INHOUD

Inleiding

Je hebt inmiddels vast al een hoop over Big Data gehoord. Over de geweldige mogelijkheden, over de gevaarlijke consequenties en het nieuwe paradigma dat alle eerdere paradigma's tenietdoet, en wat dat voor de mensheid en haar o zo geliefde websites betekent. Het is bijna niet te bevatten, alsof je net een flinke klap op je hoofd hebt gekregen. Daarom gaat dit boek dan ook niet over de zoveelste hype of het zoveelste spannende verhaal over dit hele datagebeuren. Wat ik te bieden heb is waar het daadwerkelijk om gaat: om de data zelf, ontdaan van het schreeuwerige fenomeen. Ik ga je een enorme hoeveelheid van de informatie geven die wordt verzameld, omdat ik door mazzel, mijn werk, soms door wat slinkse handigheidjes en door nóg meer mazzel in de unieke positie verkeer dat ik niet alleen over die gegevens beschik, maar ze ook nog eens kan analyseren.

Ik ben een van de oprichters van OkCupid, een Amerikaanse datingsite die, na een lange en verre van bruisende periode van tien jaar, inmiddels tot een van de grootste ter wereld is uitgegroeid. Ik ben die website met drie vrienden begonnen. We hadden allemaal een wiskundetic, en dat onze website een succes werd, had voor een groot deel te maken met het feit dat we die kennis op dating toepasten. Dat wil zeggen, we voegden wat analyse en nauwkeurigheid toe aan een gebied dat in het verleden vooral het terrein was van 'liefdesexperts' en grijnzende tovenaars à la Dr. Phil. De website zelf zit echt niet bijster ingewikkeld in elkaar. Wat sobere rekenkunde is het enige wat je nodig hebt om een model te maken van het proces waarbij twee mensen elkaar leren kennen. Om de een of andere reden sloeg onze aanpak aan, en alleen al dit jaar zullen 10 miljoen mensen via onze site naar iemand op zoek gaan.[1]

Zoals ik maar al te goed weet, vinden websites (en de bedenkers van die sites) het heerlijk om met indrukwekkende cijfers te strooien, en de meeste weldenkende mensen hebben geleerd daar geen acht meer op te slaan. Wanneer je 'miljoenen hiervan' en 'miljarden daarvan' hoort, weet je dat het in

feite neerkomt op 'hoera voor mezelf', en dat dan gevolgd door een hele rij nullen. Anders dan Google, Facebook, Twitter en de andere bronnen waarvan de data in dit boek uitvoerig aan de orde zullen komen, is OkCupid geen website die iedereen zal kennen. Als jij en je vrienden al jaren gelukkig getrouwd zijn, heb je waarschijnlijk zelfs nog nooit van ons gehoord. Ik heb echt even moeten nadenken hoe ik het bereik van de website kan uitleggen aan iemand die 'm nog nooit heeft gebruikt en die – terecht – geen belangstelling heeft voor de *user-engagement*-getallen van de start-up van een of andere gozer. Laat ik het daarom maar gewoon in persoonlijke termen uitdrukken: dankzij OkCupid zullen pakweg 30 000 stellen vanavond een eerste afspraakje hebben.[2] Ongeveer drieduizend daarvan zullen een langdurige relatie krijgen. Tweehonderd ervan zullen trouwen en uit veel van die relaties zullen, uiteraard, kinderen voortkomen. Kort gezegd: er lopen vandaag de dag springlevende, luid jengelende kinderen rond – tegendraadse wezentjes die hun schoenen 'niet nú' willen aandoen – die zonder de grillen van onze HTML nooit zouden hebben bestaan.

Ik verbeeld me niet dat we iets hebben geperfectioneerd en hoewel ik graag wel wil zeggen dat ik trots ben op de site die mijn vrienden en ik hebben opgezet, kan het me eerlijk gezegd niets schelen of je lid bent, een account gaat aanmaken of wat ook. Ik heb nog nooit via internet gedatet en mijn medeoprichters hebben dat evenmin. Mocht het niks voor je zijn, dan snap ik dat maar al te goed. En ik hou al helemaal niet van mensen die heilig geloven in alles zolang het woordje 'tech' er maar in voorkomt. Het is dan ook niet mijn bedoeling om iemand zijn prachtige eiland afhandig te maken in ruil voor mijn glanzende digitale kralen. Ik ben nog steeds op papieren tijdschriften geabonneerd. In het weekend krijg ik de *Times* in de bus. Twitteren vind ik nogal gênant. Ik ga je er ook niet van proberen te overtuigen dat je internet en sociale media meer zou moeten gebruiken, zou moeten waarderen of er meer in zou moeten 'geloven' dan je nu al doet – of juist niet doet. Blijf alsjeblieft vooral bij wat je tot nu toe van dit hele online gebeuren vond. Maar als er één ding is waarvan ik wél hoop dat je het door dit boek in heroverweging wilt nemen, dan is het wat je over jezelf denkt. Want daar gaat dit boek in feite over. OkCupid is alleen maar de manier waarop ik daartoe kwam.

Ik geef sinds 2009 leiding aan het data-analyseteam van OkCupid. Aan mij dus de taak om wijs te worden uit de data die onze gebruikers genereren. Terwijl mijn drie medeoprichters bijna al het zware bouwwerk aan de site hebben verricht, heb ik jarenlang alleen maar met de cijfertjes zitten pielen. Een deel van de dingen waaraan ik werk is nodig om ons bedrijf van dag tot

dag te runnen. Zo is snappen hoe verschillend mannen en vrouwen aankijken tegen zaken als seks en schoonheid, om maar iets te noemen, voor een datingsite tamelijk essentieel. Maar veel van wat mijn werk oplevert is niet direct bruikbaar en hooguit interessant. Zo kun je weinig doen met het gegeven dat Belle en Sebastian statistisch gezien de minst zwarte band hebben die er bestaat of dat je op een foto die met flits is gemaakt vanzelf al zeven jaar ouder lijkt, behalve dan 'huh?!' roepen en het misschien tijdens een etentje aan anderen vertellen. Dat was een tijdlang ook het enige wat we met zulke weetjes deden; de inzichten die we opdeden kwamen niet verder dan af en toe een nogal flauw persbericht. Op een gegeven moment hadden we alleen zoveel analysegegevens dat we overkoepelende trends begonnen te zien, algemene patronen die al die kleine overstegen. En vooral: ik besefte dat ik die data kon gebruiken om taboes zoals etnische afkomst van nabij te bestuderen. Dus in plaats van mensen een vragenlijst voor te leggen of kleinschalige experimenten op te zetten – wat in de sociale wetenschap tot nu toe de manier was waarop je aan gegevens kwam – kon ik gewoon kijken naar wat er daadwerkelijk gebeurt wanneer zo'n 100 000 blanke mannen met 100 000 zwarte vrouwen achter gesloten deuren met elkaar in contact komen. Die data stonden gewoon op onze servers. Dat was een onweerstaanbaar sociologisch buitenkansje.

Zodra ik me erop stortte, stapelden de ontdekkingen zich op, en zoals iedereen met meer ideeën dan publiek begon ik een blog om deze met de buitenwereld te delen. Die blog groeide uit tot dit boek, zij het na een heel belangrijke verbetering, want ik heb ditmaal veel verder gekeken dan alleen maar naar OkCupid. Ik denk zelfs dat ik een dataset van persoonlijke interacties in handen heb die meer omvat en gevarieerder is dan die waar enige andere privépersoon momenteel over beschikt. Bovendien bevat mijn dataset de meeste, zo niet alle online databronnen die er vandaag de dag toe doen. Ik zal al die gegevens in dit boek gebruiken om uitspraken te doen over de gewoontes van de gebruikers van één site, maar die daarnaast uitvergroten tot een verzameling algemene uitspraken.

Het publieke debat over data gaat vooral over twee kwesties: overheidsspionage en commerciële kansen. Ik betwijfel of ik meer weet dan jij over het eerste, namelijk alleen wat ik erover gelezen heb. Voor zover ik weet, hebben nationale veiligheidsorganisaties in elk geval nog nooit een datingsite benaderd met een verzoek om inzage. Tenzij ze het strafbaar willen stellen om een superstrak wasbord zonder bijbehorend gezicht te tonen, of jonge vrouwen uit Brooklyn gaan aanklagen die maar blijven zeuren dat ze zo dol zijn op whisky, terwijl iedereen dondersgoed weet dat dat niet zo is, kan

ik me ook niet voorstellen dat ze veel belangwekkends op zo'n site zouden aantreffen.

Over het tweede onderwerp, data omzetten in dollars, weet ik wel iets. Toen ik met dit boek begon, lag de hele 'tech'-pers in katzwijm vanwege de aanstaande beursgang van Facebook. Dat bedrijf had de persoonlijke gegevens van zijn gebruikers te gelde weten te maken en kon daar op de beurs nóg meer geld mee gaan verdienen. Een kop in de *Times* drie dagen voor de beursgang vat het denk ik aardig goed samen: FACEBOOK MOET DATA TOT GOUD SPINNEN. Je zou bijna verwachten dat Repelsteeltje op de opiniepagina zou opduiken met de uitspraak: 'Ja, Amerika, dit is een solide investering.'

Als oprichter van een site met advertenties kan ik bevestigen dat gegevens inderdaad handig zijn voor de verkoop. Elke webpagina kan de volledige gebruikerservaring opslurpen: alles waar iemand op klikt, wat hij intikt, en zelfs hoe lang hij ergens blijft hangen. Van daaruit is het vervolgens niet zo heel moeilijk om je een redelijk goed beeld te vormen van iemands behoeften en hoe je die kunt stillen. Maar hoe geweldig de mogelijkheden ook mogen zijn, ik wil het hier niet hebben over een geheime missie in ons land om bodyspray te verkopen aan mensen die iets over deodorant melden aan hun vrienden. Wat ik via dezelfde toegang tot de data wél wil doen is de gebruikerservaring – de klikjes, wat iemand intikt en de milliseconden waarin dat allemaal gebeurt – voor een heel ander doel inzetten. De twee grote en actuele verhalen over Big Data gaan over toezicht en geld, maar ik heb de afgelopen drie jaar aan een derde verhaal gewerkt, namelijk aan het menselijke.

Misschien weet Facebook dat je een van de vele mensen bent die van M&M's houden en krijg je daardoor relevante aanbiedingen voorgeschoteld. Zo weet Facebook ook wanneer het uit is met je vriendje, wanneer je naar Texas verhuist, op steeds meer foto's samen met je ex staat en het vervolgens weer aan is met hem. Google weet wanneer je naar een nieuwe auto op zoek bent en kan je het juiste merk en model tonen die precies bij jouw psychografische profiel passen. Ben je een maatschappelijk geëngageerde, sensatiebeluste persoonlijkheidstype B-man van tussen de 25 en 34 jaar? Dan krijg je een Subaru te zien. Google weet tegelijkertijd ook of je homo bent, boos, eenzaam, racistisch of bang dat je moeder kanker heeft. Twitter, Reddit, Tumblr, Instagram – al die bedrijven zijn eerst en vooral bedrijven, maar op een goede tweede plaats zijn het daarnaast demografen met een ongekend bereik, ongekende volledigheid en een ongekend belang. Zo kunnen die digitale gegevens ons bijna als een bijkomstigheid tonen hoe we ruziemaken,

liefhebben en ouder worden, wie we zijn en hoe we veranderen. Het enige wat we daarvoor hoeven te doen is kijken. Bouw een beetje afstand in en de data ontsluiten vanzelf hoe mensen zich gedragen wanneer ze denken dat er niemand meekijkt. Ik zal je in dit boek vertellen wat ik heb gezien. O ja, en, eh… lazer toch op met die deo!

Mocht je vaker populairwetenschappelijke boeken lezen, dan zullen je in dit boek misschien een paar dingen opvallen. Allereerst de kleur rood. Ten tweede dat het over geaggregeerde cijfers en grote getallen gaat en dat er weliswaar een paar individuen in voorkomen, maar dat die verder opvallend afwezig zijn in een boek dat over mensen beweert te gaan. Grafieken, figuren en tabellen staan er in overvloed in, namen daarentegen amper. Het is in de populaire wetenschap tegenwoordig bijna vaste prik om iets kleins en eigenzinnigs als een lens te gebruiken voor grote gebeurtenissen. Om de geschiedenis van de wereld bijvoorbeeld te vertellen aan de hand van een knolraap, een oorlog terug te voeren tot een vis, en een zaklampje nou net precies op zo'n manier door een prisma te laten schijnen dat je een hele regenboog op je slaapkamermuur kunt projecteren. Ik doe het juist andersom. Ik pak iets groots – een enorme verzameling van wat mensen doen, denken en zeggen, en dan heb ik het over enkele terabyte aan data – en haal daar allerlei dingen uit naar boven: wat je vriendenkring over de stabiliteit van je huwelijk zegt, met welke woorden Aziaten (en blanken en zwarten en latino's) zichzelf juist níét zouden typeren, waar en waarom homo's in de kast blijven, hoezeer onze manier van schrijven de afgelopen tien jaar is veranderd en hoe dat voor boosheid juist niet geldt. Het idee erachter is dat we bij onze opvattingen over onszelf niet zozeer naar de verhalen als wel naar de cijfers kijken, of beter gezegd: op zo'n manier gaan denken dat de cijfers zelf het verhaal worden.

Mijn aanpak is het resultaat van heel wat gezwoeg in de statistische steengroeven, en dit boek komt voort uit wat mijn collega's en ik al jaren doen. Een datingsite brengt mensen samen, en wil je dat op een geloofwaardige manier doen, dan zul je hun verlangens, gewoontes en afkeuren moeten zien te achterhalen. Je verzamelt, met andere woorden, heel veel gedetailleerde gegevens en doet je best om die te vertalen in algemeen geldende theorieën over menselijk gedrag. Wat je wanneer je met al die informatie werkt gaandeweg ontwikkelt, anders dan wanneer je bijvoorbeeld zou werken voor de huwelijkspagina van de zondagskrant, is een zekere affiniteit met de voortploeterende mensheid als geheel in plaats van alleen met twee willekeurige individuen. Je leert mensen begrijpen op een manier zoals dat voor een che-

micus zou gelden, die door zijn inzicht in al die rondzwevende moleculen in zijn maatbeker er automatisch een beetje van gaat houden.

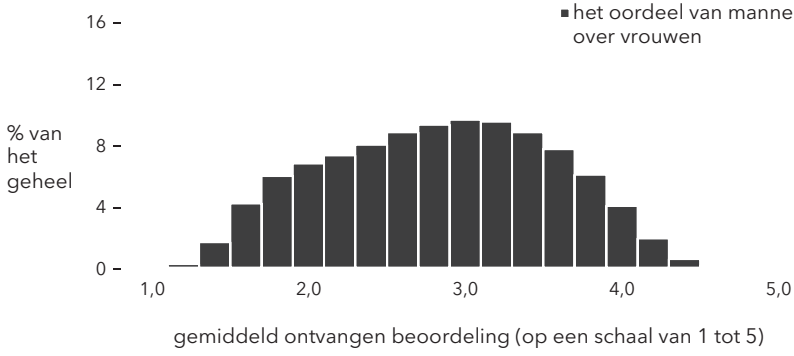
Dat gezegd hebbende: alle websites objectiveren, en sterker nog: alle datawetenschappers doen dat. Algoritmes werken nu eenmaal niet met dingen die geen getallen zijn, dus als je wilt dat een computer een idee begrijpt, moet je daar zo veel mogelijk van in getallen zien om te zetten. De uitdaging waar websites en apps voor staan is het geheel aan menselijke ervaringen, zonder dat iemand het merkt, zien op te delen en in emmertjes 1, 2 en 3 te proppen. Met andere woorden, de uitdaging zit 'm erin om een enorm groot, niet in woorden uit te drukken proces – voor Facebook zijn dat vriendschappen, voor Reddit gemeenschappen, voor datingsites liefde – in brokjes op te delen waar een server iets mee kan. Tegelijkertijd moet je alleen ook zo veel mogelijk van dat 'iets' van al die zaken zien te behouden, zodat wat je de gebruikers aanbiedt nog wel een weergave van het echte leven lijkt. Uiteindelijk is internet namelijk een broze illusie, te vergelijken met een wortel die héél minutieus gesneden is, maar zodanig dat de schijfjes op de snijplank nog wel de exacte vorm van de oorspronkelijke wortel weergeven. De spanning tussen de instandhouding van het menselijk tekort en de fragmentatie van de database kan het lastig maken om een website in de lucht te houden, maar tegelijkertijd maakt dat mijn verhaal wel mogelijk. De technologische benaderingen die tegenwoordig bestaan voor zaken als lust en vriendschap scheppen allerlei nieuwe mogelijkheden: om harde cijfers aan tijdloze raadselen te koppelen en om ervaringen te doorgronden waarvan we tot nu toe aannamen dat ze 'niet te kwantificeren' waren. Naarmate die benaderingen steeds beter worden en mensen ze verder in hun leven toelaten, neemt ons inzicht eveneens met een ontstellende snelheid toe. Laat ik een eenvoudig voorbeeld geven, maar pas nadat ik gezegd heb dat OkCupid eigenlijk als slogan had moeten kiezen: 'Het onnoembare volledig benoembaar maken.' Jammer maar helaas.

Het wemelt op internet van de beoordelingen. Of het nou de stemmen voor of tegen van Reddit zijn, de gebruikersrecensies van Amazon of de 'likes' van Facebook, websites vragen je te stemmen omdat daarmee iets wat veranderlijk en persoonlijk is – namelijk jouw mening – in iets begrijpelijks en bruikbaars kan worden omgezet. Datingsites vragen mensen elkaar te beoordelen, zodat eerste indrukken zoals deze:

> Hij heeft mooie ogen
> Hmm, hij ziet er leuk uit, maar ik hou niet van jongens met rood haar
> Hè bah

... kunnen worden omgezet in eenvoudige getalletjes, bijvoorbeeld 5, 3 en 1 op een schaal van vijf sterren. Via allerlei websites zijn er inmiddels miljarden van die micro-oordelen verzameld over de allereerste indruk die iemand van een ander heeft. Bij elkaar vormen al die minigedachten een bron van immens veel inzicht in de manier waarop mensen zich een mening over elkaar vormen.

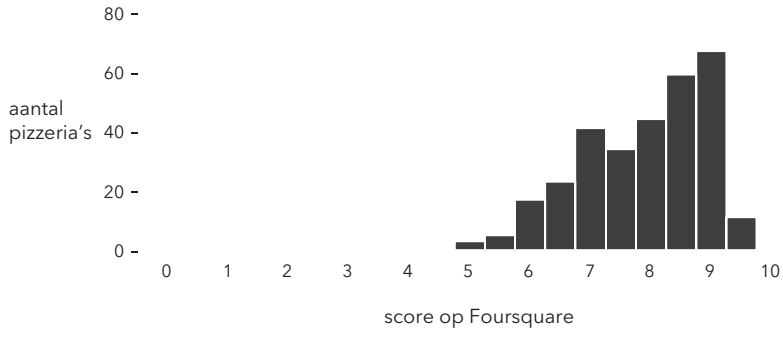
Het meest basale wat je met zulke persoonlijke beoordelingen kunt doen, is ze optellen. Tel hoeveel mensen gemiddeld één ster, twee sterren enzovoort kregen, en vergelijk de scores. Hieronder staat een diagram waarin ik dat heb gedaan met de gemiddelde scores die heteroseksuele mannen aan heteroseksuele vrouwen gaven. Dat ziet er zo uit:



Dat is waar 51 miljoen voorkeuren op neerkomen: op deze eenvoudige verzameling staafjes. Dat is, au fond, de gezamenlijke mannelijke mening over vrouwelijk schoon op OkCupid. Zo worden alle miniverhaaltjes (wat een man van een vrouw vindt, en dat dan keer een paar miljoen) en alle anekdotes (en elk daarvan had ik hier, als dit een ander soort boek was geweest, uitvoeriger kunnen bespreken) een begrijpelijk geheel. Wanneer je op die manier naar mensen kijkt, valt dat te vergelijken met vanuit de ruimte naar de aarde kijken: de details gaan weliswaar verloren, maar je krijgt een totaal nieuwe blik op iets heel vertrouwds.

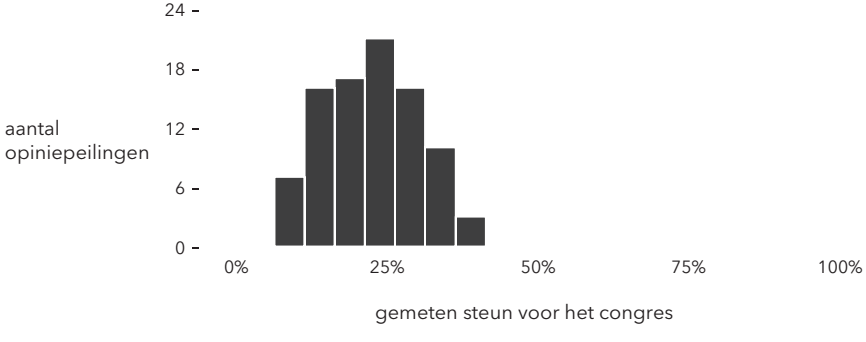
Wat kunnen we uit de bovenstaande grafiek opmaken? Je zou deze basisvorm – de normale verdeling – maar al te gemakkelijk klakkeloos voor waar kunnen aannemen, omdat je die op grond van voorbeelden uit leerboeken waarschijnlijk al verwachtte. Toch hadden de scores net zo goed heel sterk naar een van de twee kanten kunnen overhellen. Dat is bij persoonlijke voorkeuren namelijk vaak het geval. Neem bijvoorbeeld de beoordelingen van pizzarestaurants op Foursquare, die meestal heel positief zijn:[3]

bezoekersbeoordelingen van pizzarestaurants in New York op de Foursquare-schaal van 0 tot 10

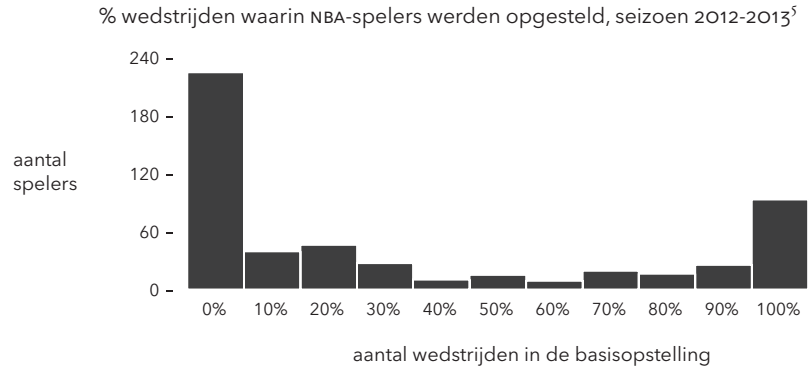


Of neem de recente steun voor het Amerikaanse Congres, die juist scheef naar links is verdeeld, omdat politici de morele tegenpool van pizza's vormen:[4]

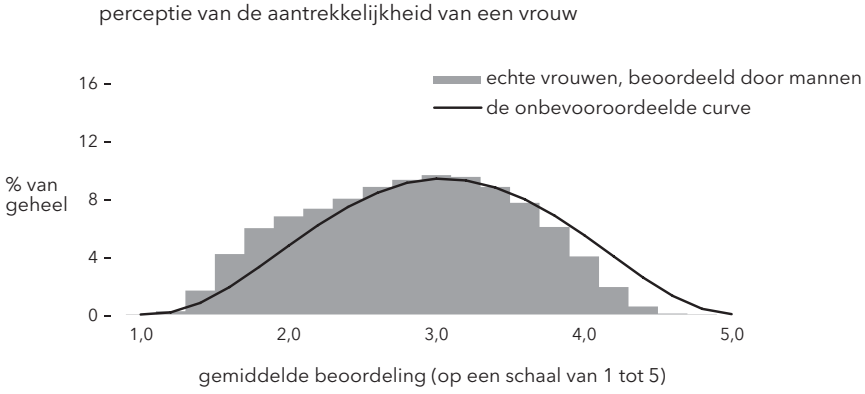
steun voor het Congres in voornaamste opiniepeilingen sinds november 2008



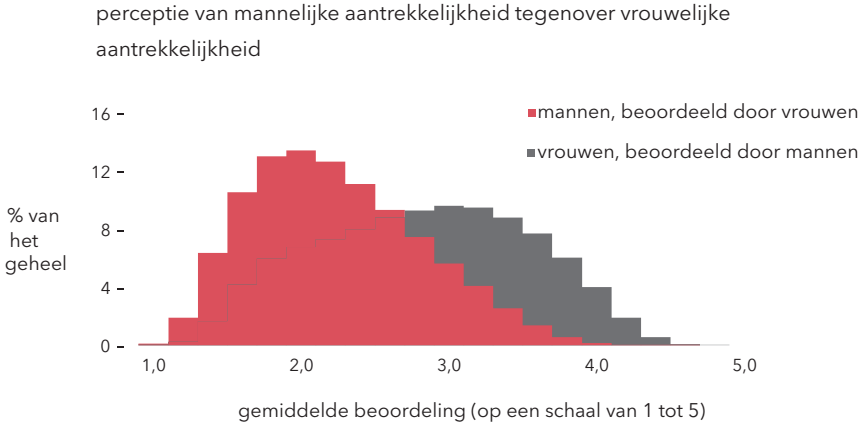
Onze eerdere beoordelingsgrafiek van mannen over vrouwen is overigens unimodaal, wat betekent dat de vrouwenscores vooral rond één specifieke waarde zijn gegroepeerd. Dat is op zich niet zo heel opzienbarend, maar er zijn heel wat situaties die meerdere modi, oftewel 'standaardwaarden', hebben. Als je Amerikaanse NBA-spelers grafisch weergeeft, gemeten naar het aantal maal dat ze in het vorige seizoen werden opgesteld, krijg je aan beide kanten een heel stel atleten op een kluitje, met bijna niemand in het midden:



Wat je daar ziet, is dat coaches blijkbaar denken dat een goede speler meteen al wel, of niet, goed genoeg is om op te stellen en daardoor dus wel of niet op de bank zit. Het is overduidelijk een binair systeem. In de eerdere beoordelingsdata zou je dat kunnen vergelijken met mannen die vrouwen als groep of 'bloedmooi', of 'spuuglelijk' vinden, maar dat is het dan ook wel. Alsof schoonheid net als toptalent in basketbal een kwestie is van 'je hebt het of je hebt het niet'. Uit de grafiek waarmee we begonnen, blijkt alleen iets anders. Ben je in de data op zoek naar inzichten, dan is het vaak een kwestie van de uitkomsten tegen dit soort tegenfeitelijkheden afzetten. Heb je een veelvoud aan mogelijkheden, dan is een voor de hand liggend resultaat soms des te opzienbarender. De grafiek hierboven lijkt bijna op een zogenoemde symmetrische ronde bètaverdeling, een vorm die vaak als model wordt gebruikt om onbevooroordeelde basisbesluiten weer te geven. Voor de duidelijkheid voeg ik die hier toe:



Onze data uit de echte wereld wijken maar een klein beetje van de ideale formule af (6 procent),[6] wat wil zeggen dat deze grafiek van mannelijke wensen min of meer neerkomt op wat je op je klompen had kunnen aanvoelen. Het is zelfs een van die standaardlesvoorbeelden die ik zo even terloops noemde. De curve is namelijk voorspelbaar, gecentreerd en misschien zelfs een tikkeltje saai. Maar wat zou dat? Nou, in dit geval maakt die saaiheid het juist bijzonder, want die impliceert dat de individuele mannen die de beoordelingen gaven evengoed voorspelbaar, keurig normaal verdeeld en bovenal onbevooroordeeld zijn. Wanneer je bedenkt dat mannen waarschijnlijk elke dag topmodellen, porno, hoezenpoezen en Lara Croft-achtige fembots voorgeschoteld krijgen, en niet te vergeten de bierreclames van Bud Light, en de meest onbetrouwbare van allemaal: Photoshop-bewerkingen, dan is het volgens mij eerder een klein wonder dat de mening van mannen over de aantrekkelijkheid van een vrouw er nog zo normaal uitziet. Je zou met je gezonde verstand bijna denken dat mannen er irreële verwachtingen van vrouwelijk schoon op na zouden móéten houden, en toch zien we hier dat dat niet het geval is. Ze zijn in elk geval een stuk milder dan vrouwen, die op hun beurt als volgt oordelen:



De rode curve is gecentreerd rond een waarde die maar net het onderste kwart van de schaal beslaat; met andere woorden, slechts één op de zes mannen is in absolute zin 'bovengemiddeld'. Sexappeal wordt niet vaak op zo'n manier gekwantificeerd, dus laat ik een wat bekendere context nemen. Stel je voor dat deze grafische voorstelling over IQ-scores ging. In dat geval zouden we het hebben over een wereld waarin vrouwen denken dat 58 procent van de mannen een hersenbeschadiging heeft opgelopen.[7]

Nu zijn de mannen op OkCupid niet direct lelijk te noemen. Dat heb ik onderzocht door een aselecte steekproef van onze gebruikers af te zetten tegen een vergelijkbare aselecte steekproef van een sociaal netwerk, waarbij ik voor beide groepen dezelfde scores kreeg. Wat blijkt? Patronen zoals die hierboven zie je terug op elke Amerikaanse datingsite die ik heb bekeken: op Tinder, op Match.com en op DateHookup. Daarbij hebben we het dan over ongeveer de helft van alle singles in Amerika, want dat is het gezamenlijke bereik van deze sites.[8] Je zou zeggen dat mannen en vrouwen er op seksueel gebied iets andere rekensommetjes op na houden. *Harper's* verwoordde dat heel raak: 'Vrouwen zijn geneigd spijt te hebben van de seks die ze hebben gehad, mannen van de seks die ze niet hebben gehad.'[9] In de data zie je mooi geïllustreerd hoe dat werkt. En laat ik daar dit aan toevoegen: de mannen hierboven hebben zeker weten vreselijke spijt.

Een bètaverdeling geeft grafisch weer wat je zou kunnen zien als de uitkomst van heel veel keer een muntje opgooien.[10] De curve geeft de overlappende waarschijnlijkheid weer van talloze op zichzelf staande binaire gebeurtenissen. In dit geval is het mannelijke muntstuk eerlijk en komt het even vaak neer op kop (wat ik in dit geval gelijkstel aan positief) als op munt. Alleen is er in onze data blijkbaar geknoeid met het vrouwelijke muntstuk, want bij hen levert elke viermaal opgooien maar één keer kop op. Een groot aantal natuurlijke processen, waaronder het weer, kan met een bètaverdeling worden weergegeven, en dankzij de nogal obsessieve archiveringsmanie van de bouwer van een bepaalde weersite konden we onze persoonlijke beoordelingen aan historische klimaatpatronen koppelen. Wat blijkt? Het mannelijk oordeel ligt dicht bij de rekenkundige functie die de bewolking boven New York voorspelt, en volgens diezelfde cijfers houdt de vrouwelijke psyche zich ergens op waar het nog net een tikkeltje duisterder is dan in Seattle.

Bij het eerste van de drie hoofdonderwerpen die in dit boek aan de orde komen – de data over het contact tussen twee mensen – zal ik die aanpak volgen. We beginnen bij sexappeal: hoe verandert en ontstaat die? We zullen zien dat een vrouw op haar eenentwintigste in feite al haar beste tijd heeft gehad, en hoe belangrijk een opvallende tatoeage kan zijn. Daarna laten we de vleselijke verbintenissen achter ons en kijken naar wat tweets ons kunnen vertellen over moderne communicatie, en vriendschappen op Facebook over de stabiliteit van een huwelijk. Profielfoto's op internet zijn zowel een zegen als een vloek, want bijna elke site (Facebook, vacaturesites, en ja, uiteraard ook datingsites) verwordt daardoor tot een schoonheidswedstrijd. We zullen zien wat er gebeurt wanneer OkCupid de foto's – op hoop van ze-

gen – een dag van de site verwijdert. Liefde is inderdaad niet blind, terwijl dat juist wel het geval zou moeten zijn.

In deel 2 richten we ons op de data over verdeeldheid. We beginnen daarbij met de grootste splijtzwam tussen mensen: etnische afkomst. We kunnen dat onderwerp nu voor het eerst op individueel niveau bekijken, en onze vertrouwelijke data brengen opvattingen aan het licht waar de meeste mensen in het openbaar nooit voor zouden durven uitkomen. We zullen zien dat racistische vooroordelen niet alleen hardnekkig zijn, maar bovendien stelselmatig en bijna woordelijk (nou ja, cijferlijk) op elke site terugkomen. Racisme kan daarbij ook iets heel innerlijks zijn – van één man, zijn vooroordeel en een toetsenbord. We zullen zien wat zoekopdrachten op Google ons kunnen vertellen over het meest gehate woord in heel Amerika, en wat dat woord zegt over ons land. Vervolgens kijken we aan de hand van een dataset die duizendmaal krachtiger is dan alles wat tot nu toe voorhanden was, naar de mate waarin fysieke schoonheid verdeeldheid kan zaaien. Lelijkheid brengt ontstellende maatschappelijke kosten met zich mee, die we nu eindelijk kunnen kwantificeren. Dat brengt ons bij wat Twitter over onze korte lontjes onthult. Dankzij dit platform kunnen mensen van minuut tot minuut met elkaar verbonden zijn, maar het kan ons even snel uit elkaar drijven. De collectieve woede die Twitter kan ontketenen geeft die al eeuwenoude menselijke vorm van samenscholing – de woedende meute – een nieuw geweldsmiddel in handen. We zullen zien of dat ons misschien ook nieuwe inzichten oplevert.

Eenmaal aangekomen bij het derde deel laten we de data van twee mensen die, goed- of kwaadschiks, met elkaar in contact komen achter ons en richten ons op het individu. We zullen erachter komen hoe onze etnische, seksuele en politieke identiteit tot uitdrukking komt, waarbij we vooral kijken naar de woorden, de beelden en de culturele pijlers waarmee mensen zich presenteren. Hieronder zie je de vijf meest kenmerkende frases waarmee een blanke Amerikaanse vrouw zichzelf omschrijft:

mijn blauwe ogen
rood haar en
vierwielaandrijving
country girl
houdt van het buitenleven

Is dit een haiku van Carrie Underwood of zijn het mijn data? Zeg jij het maar! We zullen kijken naar wat mensen in het openbaar zeggen. We zul-

len zien hoe mensen zich gedragen en zich uiten wanneer ze zich onbespied wanen, en daarbij speciaal letten op de momenten dat de 'etiketten' en het gedrag niet met elkaar stroken, zoals bij biseksuele mannen, die onze ideeën over keurig afgebakende identiteiten danig op de proef stellen. Daarna zullen we aan de hand van een groot aantal bronnen – Twitter, Facebook, Reddit, en zelfs Craigslist – naar onze thuissituaties kijken, zowel fysiek als anderszins. En we sluiten af met een voor de hand liggende vraag voor een boek als dit: hoe behoud je je privacy in een wereld waarin dit soort speurwerk mogelijk is?

Wat we telkens weer zullen zien is dat internet opwindend, keihard, liefdevol, vergevingsgezind, onbetrouwbaar, sensueel en woedend kan zijn. Dat is ook niet zo raar, want internet bestaat uit mensen. Terwijl ik al deze informatie verzamelde werd ik me ervan bewust dat deze data niet ieders leven weten te vangen. Als je geen computer of smartphone hebt, kom je in dit boek niet voor. Ik kan dat probleem op dit moment alleen maar erkennen, omzeilen en wachten tot het zich oplost.

Laat ik daarbij meteen aantekenen dat het bereik van sites als Twitter en Facebook, en zelfs van mijn datinggegevens, opmerkelijk volledig is. Als je weinig ervaring hebt met een van deze sites weet je dat misschien niet direct op waarde te schatten. In de Verenigde Staten heeft pakweg 87 procent van de bevolking toegang tot internet,[11] en dat cijfer geldt voor alle demografische groepen.[12] Van stad tot platteland, rijk of arm, zwart, Aziatisch, blank of van Latijns-Amerikaanse afkomst, iedereen zit op internet. Alleen onder bejaarden en laagopgeleiden ligt het internetgebruik beduidend lager (zo'n 60 procent). Dat is dan ook de reden dat ik de leeftijdsgrens in dit boek behoorlijk laag heb gelegd, namelijk al bij vijftig jaar, en waarom ik geen uitspraken doe over opleidingsniveau.

Meer dan één op de drie Amerikanen kijkt elke dag op Facebook, wereldwijd heeft de site inmiddels circa 1,3 miljard gebruikers.[13] Als je bedenkt dat ongeveer een kwart van de wereldbevolking jonger dan veertien is, wil dat dus zeggen dat rond de 25 procent van alle volwassenen ter wereld een Facebook-account heeft. Bij de datingsites die in dit boek aan de orde komen hebben zich in de afgelopen drie jaar zo'n 55 miljoen Amerikanen geregistreerd, en zoals ik hierboven al zei, betekent dit dat de helft van alle singles een account heeft. Twitter is demografisch gezien al helemaal interessant, en niet alleen omdat het een blits technologisch succesverhaal is. Dat een groot aantal wijken in San Francisco is opgeknapt, is bijna geheel op het conto van dit bedrijf te schrijven. Twitter is een buitengewoon populistisch medium in die zin dat het platform heel 'open' is, maar ook vanwege de

mensen die er actief zijn.[14] Je ziet geen duidelijke verschillen tussen mannen of vrouwen, en mensen met alleen een middelbareschoolopleiding twitteren even vaak als afgestudeerden. Latino's gebruiken Twitter evenveel als blanken, maar zwarten juist dubbel zoveel. En dan heb je Google natuurlijk nog. Als 87 procent van de Amerikanen internet gebruikt, dan gebruikt 87 procent daarvan Google.

Deze hoge aantallen bewijzen allerminst dat ik ergens een volledig beeld van heb, maar ze wijzen er wel op dat zo'n beeld eraan zit te komen. Bovendien zou volledigheid hoe dan ook niet de vijand moeten zijn van 'beter dan ooit tevoren'. De dataset die ik gebruik omvat duizendmaal meer mensen dan die waarover grote onderzoeksbureaus als Gallup of Pew beschikken, dat spreekt voor zich. Wat echter minder voor de hand ligt, is dat die vergeleken met die van het merendeel van wetenschappelijk gedragsonderzoek ook veel breder en veelzijdiger is.

Het is binnen de sociale wetenschappen een bekend probleem (hoewel het zelden in de openbaarheid wordt gebracht) dat bijna alle grondbeginselen gebaseerd zijn op experimenten met groepjes studenten. Zo kon ik tijdens mijn studie 25 dollar verdienen door naar het Mass General-ziekenhuis te gaan en daar een uur lang gas met een licht radioactieve marker in te ademen, om daarna een of ander psychotechnisch testje uit te voeren terwijl er foto's van mijn hersenen werden gemaakt. Je houdt er niets aan over, zeiden ze. Vergelijk het maar gewoon met een jaar lang in een vliegtuig zitten, zeiden ze. Het stelt echt niks voor, zeiden ze. Wat ze niet zeiden – en wat ik toentertijd niet doorhad – was dat ik, toen ik daar met een halve kater in een of ander CAT-scangeval lag en woordjes moest oplezen terwijl ik met mijn voet op knopjes drukte, model stond voor de 'gemiddelde' man. Een vriend van me deed ook mee aan het onderzoek. Hij was net als ik een blanke student. Ik durf te wedden dat dat voor de meeste proefpersonen gold. We waren, met andere woorden, allesbehalve gemiddeld.

Ik snap heel goed hoe dat gebeurt. In de praktijk is het meestal moeilijker om een representatieve dataset te krijgen dan het experiment dat je op je werkkamer hebt bedacht. Dus als ambitieuze hoogleraar of postdoctoraal onderzoeker neem je dan iets wat een 'gelegenheidssteekproef' wordt genoemd – kortom, de studenten op de universiteit waar je werkt. Dat levert alleen wel een heel groot probleem op, vooral wanneer je onderwerpen als opvattingen en gedrag wilt onderzoeken. Er bestaat zelfs een naam voor: WEIRD-onderzoek, waarbij het Engelse woord (dat onder meer 'raar' betekent) meteen ook verwijst naar de afkorting van de woorden voor 'blank', 'hoogopgeleid', 'technisch', 'rijk' en 'democratisch'.[15] Het gros van de onder-

zoeksartikelen in de sociale wetenschap is WEIRD.*

Bij mijn data speelt een aantal van dezelfde problemen. Zo zal het nog wel even duren voor je bij digitale data 'technisch' helemaal van het lijstje kunt schrappen. Maar omdat alles wat met technologie te maken heeft vaak als iets elitairs wordt gezien – een beeld dat veel mensen in de sector zelf overigens maar al te graag in stand houden –, wil ik graag wat nuances aanbrengen tussen de ondernemers en de durfkapitalisten die je op de publieke podia grote gebaren ziet maken en met hoofdmicrofoontjes ronkende verhalen hoort opdissen (mensen die meestal inderdaad behoorlijk WEIRD zijn) en de gebruikers van de sites zelf. Die laatste groep is juist heel normaal, en daar kunnen ze ook niets aan doen, want gebruikmaken van sites als Twitter, Facebook en Google ís juist de norm.

Wat de authenticiteit van de data betreft: omdat internet tegenwoordig zo'n belangrijke plaats in het dagelijks leven inneemt, reguleert die zichzelf als het ware. Neem de data van OkCupid. Als gebruiker vertel je de site waar je woont, of je man of vrouw bent, hoe oud je bent en naar wie je op zoekt bent. Vervolgens helpt de site je iemand te vinden met wie je kunt afspreken voor een kop koffie of een biertje. Je profiel zou een goede weergave moeten zijn van wie je bent, van de echte jij dus. Upload je een foto van iemand die er knapper uitziet dan jij of doe je je jonger voor dan je bent, dan levert dat je waarschijnlijk meer afspraakjes op. Maar stel je dan meteen even het moment voor dat je die ander daadwerkelijk ontmoet, die op basis van je online profiel bepaalde verwachtingen koestert. Als je daar in het echt niet op lijkt, is het afspraakje eigenlijk al mislukt. Dit is maar één voorbeeld van een veel bredere trend: dat naarmate de online- en de offlinewereld meer met elkaar samenvallen, de ingebouwde sociale druk op internet ervoor zal zorgen dat de ergste impulsen tot fabuleren worden ingetoomd.

De mensen die gebruikmaken van al deze sites – van de datingsites, de socialenetwerksites en de nieuwsvergaarbakken – zoeken zich op de tast een weg door het leven, zoals mensen dat altijd al hebben gedaan. Alleen doen ze dat nu op mobieltjes en laptops. Zo hebben ze bijna onbedoeld een uniek archief gecreëerd. Overal ter wereld bevatten talloze databases jarenlange gegevens over verlangens, meningen en chaos. Omdat deze data met een kristalheldere precisie worden bewaard, kunnen ze op den duur worden geana-

* Het tijdschrift *Slate* schreef daar dit over: 'WEIRD-proefpersonen, uit landen die amper 12 procent van de wereldbevolking vertegenwoordigen, verschillen van andere bevolkingsgroepen wat betreft morele-besluitvormingsprocessen, redenatiestijl, rechtvaardigheidsgevoel en zelfs visuele waarnemingen. Dat komt doordat een groot deel van het menselijk gedrag en van onze waarnemingen gebaseerd is op de leefomgeving en de context waarin we zijn opgegroeid.'

lyseerd, en dat ook nog eens met een reikwijdte en flexibiliteit die we tien jaar geleden niet hadden kunnen bevroeden.

Ik ben er een paar jaar mee bezig geweest die data te verzamelen en uit te pluizen, niet alleen van OkCupid, maar ook van bijna elke andere vooraanstaande site. Toch heb ik nooit echt een einde kunnen maken aan de knagende twijfel – die gezien mijn zwak voor luddieten des te irritanter is – dat een boek over internet schrijven net zoiets is als een mooie tekening maken over een film. Waarom zou je? Tja, van zo'n vraag kan ik dus 's nachts wakker liggen.

Er bestaat een geweldige documentaire over Bob Dylan, *Don't Look Back*, die ik tijdens mijn studie een paar keer heb gezien; mijn beste vriend, Justin, studeerde filmwetenschappen. Ergens in die film, op een afterparty, raakt Bob in een felle discussie verwikkeld met iemand die misschien wel of misschien ook niet glas op straat heeft gegooid. Ze hebben allebei duidelijk al iets te veel op. Het hoogtepunt van dat gesprek is de volgende dialoog, en die herinner ik me na vijftien jaar dus nóg:

> Dylan: Ik ken wel duizend gasten die er net zo uitzien als jij en net zo praten als jij.
> Gozer op feestje: Lazer toch op, man. Jij bent een hotemetoot. Hè?
> Dylan: Ja, dat weet ik toch? Ik weet ook wel dat ik een hotemetoot ben.
> Gozer op feestje: Ja, ik weet dat jij dat weet.
> Dylan: Ik ben een veel grotere hotemetoot dan jij, man.
> Gozer op feestje: Ik ben een kleine hotemetoot.
> Dylan*: Right.*

En dan komt er iemand tussenbeide, waarna ze het met z'n allen over poëzie gaan hebben. Zo'n feestje dus. Waar het me hier om gaat is het volgende: popster of niet, tot nu toe waren het uitsluitend de hotemetoten die het tot de geschiedenisboekjes schopten. Zo werd ons overkoepelende verhaal altijd verteld: aan de hand van de levens van de veroveraars, de magnaten, de martelaars, de bevrijders en zelfs van de schurken (of juist vooral van de schurken!). Zo hebben we onze vooruitgang vastgelegd, al sinds de tijd dat we aan de oevers van een paar ziltige rivieren zaten tot waar we nu dan ook zijn. Van farao Narmer uit 3100 v.Chr., van wie we de naam nu nog kennen, tot aan Steve Jobs en Nelson Mandela.[16] Zo heeft de mens de wereld tot nu toe altijd geordend: via een heroïsch kader. Narmer was de eerste op een eeuwenoude lijst van koningen. De schriftgeleerden veranderen en de lijst wordt alleen

maar langer. Neem de jaren zestig, met *power to the people* en nog zo wat. Dat is denk ik het perfecte voorbeeld: het tijdperk van Lennon en McCartney, van Dylan en Hendrix, en niet van 'een of andere gozer op een feestje'. Het bestaan van de man in de straat was niet de moeite waard om vast te leggen, behalve wanneer hij het pad van een legende kruiste.

Aan die asymmetrie komt nu langzaam een einde, want het gemor en geroezemoes van het gepeupel wordt nu voor het eerst wél bewaard. Internet heeft allerlei zaken waarvoor mensen zich inspannen gedemocratiseerd: journalistiek, fotografie, porno, liefdadigheid en comedy. Hopelijk zal dat met ons overkoepelende verhaal uiteindelijk ook gebeuren. Het geluid van 'gewone mensen' mag dan nog wat embryonaal en ongeraffineerd klinken, ik heb dit boek geschreven om de vage patronen die ik wel degelijk bespeur te tonen, en ik ben niet de enige. Het is als het geluid van de naderende trein wanneer je een oor tegen de spoorrails drukt. De datawetenschap is verre van perfect en alleen al door selectiecriteria is er sprake van vertekening. En zo zijn er nog wel meer tekortkomingen die we zullen moeten begrijpen, onderkennen en zien te omzeilen. Maar de afstand tussen wat zou kunnen zijn en wat er al is, wordt met de dag kleiner, en het gaat mij om het uiteindelijke moment waarop die twee zullen samenvallen.

Ik weet dat er een heleboel mensen zijn die allerlei hoogdravende uitspraken doen over data, en ik ga hier echt niet beweren dat dit cijfermateriaal de loop van de geschiedenis zal veranderen – zeker niet op de manier waarop dat gold voor de verbrandingsmotor of het staal –, maar ik denk wel dat data zullen veranderen wat geschiedenis ís. Omdat we het verleden nu verder kunnen uitdiepen. Het kan méér worden. In tegenstelling tot kleitabletten, tot papyrus, tot papier, kranten, celluloid of fotostock, is diskruimte goedkoop en bijna onuitputtelijk. Op een harde schijf is er niet alleen maar plaats voor de helden. Aangezien ik zelf geen held ben en eigenlijk het liefst dingen doe met mijn vrienden en familie, en geniet van de kleine dingen in het leven, betekent dat voor mij ook echt iets.

Hoe graag ik ook zou willen dat WhoBeefed81 samen met jou en mij straks zij aan zij op dezelfde pagina zal worden genoemd als de president wanneer er in de toekomst over dit decennium wordt geschreven, ik ga ervan uit dat gewone mensen zo goed als anoniem zullen blijven, wat hier in dit boek trouwens ook het geval is. Daar kunnen zelfs de beste data geen verandering in brengen. Maar ook wij zullen worden meegeteld. Als iemand over tien, twintig of honderd jaar deze tijdsperiode probeert te duiden en de ontwikkelingen wil begrijpen – dus bijvoorbeeld wil snappen hoe het legaliseren van het homohuwelijk er zowel voor zorgde dat homoseksualiteit

meer geaccepteerd werd als dat deze legalisatie die acceptatie weerspiegelde, of hoe het dorpsleven in Azië werd ontworteld en vervolgens in grote stedelijke gebieden weer werd opgebouwd –, zullen al die verhalen data bevatten. Sterker nog: ze zullen daar zelfs uit bestaan, en van Facebook, Twitter en van Reddit en consorten afkomstig zijn. En mocht dat niet het geval zijn, dan heeft onze vermeende schrijver gefaald.

Met *Dataclysm*, de oorspronkelijke porte-manteautitel voor dit boek, heb ik gepoogd dit allemaal te ondervangen. *Kataklysmos* is het Griekse woord voor de zondvloed en ik zinspeel daar om twee redenen op. Allereerst vormen de data zelf een ongeëvenaarde vloedgolf. Wat vandaag de dag verzameld wordt, reikt zo diep dat het onuitputtelijk is; dat kun je met gemak vergelijken met veertig dagen en veertig nachten stortregen in plaats van de paar spetters van vroeger. Daarnaast er is ook sprake van hoop dat het een omslag zal betekenen en dat zowel het in groei achtergebleven begrip van gisteren als de beperkte visie van tegenwoordig door de vloedgolf zal worden weggespoeld.

Dit boek is een aaneenschakeling van vignetten, kleine vensters die ons een inkijkje geven in onze levens, in wat ons samenbrengt, wat ons uit elkaar drijft en wat ons maakt tot wie we zijn. Naarmate de data zich opstapelen zullen die venstertjes almaar groter worden, maar er is nu al voldoende te zien, en de eerste glimp is meestal toch het spannendst. Dus klim maar op de vensterbank, ik geef je wel een knietje.