Inhoud

1

Nu

8 In 2009 werd er een nieuw griepvirus ontdekt. De nieuwe stam, die als H1N1 werd aangeduid, bevatte elementen van de virussen die vogelgriep en varkensgriep veroorzaken en verspreidde zich snel. Binnen enkele weken vreesden volksgezondheidsinstellingen overal ter wereld dat er een verschrikkelijke pandemie aanstaande was. Sommige commentatoren waarschuwden voor een griepuitbraak van dezelfde omvang als de Spaanse griep van 1918, waarbij een half miljard mensen besmet raakten en waaraan tientallen miljoenen overleden. En, wat nog erger was, er was geen onmiddellijk beschikbaar vaccin tegen het nieuwe virus. De enige hoop van de verantwoordelijken voor de volksgezondheid was dat het zou lukken de verspreiding van het virus te vertragen. Maar daarvoor moesten ze eerst weten hoe ver het virus zich al had verspreid.

In de Verenigde Staten verzochten de Centers for Disease Control and Prevention (centra voor ziektebestrijding en -preventie, CDC's) artsen hen te informeren over nieuwe griepgevallen. Maar het beeld van de pandemie dat daaruit naar voren kwam, was al-

tijd dat van een paar weken geleden. Mensen voelden zich misschien al dagenlang ziek, maar wachtten nog met naar de dokter gaan. Het doorgeven van de informatie aan de centrale instanties kostte tijd en de CDC's verwerkten de gegevens maar één keer per week. Bij een ziekte die zich snel verspreidt is een vertraging van twee weken veel te veel. Door die vertraging tastten de instanties voor gezondheidszorg op de meest cruciale momenten volledig in het duister.

Slechts een paar weken voordat het H1N1-virus de voorpagina's haalde, hadden een paar dataspecialisten van de internetgigant Google een opmerkelijk artikel gepubliceerd in het wetenschappelijke tijdschrift *Nature*. Het sloeg als een bom in bij verantwoordelijken voor gezondheidszorg en informatici, maar bleef verder onopgemerkt. De auteurs legden uit dat Google de verspreiding van de wintergriep in de VS kon 'voorspellen', niet alleen op landelijk niveau, maar per regio en zelfs per deelstaat. Dat kreeg het bedrijf voor elkaar door te kijken waar mensen op internet naar zochten. Aangezien Google elke dag meer dan drie miljard zoekopdrachten krijgt en die allemaal bewaart, waren er meer dan genoeg gegevens om mee te werken.

Google bekeek de 50 miljoen zoektermen die het meest door Amerikanen werden ingetypt en vergeleek die lijst met de gegevens van de CDC's over de verspreiding van de seizoensgriep tussen 2003 en 2008. De bedoeling was de mensen die met het griepvirus geïnfecteerd waren aan te wijzen op basis van dingen waar ze op internet naar zochten. Datzelfde hadden ook anderen al geprobeerd, maar niemand beschikte over zoveel gegevens, rekenkracht en statistische knowhow als Google.

De mensen van Google hadden erop gegokt dat de zoekers uit waren op het verkrijgen van informatie over griep – met behulp van zoektermen als 'medicijn voor hoest en koorts' –, maar dat bleek niet de crux te zijn. Daarom ontwierpen ze een systeem waarbij het er niet toe deed waar mensen naar zochten. Het systeem zocht alleen maar naar correlaties tussen de frequentie van

bepaalde zoekopdrachten en de verspreiding van de griep in tijd en ruimte. In totaal verwerkten ze maar liefst 450 miljoen verschillende wiskundige modellen voor het testen van de zoektermen en vergeleken ze de voorspellingen met de gegevens van de CDC's over werkelijke griepgevallen in 2007 en 2008. En ze stuitten op een goudmijn: hun software vond een combinatie van vijfenveertig zoektermen die, indien ze samen werden ingevoerd in een wiskundig model, een sterke correlatie lieten zien tussen de voorspellingen en de officiële landelijke cijfers. Ze konden net als de CDC's aangeven waar de griep zich had verspreid, maar anders dan de CDC's konden ze dat bijna realtime in plaats van pas een paar weken later.

Toen de H1N1-crisis in 2009 toesloeg, bleek het Google-systeem een bruikbaardere en snellere indicator dan de overheidsstatistieken met hun inherente meldingsvertragingen. Zo kregen de verantwoordelijken voor de volksgezondheid waardevolle informatie ter beschikking.

De Google-methode is opmerkelijk omdat er geen speekselmonsters worden verzameld en geen artsenpraktijken worden bevraagd. In plaats daarvan maakt ze gebruik van *big data* – het vermogen van de samenleving om informatie op nieuwe manieren in te zetten voor het verkrijgen van nuttige inzichten of waardevolle goederen en diensten. Daarmee heeft de wereld tegen de tijd dat de volgende pandemie zich aandient een beter instrument ter beschikking om de verspreiding van de ziekte te voorspellen en dus te voorkomen.

Gezondheidszorg is maar een van de terreinen waarop de big data-revolutie haar invloed doet gelden. Hele sectoren van het zakenleven veranderen op dit moment ingrijpend door big data. De aankoop van vliegtickets is een goed voorbeeld.

In 2003 moest Oren Etzioni van Seattle naar Los Angeles vliegen voor de bruiloft van zijn jongste broer. Een paar maanden voor de grote dag ging hij online en kocht hij een vliegticket, in de ver-

onderstelling dat je minder betaalt als je vroeg boekt. Tijdens de vlucht kreeg zijn nieuwsgierigheid de overhand, en hij vroeg de man die naast hem zat hoeveel zijn ticket had gekost en wanneer hij het had gekocht. De man bleek aanzienlijk minder te hebben betaald dan Etzioni, hoewel hij zijn ticket veel later had gekocht. Dat zat Etzioni helemaal niet lekker, en hij sprak nog meer medepassagiers aan. De meesten hadden minder betaald dan hij.

De meeste mensen zouden hun verontwaardiging over hun te dure ticket waarschijnlijk zijn vergeten tegen de tijd dat ze hun tafeltje opklapten en hun vliegtuigstoel weer rechtop zetten. Maar Etzioni is een van de meest vooraanstaande Amerikaanse informatici. Hij beschouwt de wereld als een reeks big data-problemen – die hij kan oplossen. En daar is hij al mee bezig sinds hij in 1986 aan Harvard afstudeerde als eerste student met het hoofdvak informatica.

Vanaf zijn uitvalsbasis aan de universiteit van Washington had hij al een hele rits big data-bedrijven opgestart voordat de term 'big data' bekend werd. Hij hielp bij het bouwen van een van de eerste zoekmachines op het web, MetaCrawler, die in 1994 werd gelanceerd en kort daarna overgenomen door InfoSpace, destijds een groot internetbedrijf. Hij is een van de oprichters van Netbot, de eerste prijsvergelijkingssite voor consumenten, die hij later doorverkocht aan Excite. Zijn bedrijfje voor het distilleren van betekenis uit tekstdocumenten, ClearForest, werd later overgenomen door Reuters.

Zodra Etzioni weer met beide benen op de grond stond, besloot hij een manier te bedenken om mensen te laten weten of een ticketprijs die ze online zien een goede koop is of niet. Een vliegtuigstoel is een massaproduct: de ene stoel is in principe niet te onderscheiden van alle andere op dezelfde vlucht. Toch variëren de stoelprijzen enorm onder invloed van ontelbare factoren waarvan de meeste uitsluitend aan de luchtvaartmaatschappijen zelf bekend zijn.

Etzioni kwam tot de conclusie dat hij de precieze mechanismes

achter de prijsverschillen niet hoefde uit te pluizen. Hij hoefde alleen maar te voorspellen hoe waarschijnlijk het was dat de weergegeven prijs in de toekomst zou stijgen of dalen. Dat is mogelijk, zij het niet eenvoudig. Het enige wat je ervoor hoeft te doen, is alle ticketverkopen voor een bepaalde route analyseren en onderzoeken hoe de betaalde prijzen afhangen van het aantal dagen voor vertrek.

Als de gemiddelde prijs van een ticket de neiging had te dalen, zou het verstandig zijn te wachten en het ticket later te kopen. Als de gemiddelde prijs meestal steeg, gaf het systeem de aanbeveling het ticket direct te kopen voor de weergegeven prijs. Met andere woorden, er hoefde alleen maar een veredelde versie te worden uitgevoerd van Etzioni's informele onderzoekje op tien kilometer hoogte. Voor de goede orde: ook dit was weer een moeilijk informaticaprobleem. Maar het was een probleem dat hij kon oplossen. En dus ging hij aan het werk.

Etzioni werkte met een steekproef van 12.000 prijsobservaties
12 die was verkregen door gedurende een periode van eenenveertig dagen informatie te 'oogsten' van een reiswebsite. Daarvan uitgaande bouwde hij een voorspellingsmodel dat de denkbeeldige passagiers een aardig voordeel opleverde. Het model hield zich niet bezig met het 'waarom', maar uitsluitend met het 'wat'. Het kende geen van de variabelen die een rol spelen bij de prijsbeslissingen van luchtvaartmaatschappijen, zoals het aantal nog onverkochte stoelen, seizoensinvloeden of de vraag of een overnachting in het weekend de prijs misschien op magische wijze drukte. Het baseerde zijn voorspellingen op wat het model wél kende: waarschijnlijkheden die moeizaam bijeen waren gesprokkeld uit gegevens van andere vluchten. 'Kopen of niet kopen, dat is de vraag,' zo redeneerde Etzioni, en hij gaf zijn onderzoeksproject de toepasselijke naam Hamlet.

Uit dit kleinschalige beginproject ontstond een met risicodragend kapitaal gefinancierde onderneming, Farecast. Door te voorspellen of de prijs van een vliegticket waarschijnlijk zou stijgen of dalen, en hoeveel, stelde Farecast consumenten in staat te bepalen

wanneer ze op de knop 'Koop nu' moesten klikken. Ze hadden nu de beschikking over informatie waartoe ze tot voor kort nooit toegang hadden gehad. Het bedrijf gaf, als toppunt van transparantie, zelfs aan hoeveel vertrouwen het in zijn eigen voorspellingen had en liet die informatie op het scherm aan de gebruikers zien.

Om goed te kunnen werken moest het systeem over heel veel gegevens beschikken. Etzioni wist toegang te krijgen tot een van de reserveringsdatabases van de luchtvaartbranche om de prestaties van zijn systeem te verbeteren. Op basis van die informatie kon het model voorspellingen doen aan de hand van gegevens over alle stoelen op alle vluchten op het merendeel van de routes in de Amerikaanse burgerluchtvaart over een periode van een jaar. Farecast verwerkte voor het doen van voorspellingen inmiddels bijna 200 miljard vluchtprijsrecords, waarmee het bedrijf consumenten een smak geld bespaarde.

Etzioni maakt met zijn zandkleurige haar, zijn brede grijns en zijn engelachtige voorkomen niet bepaald de indruk iemand te zijn die in staat is de luchtvaartbranche miljoenen dollars aan potentiele inkomsten door de neus te boren. Maar in werkelijkheid heeft hij zich zelfs een nog hoger doel gesteld. In 2008 had hij plannen om dezelfde methode ook toe te passen op andere goederen, zoals hotelkamers, concertkaartjes en tweedehandsauto's: allemaal zaken die worden gekenmerkt door weinig productdifferentiatie, sterke prijsvariaties en een enorme hoeveelheid gegevens. Maar voordat hij die plannen verder had kunnen uitwerken klopte Microsoft bij hem aan, slokte Farecast voor circa 110 miljoen dollar op en integreerde het in de zoekmachine Bing. In 2012 deed het systeem in driekwart van de gevallen correcte voorspellingen en bespaarde het reizigers gemiddeld 50 dollar per ticket.

Farecast is een schoolvoorbeeld van een bedrijf dat zijn geld verdient met big data, en het biedt ons een kijkje in de nabije toekomst. Etzioni zou het bedrijf vijf of tien jaar geleden niet hebben kunnen opzetten. 'Dat zou onmogelijk zijn geweest,' zegt hij. De hoeveelheid rekenkracht en opslagruimte die hij nodig had, was

toen nog te duur. Maar hoewel de technologische ontwikkelingen een cruciale factor zijn geweest, veranderde er ook nog iets veel belangrijkers, iets subtiels: de manier waarop tegen het mogelijke gebruik van gegevens werd aangekeken.

Gegevens werden niet meer beschouwd als statisch of triviaal, iets wat geen nut meer had zodra het doel was bereikt waarvoor ze oorspronkelijk waren verzameld, bijvoorbeeld nadat het vliegtuig was geland of – in het geval van Google – een zoekopdracht was afgewerkt. Gegevens veranderden in een ruwe grondstof voor bedrijven, een belangrijk economisch goed dat kon worden gebruikt om een nieuwe vorm van economische waarde te creëren. Sterker nog, met de juiste invalshoek kunnen gegevens op een slimme manier worden hergebruikt en zo uitgroeien tot een rijke bron van innovaties en nieuwe diensten. Ze kunnen geheimen onthullen aan degenen die beschikken over de nederigheid, de bereidheid en de juiste gereedschappen om te luisteren.

DE GEGEVENS LATEN SPREKEN

De smartphone is in ieders hand, in menige rugzak zit een laptop en in elk bedrijf staat een groot IT-systeem – de verworvenheden van de informatiesamenleving zijn moeilijk over het hoofd te zien. Maar de informatie zelf springt minder in het oog. Een halve eeuw nadat computers een rol gingen spelen in het dagelijkse maatschappelijk verkeer is de hoeveelheid gegevens zo groot geworden dat er een bijzondere nieuwe ontwikkeling op gang is gekomen. Niet alleen wordt de wereld door meer informatie overspoeld dan ooit tevoren; die hoeveelheid groeit ook sneller. Die schaalverandering heeft tot een geheel nieuwe situatie geleid: de kwantitatieve verandering heeft een kwalitatieve verandering veroorzaakt. In wetenschappen als astronomie en genetica, die de explosie in het eerste decennium van de eenentwintigste eeuw als eerste meemaakten, is hiervoor de term ‘big data’ bedacht. Dit begrip vindt nu zijn weg naar alle andere maatschappelijke terreinen.

Er bestaat geen exacte definitie van big data. Aanvankelijk had men het idee dat de omvang van de informatie zo groot was geworden dat de te onderzoeken hoeveelheid gegevens niet meer in het geheugen paste dat computers voor de verwerking ter beschikking hadden, zodat de dataspecialisten de gereedschappen die ze voor hun analyses gebruikten ingrijpend moesten herzien. Dat heeft geleid tot nieuwe verwerkingstechnieken zoals MapReduce van Google en het open source-equivalent Hadoop, dat afkomstig is van Yahoo. Hiermee konden veel grotere hoeveelheden gegevens worden beheerd dan daarvoor, en – dat was belangrijk – die gegevens hoefden niet meer in nette reeksen of databasetabellen te worden geordend. Er wordt al gewerkt aan andere technieken voor het verwerken van grote hoeveelheden gegevens waarin de rigide hiërarchische structuren en de gelijkvormigheid van oude systemen eveneens geen vereiste meer zijn. Tegelijkertijd groeiden internetbedrijven, die immers een enorme schat aan gegevens konden verzamelen en een dringende financiële noodzaak voelden om daar iets zinnigs mee te doen, uit tot de grootste gebruikers van de nieuwste verwerkingstechnieken, waarmee ze traditionele bedrijven overvleugelden die soms decennia meer ervaring hadden.

Een manier om op dit moment tegen deze kwestie aan te kijken – de visie die wij ook in dit boek propageren – is de volgende: de term 'big data' verwijst naar dingen die je op een grote schaal kunt doen en die op een kleinere schaal niet mogelijk zijn, en waarmee je nieuwe inzichten verkrijgt of nieuwe vormen van economische waarde creëert op een manier die invloed heeft op onder andere markten, organisaties en de relatie tussen burgers en overheden.

Maar dat is nog maar het begin. Het tijdperk van de big data grijpt in op onze manier van leven en onze interactie met de wereld. Het meest in het oog springende is dat de samenleving een deel van haar obsessie met causaliteit zal moeten afdanken in ruil voor simpele correlaties: niet weten waaróm, maar alleen maar wát. Dit gaat lijnrecht in tegen eeuwenoude gewoontes en gebruiken en raakt de kern van ons begrip van de manier waarop we