missie in Afghanistan zijn geweest. Relatief veel militairen hebben daarna zelfmoord gepleegd. De Denen willen weten hoe dat kan. Bij zo’n onderzoek met een kleine dataset is de vraagstelling heel belangrijk.’

Waarom richt u zich op small data? ‘Kleine datasets zitten vaak dicht op de maatschappij, maar het gaat wel om grote vragen. Zo doe ik ook met het brandwondencentrum in Beverwijk onderzoek naar zwaar verbrande baby’s. We willen weten hoe het met hen gaat en met hun ouders, die ook vaak getraumatiseerd zijn. Net als in Noorwegen kunnen we hier niet zeggen: we wachten totdat er nog eens vijftig kinderen zijn verbrand.’

Wordt u ook ingehuurd door bedrijven? ‘Ik werk voor consultant Profitwise. Zij doen veel op het gebied van leiderschapstrainingen en willen weten welke trainingen ze aan welke personen moeten geven, en hoe we achteraf kunnen zien of de juiste training aan de juiste personen is gegeven. Dan moet je denken aan een bepaald type verkoper. Verder hebben we bij alle financiële instellingen in Nederland uitgezocht welke ceo’s de cijfers voor het komende kwartaal goed konden voorspellen. Dit gebeurde anoniem. Sommigen konden dat heel goed, anderen konden dat veel minder.’

Het zou de economie vermoedelijk meer rust geven als we weten wie het beste een financiële instelling kan leiden. ‘Die specifieke vraag hebben we nog niet beantwoord. We waren allang blij dat

Bij small data is de fout snel te vinden, bij big data niet

Big data is voor veel bedrijven een toverterm. Zij zoeken in gegevens die ze jarenlang hebben verzameld naar kennis over hun klanten, leveranciers of andere betrokkenen die hun businessmodel onderscheidend kan maken.

Over small data horen we zelden iemand, terwijl voor veel bedrijven kleine datasets van enorme waarde kunnen zijn. Neem hr-managers. Zij zoeken naar werknemers die hun vak verstaan, maar vooral naar werknemers die in het team passen. Diversiteit in achtergrond en ervaring is daarbij van belang. Maar veel algoritmen kunnen deze variabelen niet aan. Zodoende worden vooral de jonge techies met een

opleiding op een van de beste universiteiten eruit gevist. Terwijl eenheidsworst in een team niet per se het beste resultaat geeft.

Het advies van veel datacrunchers is om voordat je data gaat verzamelen goed te bedenken waar je naar zoekt en uit welke data je het antwoord kunt halen. Zo schreef Frederik Nieuwenhuys onlangs in het FD over een onderzoek door zijn collega bij de boodschappenbezorgdienst Picnic. Om te begrijpen hoe de klanten deze dienst ervaren, dook Picnic niet in de data die

Bij big data is de herkomst van de data vaak niet meer te herleiden, waardoor ook de bias moeilijk is te controleren

het verzamelde over bijvoorbeeld kopersgedrag, maar ging men bij de klanten thuis langs en stelde daar vragen. Het grote voordeel bij deze aanpak is dat een bedrijf niet alleen een antwoord krijgt op de vraag, maar ook informatie (data) krijgt over andere zaken die de klant kwijt wil.

Big data kan ook veel ‘bias’ of ruis bevatten, omdat de personen die de data hebben ingevoerd hier fouten mee hebben gemaakt. Bij small data gebeurt dat ook, maar is de fout sneller te herkennen. Bij big data is de herkomst van de data vaak niet meer te herleiden, waardoor ook de bias moeilijk is te controleren. Het gevolg is dat de uitkomsten niet altijd betrouwbaar zijn.

Ceo’s en rampen



Breivik en Utøya

Bij de aanslag van Anders Breivik op 22 juli 2011 op het Noorse eiland Utøya, dat werd gebruikt voor jongerenkampen, vielen 77 doden. Rens van de Schoot onderzoekt met Noorse experts hoe het de 495 jongeren die de ramp overleefden sindsdien is vergaan en of ze nog last hebben van een post-traumatisch stressyndroom.



Ceo’s die voorspellen

Van de Schoot onderzocht bij financiële instellingen in Nederland welke ceo’s de cijfers voor het komende kwartaal goed konden voorspellen. Sommigen konden dat heel goed, anderen veel minder. Wie dat zijn, blijft geheim.



Brandwonden bij baby’s

Met het brandwondencentrum in Beverwijk doet Van de Schoot onderzoek naar zwaar verbrande baby’s die zijn opgenomen in het brandwondencentrum in Beverwijk. Omdat baby’s bijvoorbeeld niet kunnen aangeven dat hun luiër in brand staat, lopen ze vaak extra brandwonden op. Zowel de baby’s als hun ouders worden hierbij gevolgd.

negen van de twintig ceo’s welwillend waren om aan ons deze gevaarlijke uitspraken te doen.’

Negen ceo’s is wel een erg klein onderzoek. Hoe kunt u ervoor zorgen dat de uitkomst toch betrouwbaar is? ‘Het antwoord dat we krijgen uit zo’n kleine dataset is altijd een eerste stap naar het uiteindelijke antwoord. Dus na de aanslag in Noorwegen doen we ook onderzoek op andere locaties waar aanslagen zijn gedaan. Zoals bij de slachtoffers van 9/11 in New York. Daar kijken we hoe het de politie- en brandweermensen is vergaan. Door meer onderzoek naast elkaar te leggen kun je weer hardere uitspraken doen.’

Haalt u ook andere data erbij? ‘Voorkennis is belangrijk. Neem de ramp in café het Hemeltje in Volendam op 1 januari 2001. In de oudejaarsnacht ontstond brand, waardoor veel jongeren ernstige brandwonden opliepen. Toevallig waren ze in de regio Volendam al met een onderzoek bezig naar brandwonden. Die resultaten hebben we meegenomen in ons onderzoek. Daaruit bleek dat de post-traumatische stressreacties bij rampen zoals in Volendam intensiever en langduriger zijn dan bij individuele gevallen.’

Is dit een beroep in opkomst? ‘Absoluut. 100% van onze promovendi krijgt een baan. Dat is bij het CBS, het Cito, bij bedrijven, bij PwC — er is onwijs veel behoefte aan iemand die naast de onderzoeker zit om de onderzoeksvraag scherp te krijgen.’

Zou u bij een bedrijf willen werken? ‘Je moet niet op de universiteit werken voor het salaris, maar ik houd van de vrijheid. Ik hoef geen uren te schrijven. Ik hoef me niet te verantwoorden waarom ik naar Oslo ga. En ik blijf “on top of mind” met nieuwe ontwikkelingen. Ik vind de maatschappelijke relevantie van mijn werk belangrijk. Maar het kan ook gaan om ceo’s die hun bedrijf beter willen vormgeven. Die variatie heb ik bij de universiteit. Maar ik zie ook dat we in de wetenschap soms vergeten de grote vragen te beantwoorden. We zijn zo veel mogelijk aan het publiceren en beurzen aan het binnenhalen. We zijn meer bezig met onze baan te behouden dan dat we ons afvragen wat de onderzoeksvragen zijn waaraan we werken.’

Zonder financiering werk je nergens. ‘Klopt. Ik zit aan de goede kant van de lijn, want ik heb een vaste baan en genoeg financiering. Maar ik zie wel dat anderen te veel rennen. Je hoeft niet tientallen publicaties op je naam te hebben om je baan te houden. Misschien zijn we wel met te veel mensen in de wetenschap. Dat kan ook.’

Maaik Kroon, hoogleraar chemische technologie, ging naar Abu Dhabi omdat ze daar geen tijd verloor aan het binnenhalen van financiering. Zou u zo iets willen? ‘Ik denk dat ik met Maaik zou moeten praten als ik een aanbod krijg. De vraag speelt bij mij nog niet echt. Ik heb wel wat voorzichtige vragen vanuit Amerika gehad. Maar daar zijn ze nog meer op publiceren en elkaar beconcurreren gericht dan hier. Daarbij zou ik me niet prettig voelen.’

Franka Rolvink Couzy is redacteur van Het Financieele Dagblad.



Van de Schoot onderzocht ook waarom zoveel Deense soldaten zelfmoord pleegden na hun missie in Afghanistan.
FOTO: GETTY