Big Data in HR– een typisch geval van 'Dust Bowl empiricism'?

Patrick Vermeren en Tom Speelman
Oktober 2014

Big Data en HR Analytics zijn hot

De laatste maanden boomt een nieuwe hype binnen HR: het gebruik van Big Data om betekenisvolle uitkomsten te voorspellen zoals toekomstige prestaties, risico op vrijwillig ontslag enzovoort. 'Big data' is de term die men gebruikt voor een verzameling data die enorm groot en vaak ongestructureerd en complex zijn en die tot op heden vaak niet werden gebruikt of niet konden worden gebruikt, omdat de software of database management systemen deze simpelweg niet konden behappen. Big Data analyses worden al gebruikt in een aantal domeinen zoals de distributie (om klanten te segmenteren en gebeurtenissen of veranderingen in hun leven te ontdekken – zo kunnen ze vrij accuraat afleiden uit het koopgedrag of iemand recent een kind kreeg, helpen ze bij weersvoorspellingen en in de fysica (bijvoorbeeld het verwerken van de miljoenen data van de sensoren van de Large Hadron Collider van CERN in Geneve). Merk daarbij op dat in de meeste gevallen Big Data worden gebruikt voor 'detectie' – terwijl we in HR vooral claims horen over 'predictie' (voorspellen). In het experiment van CERN bijvoorbeeld was het Engler-Brout-Higgs deeltje reeds voorspeld en ging men op zoek naar bewijzen in de gigantische hoeveelheid data (detectie) die het experiment met de deeltjesversneller opleverde.

Bijna alle aanbieders voor het verwerken van uw Big Data ('Business Analytics') beweren dat ze *evidence based* werken, waardoor ze alleszins bij een aantal mensen aan geloofwaardigheid winnen. Evidence based begint immers stilaan te gelden als een autoriteitsargument. Maar is dat echt zo? Tenslotte zijn de meesten zuivere IT-bedrijven waarin geen psychometrici of statistici met kennis van de problemen van willekeurige verbanden in de sociale wetenschappen werken. In elk geval, dat borstgeklop met alle grote claims over Big Data en de voorspelkracht die moet leiden tot betere beslissingen, deden onze alarmbellen rinkelen en we konden ons niet ontdoen van de gedachte dat dit wel eens een typisch geval van 'Dust Bowl empiricism' zou kunnen zijn en dus begonnen we aan een speurtocht.

Een conferentie van AHRI in Melbourne ("Australian Human Resources Institute"- het grootste HR-instituut in Australië) eind augustus 2014 bood een perfecte gelegenheid om een en ander uit te zoeken. De eerste auteur van dit artikel woonde een presentatie bij die werd gehouden door een bedrijf dat door een Indische zakenman werd opgericht – verder BDC genoemd. We hadden ook een gesprek met een Belgische marketeer van een ander bedrijf dat we BDC2 zullen noemen. Deze laatste gaf ons het meeste inzicht in hun programmatie en gebruikte algoritmes. Maar laat ons u eerst iets vertellen over Dust Bowl empiricism en rainforest empiricism…

Dust Bowl empiricism zeg je?

De term "Dust Bowl" verwijst naar de prairievlakten in de Verenigde Staten en Canada waar grote stofstormen woedden in de jaren 30. In het centrum van de Verenigde Staten was het tussen 1925 en 1950 enorm populair om empirische observaties te doen en

data te verzamelen zonder aan theorievorming te doen. Notoire voorbeelden van 'wetenschappers' die zo te werk gingen zijn Elton Mayo en Frederick Taylor- de zogenaamde Hawthorne studies bevatten heel wat fouten door het gebrek aan duidelijke voorafgaande theorievorming en de oorzaken voor minder of meer produceren werden aan foute zaken toegeschreven. Immers, wanneer je veel meetdata verzamelt zoals hij deed, kan je wel correlatietechnieken en factoranalyses of aanverwante procedures hanteren, meer dan enkele verbanden veronderstellen kan je nog altijd niet. Deze zaken kunnen in een exploratieve fase wel helpen om een theorie, model en hypotheses te ontwikkelen, die je dan vervolgens gaat onderzoeken met andere datasets, nooit dezelfde (anders draai je in een cirkel). Het onderzoeken van data zonder een theoretisch verklaringskader wordt daarom "Dust Bowl empiricism" genoemd: je hebt een enorme massa data – zoals de stofwolk – maar daar echt relevante verbanden in te vinden is bijzonder moeilijk en riskant.

Verschillende domeinen in de menswetenschappen worden fel bekritiseerd omdat ze zich voornamelijk baseren op observaties en het verzamelen van data in plaats van het uitwerken van een goede theorie die ze vervolgens testen door hypotheses te ontwikkelen en te toetsen. Zulke hypotheses zijn het voorwerp van experimenten die ertoe leiden dat de hypotheses ofwel (voorlopig) worden bevestigd (confirmation) of worden verworpen (falsification).

Heel wat wetenschappers en wetenschapsfilosofen beschouwen het louter verzamelen van data en daar dan vervolgens trachten correlaties in te vinden, om daaruit uiteindelijk "voorschriften" uit te distilleren als een enorm probleem en slechte wetenschap. Eerst en vooral vormt dit het zogenaamde "is/ought" probleem (Hume): het is niet omdat iets op een bepaald ogenblik kan teruggevonden worden in de natuurlijke wereld dat het daarom ook geldt als een voorschrift voor later, geldt in andere omstandigheden of wat dan ook. Deze werkwijze vormt ook het bijzondere probleem dat wetenschapsfilosofen onder-determinatie (under determination) noemen.

Ten tweede is het niet omdat je op een bepaald tijdstip correlaties vindt, dat deze correlaties zullen teruggevonden worden in een volgend onderzoek, betekenisvol zijn of iets zeggen over oorzaak en gevolg. **Integendeel, hoe meer data, hoe waarschijnlijker het is dat je willekeurige correlaties vindt en dat je andere correlaties mist.** Dit laatste is onder-determinatie of het idee dat de 'gevonden bewijzen' (evidence) niet volstaan om ons te helpen begrijpen wat deze nu eigenlijk betekenen. Het zijn dus niet de feiten of data die problematisch zijn, het is de relatie tussen de feiten die we menen te zien die problematisch is. Die neiging tot het trachten te ontwaren van correlaties is een goed bestudeerd facet van onze menselijke psychologie: wij zijn correlatie- en verklaring zoekende wezens.

Een typisch voorbeeld van Dust Bowl empiricism kan de bovenstaande definitie helpen duiden: het valt iemand (X) op dat op warme (zomer)dagen vrouwen meer korte rokken dragen maar ook dat er meer mensen ijsjes eten. X besluit dat '*het dragen van korte rokjes ijskreemconsumptie verhoogt'.* Dit is een eenvoudig voorbeeld van Dust Bowl empiricism gebaseerd op twee verschillende waargenomen correlaties:

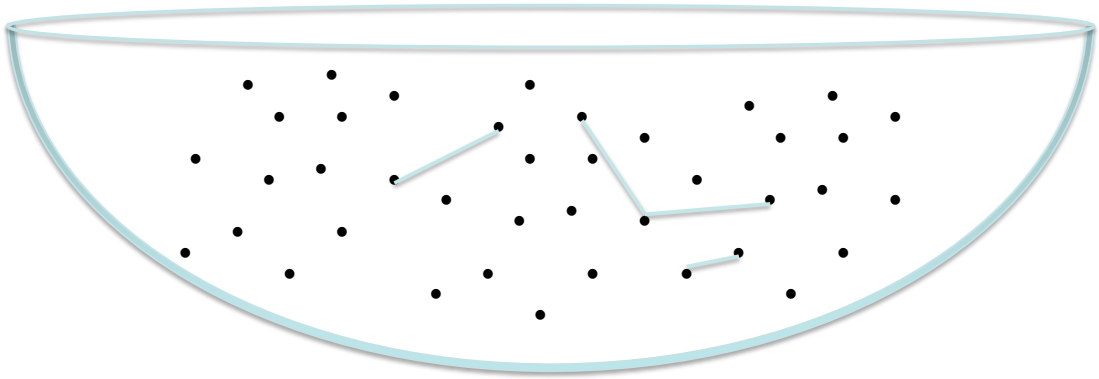
1. als het buiten warm is, dragen meer vrouwen korte rokken
2. als het buiten warm is, eten mensen meer ijskreem

De gemeenschappelijke variabele is echter de buitentemperatuur, maar dat betekent nog niet dat korte rokken en ijskreem gecorreleerd zijn.

Een wetenschappelijk methodologisch juiste manier zou er in bestaan eerst een hypothese te ontwikkelen en te formuleren, zoals '*het dragen van korte rokken zal het eten van ijskreem verhogen'* (of omgekeerd). Vervolgens dient een goed experiment te worden ontworpen dat de hypothese kan bevestigen of verwerpen. Indien bevestigd, mag men op basis van één methodologisch correcte studie nog steeds geen conclusies trekken. Een andere wetenschapper of groep wetenschappers zou dan het experiment moeten overdoen om te kijken of de resultaten kunnen 'gerepliceerd' worden: krijgt men met dezelfde methode hetzelfde resultaat? Indien meerdere pogingen tot replicatie mislukken, moeten ofwel de theorie en de er bij horende hypotheses worden aangepast of de theorie worden verlaten. Deze wetenschappelijke methode wordt ook wel de hypothetisch-deductieve methode genoemd en biedt de beste waarborg dat we onszelf niet bedriegen en verstandig omgaan met onze menselijke neiging tot vooroordelen en confirmatieneiging.

Het probleem met Dust Bowl empiricism is dat je de data 'zo hard kan martelen', dat je bijna alles kan 'bewijzen' of beweren… Dat is dus een groot gevaar voor grote databases waarin zowel gestructureerde als ongestructureerde data zitten.

Samengevat: correlaties kunnen eenvoudigweg willekeurig zijn en/of kunnen onder-determinatie verbergen. Big Data leidt tot apofenie – de aandoening waarbij mensen in willekeurige zaken patronen menen te zien (waar er geen zijn). Elke zuivere empirische benadering is per definitie ondergedetermineerd: er zijn waarschijnlijk feiten die je tot op heden nog niet vond. In de meeste gevallen 'verdwijnen' de bewijzen (in feite slechts correlaties) wanneer men probeert te repliceren of wanneer wordt vergeleken met bevindingen in andere wetenschappelijke domeinen.



Figuur: *Een grote database kan je vergelijken met een kom vol stof waarin verschillende korreltjes ("data") willekeurige correlaties en échte, betekenisvolle correlaties kunnen vertonen. Sommige korreltjes kunnen meervoudige correlaties vertonen die in alle richtingen kunnen uitstralen, maar dit betekent nog niet dat de korreltjes ook onderling gecorreleerd zijn.*

Rainforest empiricism? Wat is dat nu weer…

Een variant van Dust Bowl empiricism wordt 'rainforest empiricism' genoemd (Mael & Hirsch, 1993 – waarom ze deze term 'uitvonden' is ons niet bekend). Het enige verschil met Dust Bowl empiricism is dat je nadat je correlaties hebt gevonden (ook

willekeurige), je de wetenschappelijke literatuur screent om een theorie te vinden die bij de data zouden passen en hieruit vervolgens conclusies trekt. Dit wordt in de wetenschapsfilosofie tot een van de ergste vormen van zelfbedrog gerekend (waarbij je jezelf wijsmaakt dat je dit succes aan je eigen capaciteiten te danken hebt). Je doet volop aan confirmatieneiging en hanteert een hoogst onwetenschappelijke benadering. Het is een typisch geval van *post hoc ergo propter hoc* (na dit, dus voor dit): je vindt een theorie die klopt met je gevonden correlaties en je probeert het verleden uit te leggen door het heden te gebruiken.

Het probleem met rainforest empiricism is dat je nu de onmogelijkheid hebt ingebouwd om de theorie te falsifiëren want je hebt je correlaties nu al gelinkt aan een theorie en beide bevestigen nu elkaar in een duivelse kringloop. Het is een *catch 22* geworden die door niemand nog kan worden bekritiseerd.

Eén medewerker als benchmark

De ingenieurs die de presentatie houden voor BDC claimen dat hun investeerders een verbazingwekkende 8 miljoen USD hebben opgehoest om het bedrijf op te richten. In hun verkooppraatje beweren ze dat ze daardoor een groep erg performante ingenieurs hebben aangeworven die nu een erg krachtige IT-tool hebben ontworpen om zowel gestructureerde data (bijvoorbeeld data in een cv) als 'ongestructureerde' data (bijvoorbeeld een videofilmpje) te verwerken. Ze beweren dat hun analyses van de Big Data het mogelijk maakt belangrijke beslissingen te nemen voor HR-aangelegenheden, zoals het aanwerven van talenten, talent management en payroll management.

BDC stelde dat organisaties heel wat ruwe data hebben waarvan de meeste ongestructureerd zijn en dat dit een echte goudmijn is. Het klinkt verlokkelijk. Big Data kunnen uit word-documenten, blogs en video worden gehaald. Dit wordt dan door hen "vertaald" in "begrijpbare taal" – wat dat ook moge betekenen: Het hoeft niet te verwonderen dat deze "taal" enkel door hun ingenieurs wordt gesproken. Hoe ze het doen blijft in hun geheime zwarte doos. Ze kunnen naar eigen zeggen ook de 'houdbaarheidsdatum" van een medewerker voorspellen gebaseerd op hun analyses. BDC zegt dat dit ondermeer te danken is aan hun capaciteit om "contextuele intelligentie" aan de data toe te voegen. Contextuele intelligentie die wordt aangeleverd door het bedrijf. Als ze iets specifieker worden, blijkt het (ondermeer?) te gaan om de visie van het bedrijf en wat het bedrijf wil bereiken. Vervolgens proberen ze dit te "capteren" en dan in de Big Data op zoek te gaan naar competenties die passen bij die context en doelen ("look for competencies that match that context and goal"). Hoe ze dit dan weer doen? Opnieuw is dit hun zakengeheim. Het maakt het onderzoeken en testen van hun beloftes onmogelijk en wie hun diensten koopt moet dus een blind vertrouwen in hen hebben.

Het betoog werd nog buitenissiger. In een voorbeeld over een recruteringsopdracht vertelden ze dat ze het cv van de meest capabele medewerker (hoe en door wie dit bepaald is werd niet verteld) in hun software hebben ingevoerd. Hun intelligent algoritme "herkent" deze data en analyseert deze (het lijkt nog veelbelovender dan kunstmatige intelligentie). Het cv van deze ene medewerker is de basis van een recruteringsprofiel en een 'recept' om potentiële kandidaten te screenen.

Het doet wat denken aan de tijd- en bewegingsstudies uitgevoerd door ingenieurs in een poging om een standaardisatie in hoog repetitief werk aan te brengen. De praktijk hield geen rekening met het feit dat mensen van vlees en bloed zijn en geen robotten. Mensen hebben verschillende persoonlijkheden, een verschillend energiepeil, recuperatievermogen en interesses bijvoorbeeld. Door het werk robotachtig te organiseren, voelden mensen zich dodelijk verveeld (bore-out), ervoeren ze een totaal gebrek aan autonomie en job control en gedroegen ze zich na verloop van tijd vijandig tegenover het bedrijf en de leiding, met meer ziekteverzuim en een hoger verloop als direct meetbare gevolgen.

Het is ronduit gevaarlijk om het cv van slechts één medewerker te gebruiken als een soort benchmark voor nieuwe aanwervingen – het is simpelweg slechte wetenschap. Dit terwijl er flink wat onderzoek voorhanden is dat aangeeft wat de beste voorspellers voor toekomstige werkprestaties zijn (onderzoek op basis van vele duizenden werknemers). Over de nood aan voldoende grote aantallen voor bepaalde wetenschappelijke conclusies is er al lang consensus binnen de wetenschappelijke wereld: 'single' studies die maar een paar honderd mensen in hun studie opnemen, hebben een grote kans op "vals positieven". Vervolgstudies (replicatiestudies), systematische reviews zoals meta-analyses bieden grotere statistische zekerheden.

Kritisch denken ten aanzien van BDC

Er werden wat vragen gesteld aan de ingenieurs van BDC die de presentatie hielden. Het werd al snel duidelijk dat ze geen idee hadden van problemen zoals Dust Bowl empiricism of willekeurige correlaties, maar erger nog, het bleek al snel dat deze 'chief programmers' geen idee hadden welke goede voorspellers door psychologisch onderzoek binnen de bedrijfspsychologie werden geïdentificeerd. Gevraagd naar de predictieve validiteit (voorspelkracht) van hun modellen, algoritmes of software, gaven ze toe dat ze "nog niet over dergelijke data" beschikten maar dat deze er "weldra zouden komen". Hadden ze net niet beweerd dat ze nu al bedrijven konden helpen in het nemen van betere beslissingen? Lichtgelovige kopers kunnen niet anders dan betreurenswaardige beslissingen nemen als je het ons vraagt…

Geconfronteerd met de uitdagende stelling dat wat ze deden fel leek op Dust Bowl empiricism luidde hun verdediging dat het aan de klant was om de juiste contextinformatie te leveren (eerder door hen 'contextuele intelligentie' genoemd) en dat zij alleen maar de vertaler van die informatie zijn. Met andere woorden; ultiem is het de klant die de verantwoordelijkheid draagt als de voorspellingen slecht zijn.

We drongen aan en vroegen hoe ze op dit ogenblik dan betere recruteringsresultaten konden voorspellen. Ze herhaalden dat ze nog geen predicatieve validiteitsstudies hadden gedaan maar dat het ondertussen kon helpen om een selectie te maken tussen veel curriculum vitae. Onverstoord vervolgden ze met een voorbeeld van een bedrijf dat - let op de ronde getallen - 200 nieuwe mensen moesten aanwerven en meer dan 25.000 cv's te verwerken kreeg naar aanleiding van een advertentie in een krant. BDC voerde de cv's in hun softwareprogramma en maakte binnen de twee weken een selectie ("shift") tot slechts 400 cv's, wat veel meer behapbaar was voor de recruteerders. Maar wacht eens even – indien BDC niet weet hoe predictief hun criteria om te 'ziften' zijn, hoe kan je dan beweren dat je accuraat hebt gezift? Wij zijn er redelijk sterk van overtuigd dat je de

kwaliteit van hun werk niet kan nagaan, deels omdat hun software een echte zwarte doos lijkt die ze zorgvuldig verbergen.

Ze zaten niet om een straffe uitspraak verlegen: BDC beweerde dat dankzij hun software bedrijven uit de VS minder legale problemen hebben omdat je kan bewijzen dat je niet discrimineerde op basis van geslacht, minderheidsgroepen enzovoort. Naar hun mening is er geen "bias" op basis van huidskleur, ras, religie enzovoort.

Wat voor ons de deur helemaal dichtdeed en hun geloofwaardigheid tot nul herleidde, was dat ze zeiden dat hun systemen extreme 'modulatie' toelieten en dat de klant om het even welke data kan inbrengen zoals "data van MBTI of PAPI-testen". Dit is een enorm probleem. Laat ons kort het voorbeeld van de MBTI nemen – een test die al meermaals werd bekritiseerd door ons: (1) de theorie van Jung over archetypes is complete nonsens (het gaat om archetypes die zich in een parallel universum bevinden die via een soort van paranormaal proces zich in ons collectieve geheugen nestelen en waartoe we onbewust toegang hebben – gemengd met een aantal psychoanalytische concepten); (2) er zijn vele problemen met de test (lage test-hertest betrouwbaarheid, ipsatief karakter enz.), maar als klap op de vuurpijl (3) waarschuwt de uitgever van MBTI zelf dat de test nooit mag gebruikt worden in een recruteringscontext wegens onvoldoende betrouwbaar. We kunnen dan ook simpelweg concluderen: als je rommel in je algoritme stopt, kan je er alleen maar rommel uitkrijgen (*garbage in, garbage out*). En dan hadden we het nog niet over de vele gevaren van het overtreden van de wetgeving op de bescherming van de persoonlijke levenssfeer waar zoveel auteurs voor waarschuwen.

Ik schreef mijn eigen algoritme

Zoals reeds eerder gesteld, viel het op hoe weinig deze bedrijven afwisten over wetenschappelijke bevindingen in het domein van de arbeidspsychologie. Daarom dat bedrijven als BDC zonder schaamte durven stellen dat ze engagement bij medewerkers meten aan de hand van volgende eigen criteria:

- Heeft de medewerker al eens voor een andere interne job gesolliciteerd?
- Zijn ze lang genoeg in dezelfde job gebleven?
- Hebben ze al aan opleidingen deelgenomen.

Deze drie criteria worden gebruikt in een scoringsmechanisme – hun algoritme dat ze hiervoor schreven is uiteraard… geheim.

Toevallig werkte een Belg als marketing directeur voor het andere bedrijf (BDC2) en vonden we hem bereid om de softwaremodules te demonstreren. Ook hier stelden we gaandeweg methodologische vragen zoals '*Waar verkreeg je je wetenschappelijke informatie?', 'Wie bepaalde het algoritme?'* en *'Welke wetenschappelijke basis had dit algoritme*?'. Het mag geen verrassing heten dat we ook hier vaststelden dat een aantal IT-ingenieurs het algoritme zelf hadden geschreven zonder dat ze volgens de marketing directeur ook maar één wetenschappelijk artikel hadden geconsulteerd. In hun retentiemodule demonstreerden ze nochtans de indicator "vertrekrisico" ("departure risk"): het risiconiveau werd uitgedrukt met een kleurcode (rood = hoog risico). Geconfronteerd met onze methodologische vragen (en steeds kritischer wordende blik wellicht) haastte deze marketing man om te zeggen dat het slechts om "een advies ging, geen voorspelling". Na stevig aandringen ("maar waarom?") gaf de man uiteindelijk toe

dat de term “advies” werd gekozen om zichzelf in te dekken tegen juridische vervolging. Het was duidelijk: ze gebruikten hun eigen gokcriteria om algoritmes te schrijven. Ook voor hun andere indicator (engagement) gaf de man van BDC2 toe dat het hier ook om giswerk ging – ook hier ontbrak elk besef dat er wetenschappelijke bevindingen zijn die hen zouden kunnen helpen om nauwkeuriger “engagement”te voorspellen.

Het echte probleem

Het moge duidelijk zijn: het gaat niet om de data zelf, noch hun volume die het probleem vormen. Het is de interpretatie ervan die tot problemen leidt. Interpretatie van feiten is nooit objectief. Iedere mens, ook de wetenschapper, moet data interpreteren en daar zit de kwetsbaarheid: het interpretatieproces is onderhevig aan subjectiviteit en oordeelvervorming (bias). Nogmaals, correlaties vinden in data is één zaak, ze zinvol interpreteren is een andere. Het verhaal van BDC en BDC2 laten zien hoe subjectief zij tewerk gingen en – veel erger - geen rekening hielden met de reeds bestaande wetenschappelijke bevindingen. Zij schenen ook niet te weten dat wetenschap een methode is die in het leven werd geroepen juist om te voorkomen dat we onszelf bedriegen en ten prooi vallen aan over-interpretatie, vooroordelen, confirmatieneiging enzovoort – problemen waar wetenschappers altijd beducht moeten voor zijn maar de sociale wetenschappers nog meer.

Ben Goldacre, de Britse dokter die een knuppel in het medische en farmaceutische hoenderhok wierp, stelde in zijn boek ‘Bad Science’ het volgende: “*This breaks a cardinal rule of any research involving statistics: you cannot find your hypothesis in your results. Before you go to your data with your statistical tool, you have to have a specific hypothesis to test. If your hypothesis comes from analysing the data, then there is no sense in analysing the same data again to confirm it.”* (p. 275). Deze vorm van ‘bad science’ is net wat BDC en BDC2 beoefenden, en wellicht nog heel wat meer bedrijven. Dit zijn erg zorgwekkende vaststellingen en dit zou ons moeten alert en kritisch maken tegenover andere Big Data of HR Analytics aanbieders. Maar al te vaak denken bijvoorbeeld journalisten in de HR-media dat als er statistiek mee gemoeid is, het dan ‘evidence based’ is. Niet noodzakelijk, wel integendeel zelfs in dit geval.

Beter dan wetenschappers!

De meest hilarische bewering (nou ja, in onze ogen) kwam van BDC: “*We have determined a measure of success based on 16,000 employees: this measure allows to measure the financial contribution to the top line of the business. This is something researchers were not able to do so far, they can make direct links, but not provide the kind of intelligence we can provide*”.

We durven te hopen dat we een beetje hebben bijgedragen tot meer alertheid en een kritische houding tegenover de beweringen van genialiteit en overtrokken beloftes waarop de Big Data bedrijven een patent lijken op te hebben.

Patrick Vermeren is de voorzitter van de vzw Evidence Based HRM
Tom Speelman is filosoof en HR-consulent.

Verdere bronnen:
Boyd, D., & Crawford, K. (2012) Critical Questions for Big Data. *Information, Communication & Society,* 15(5)

Wenst u meer kritiek te lezen? Gebruik een zoekmachine op internet en tik in: “Big data critique”