Big Data

Big Data - veelal met hoofdletters geschreven, als om te onderstrepen dat we hier met iets heel bijzonders te maken hebben - is al langere tijd een veelbesproken onderwerp. Je kunt tegenwoordig nauwelijks meer naar een symposium gaan of de krant openslaan, of je wordt er mee geconfronteerd. De teneur is doorgaans dat de mensheid met Big Data in een nieuwe wereld terecht zal komen, met onbegrensde mogelijkheden en kansen.

WFZ
Waarborgfonds voor de Zorgsector

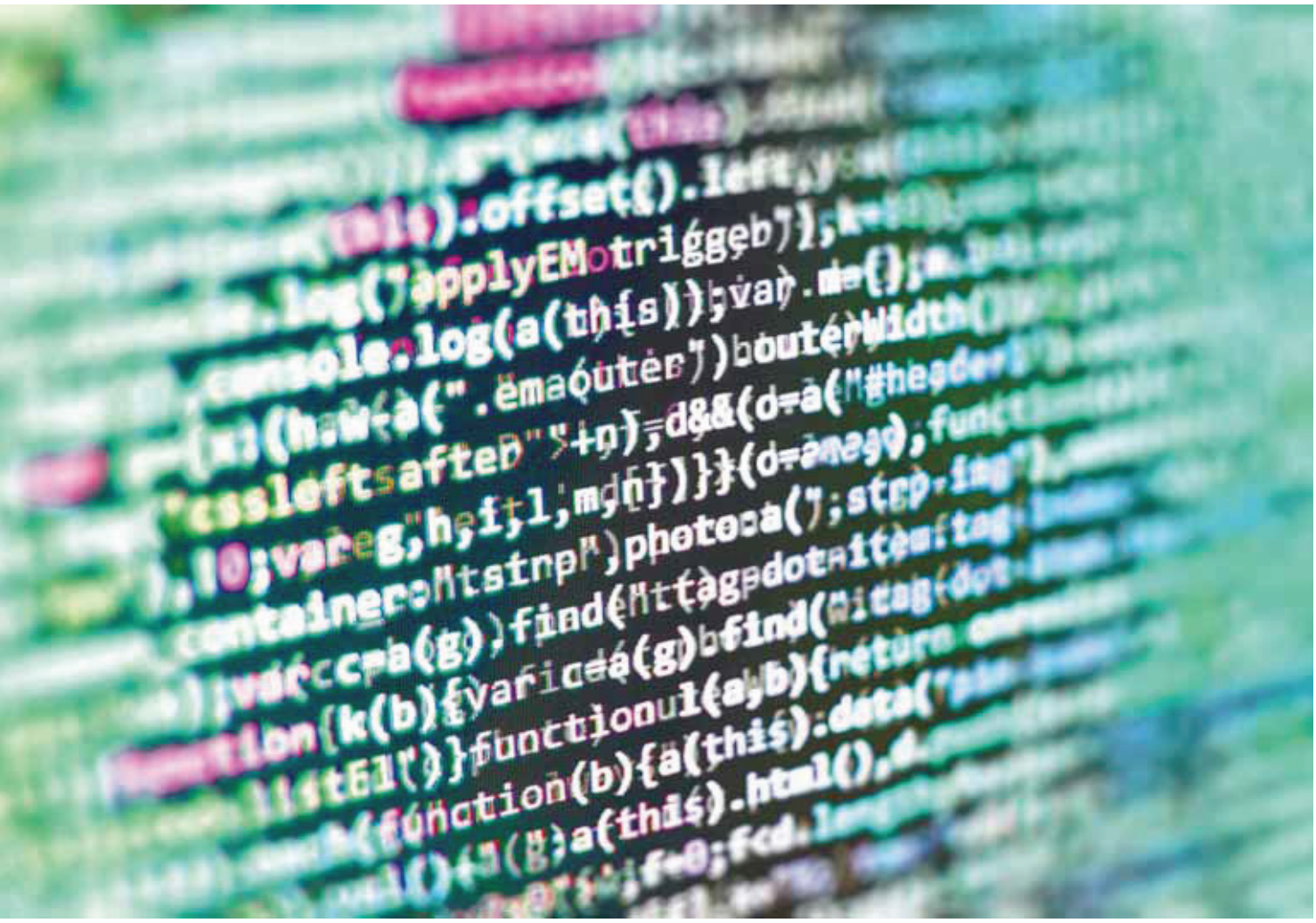
Herman Bellers, directeur WFZ

Nieuwe industrieën zullen opkomen, bestaande organisaties, werkwijzen en activiteiten zullen verdwijnen. Ook de wetenschap zal volgens sommige voorspellingen een nieuwe fase ingaan. Immers, als de datasets maar groot genoeg worden, gaan we onvermoede verbanden vinden tussen allerlei fenomenen. Niet iedereen deelt deze optimistische visioenen. Een van de meest uitgesproken sceptici is professor Nassim Taleb[1] die zijn opvatting tijdens een congres in Amsterdam als volgt verwoordde: 'Het hele idee is bullshit. Mensen die van Big Data houden zijn ofwel geen wetenschappers, ofwel ze hebben er iets bij te winnen.'

Ook als je nauwelijks weet wat er onder Big Data wordt verstaan, zal de laatste opmerking van Taleb op velen waarschijnlijk plausibel overkomen. Immers, 'omzet gedreven' advisering en communicatie is van alle tijden. Dat dus ook een 'hot' thema als Big Data wordt gehypet, als nieuwe successtrategie voor organisaties in de markt wordt gezet, en dat door consultants in hun marketingcommunicatie gouden bergen worden beloofd, mag nauwelijks verbazing wekken.

De opmerking van Taleb over het niet-wetenschappelijke karakter van Big Data is wellicht wat minder evident. Deze opmerking heeft betrekking op een van de kernelementen van Big Data, namelijk het (geautomatiseerd) zoeken naar verbanden ('correlaties') tussen grote aantallen feiten, personen en/of gebeurtenissen, en op basis hiervan beslissingen nemen (ook wel aangeduid als 'predictive analysis'). Veel van de inhoudelijke kritiek op Big Data (van Taleb en anderen) is samen te vatten als: correlatie is niet hetzelfde als causaliteit. Huiselijker gezegd: een statistisch verband is geen bewijs voor een inhoudelijk of oorzakelijk verband. Zo bestaat er een sterke statistische correlatie tussen het aantal kerken en de omvang van criminaliteit in een gemeenschap. De cijferreeksen van beide fenomenen gaan namelijk steeds keurig gelijktijdig op en neer. Maar dit geeft nog geen antwoord op de vraag hoe de oorzaak-gevolg relatie ligt: vormen kerken een bron van crimineel gedrag, of is omgekeerd juist de omvang van criminaliteit bepalend voor de bouw of sloop van kerken? In dit geval is het antwoord: geen van beide, want er is helemaal geen inhoudelijk verband. Zowel het aantal kerken als de omvang van criminaliteit hangen direct samen met een derde variabele (die in de analyse buiten beeld blijft) namelijk de bevolkingsomvang.

Een ander voorbeeld van dergelijke 'spurious correlation' is die tussen bierconsumptie en het hebben van sproeten. Ook hier is het niet zo dat men dorst krijgt van sproeten, of dat bier sproeten veroorzaakt. De verklaring is dat zowel bierconsumptie als sproeten direct samenhangen met een onzichtbare derde variabele, namelijk het aantal uren zon. En dan zijn er natuurlijk nog de gevallen van volstrekt toevallige samenloop van gebeurtenissen in de tijd. De website www.tylervigen.com geeft hiervan tal van humoristische voorbeelden, zoals de 'samenhang' tussen het aantal mensen dat overlijdt door verdrinking in hun eigen zwembad en het aantal films waarin Nicolas Cage optreedt, tussen echtscheidingen en margarineconsumptie, enzovoorts. Het verschijnsel van ogenschijnlijke verbanden is een serieus probleempunt bij Big Data. Op de eerste plaats vanwege de onderzoeksmethodiek. In de 'traditionele' wetenschappelijke aanpak wordt getracht kennis over de wereld om ons heen te verwerven door een gestructureerde werkwijze, waarbij eerst vermoedens over inhoudelijke verbanden worden opgesteld en uitgewerkt: theorieën dan wel hypothesen. Deze worden vervolgens aan feiten getoetst. Op basis hiervan wordt bezien of de feiten in overeenstemming zijn met de theorie, dan wel deze onderuithalen (falsificatie-benadering).



Bij deze aanpak komt zoiets onzinnigs als de bier/sproeten-casus dus niet eens in een onderzoek terecht, aangezien hierover geen theorie zal worden opgesteld (want er is geen vermoeden van een inhoudelijk verband). Bij Big Data analyse wordt de werkwijze echter omgedraaid. Het begint met het zoeken naar data-correlaties, die vervolgens de basis vormen voor conclusies over inhoudelijke verbanden (vandaar de opmerking van Taleb over het 'onwetenschappelijke' van Big Data). Wat anders verwoord: het vertrekpunt voor onderzoek en analyse ligt niet langer bij een probleem en de vraag hoe dit op te lossen, of bij een verondersteld verband en de vraag of dit er ook feitelijk wel is, maar bij de data die voorhanden zijn.

Op de tweede plaats (in samenhang met het voorgaande) is 'spurious correlation' bij Big Data een probleem vanwege dat 'Big'. Immers, naarmate de datasets diverser en omvangrijker worden, neemt automatisch de kans toe dat er in het zoekproces wel ergens statistische verbanden opduiken. De crux bij Big Data-analyse wordt dan ook hoe je kaf van koren kunt scheiden, en de relevante verbanden kunnen onderscheiden van de toevallige onzin. De beeldende metafoor die in dit verband wel wordt gebruikt is hoe je nog het gouden muntje kunt vinden in de steeds groter wordende berg mest.

Big Data is zonder twijfel een actueel en belangrijk thema. Begrijpelijk, want er is nauwelijks een onderwerp te bedenken waar een grote hoeveelheid zorgvuldig verzamelde data en een grondige analyse daarvan geen toegevoegde waarde heeft. Ook in de zorg zijn op dit moment tal van interessante initiatieven gaande rond gegevensverzameling en -analyse, waarbij er mede door de toenemende ICT-toepassingen nog werelden te winnen lijken. Maar dat Big Data de belofte van alomvattend wondermiddel zal inlossen lijkt vooralsnog te betwijfelen. Taleb en de zijnen hebben serieuze kritiekpunten die enige relativering van de huidige Big Data hype rechtvaardigen.

[1] Bestsellerauteur, bekend van onder meer de risicomanagement klassieker 'De Zwarte Zwaan'.