UNIVERSITY OF APPLIED SCIENCES

DATASCIENCE LECTORALE REDE VOOR DE INAUGURATIE ALS HZ-LECTOR DATA SCIENCE MISCHA BECKERS

UNIVERSITY OF APPLIED SCIENCES

DATA SCIENCE

IN EEN DUURZAME DYNAMISCHE DELTA

DO BELIEVE THE HYPE!

LECTORALE REDE VOOR DE INAUGURATIE ALS HZ-LECTOR DATA SCIENCE

MISCHA BECKERS

Vlissingen, oktober 2017

DATA SCIENCE

IN EEN DUURZAME DYNAMISCHE DELTA

DO BELIEVE THE HYPE!

LECTORALE REDE VOOR
DE INAUGURATIE ALS
HZ-LECTOR DATA SCIENCE
MISCHA BECKERS

Vlissingen, oktober 2017

HZ UNIVERSITY
OF APPLIED SCIENCES

INHOUD

Voorwoord **1**
Begrippenlijst **2**
Don't believe the hype? **3**
Nu starten met data **4**
Nut en noodzaak 5
De bomen en het bos 5
Data Science **7**
Proces 7
De science in Data Science 8
Data Science in de praktijk 8
Toepassing **12**
Duurzame Dynamische Delta 13
Predictive Analytics in de context 13
Naar Predictive Maintenance 14
Valkuilen **15**
Goochelen met resultaten 15
Privacy 15
Ethiek 16
Toekomst **18**
Data Science Lab 18
Techniek 18
Sneller rekenen 18
Nieuwe algoritmes 18
Nieuwe visualisaties 19
Context 19
Do believe the hype! **21**
Bronnenlijst **22**
Dankwoord **24**
CV **24**
Bijlage 1 – Waarom aan de slag met Data Science? **25**
Bijlage 2 - Wat is veel data? **26**
Bijlage 3 – Data Science disciplines en vaardigheden **26**
Bijlage 4 – Data Science Lab Zuid **27**
Bijlage 5 – Wanneer werken Deep Neural Networks goed? **28**

VOORWOORD

In 1989 voerde ik mijn afstudeeronderzoek voor de Hogere Laboratoriumopleiding uit bij DOW Benelux. Een boeiend onderzoek naar de fysische eigenschappen van restproducten uit het olieraffinageproces. Het doel was om hiervoor toepassingen te vinden, in plaats van het als afval te verwerken. Daarvoor waren diverse chemische analyses nodig, zowel met klassiek handwerk als geavanceerde apparatuur. Maar mijn aandacht werd voortdurend getrokken door twee analisten. Gebiologeerd tuurden ze in hun kantoor naar een computerscherm, discussieerden op gedempte toon alvorens ze resultaten invoerden. Ik stapte binnen en vroeg waar zij mee bezig waren. Data analyse, zo bleek. Door data (anders) te ordenen, samen te vatten in kengetallen, wiskundige functies te zoeken die zo goed mogelijk bij een set datapunten pasten en inventief te visualiseren kwamen ze tot nieuwe inzichten. Steeds vaker ging ik buurten. Data analyse dus. Het boeide mij mateloos. Daar wilde ik verder mee. De laboratoriumjas ging aan de kant. Op zoek naar een vervolgstudie kwam ik uit bij de Radboud Universiteit, met de major Analytische Chemie, waarbij de focus volledig op data analyse ligt, *"which we call chemometrics [...], the discipline within chemistry that develops methods to obtain relevant information from chemical data"* (Radboud University, 2017)[1]. Data analyse mondde uit in Data Mining (Buydens, Reijmers, Beckers, & Wehrens, 1999) en in diverse wetenschappelijke artikelen en een proefschrift (Beckers, 1997). Ik had mijn weg gevonden. Data laten spreken werd een rode draad in mijn werk. Met als bekroning mijn eigen lectoraat Data Science. Hoe ik onderzoek doe naar, en met data, waarom, en wat ik daar mee wil bereiken zet ik uiteen in deze inaugurele rede.

[1] Hoewel er diverse (uitgebreide) definities van chemometrie bestaan. Svante Wold besteedt in (Wold, 1995) zelfs een heel wetenschappelijk artikel aan de betekenis van Chemometrie.

BEGRIPPENLIJST

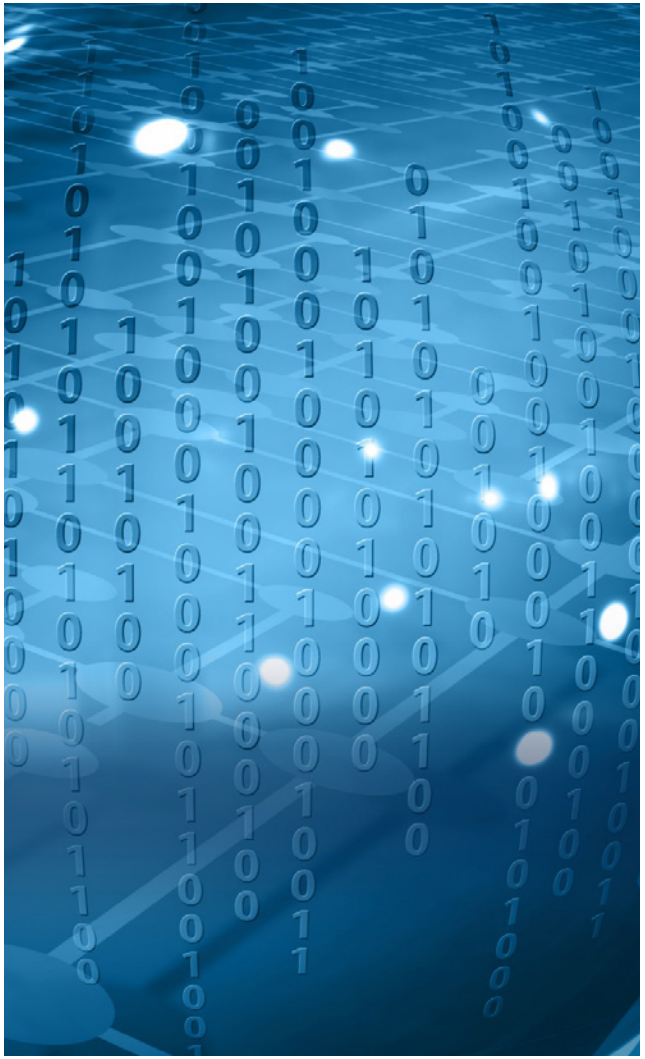
Application Programming Interface (API)	Een verzameling definities op basis waarvan een computerprogramma kan communiceren met een ander programma, of – onderdeel.
Big Data	Datasets die te groot in omvang zijn om met algemeen gebruikelijke software tools te verzamelen, verwerken en beheren, en dat proces binnen een acceptabele tijd uit te kunnen voeren.
Chemometrie	Discipline binnen de chemie die methoden ontwikkelt om relevante informatie uit chemische data te verkrijgen.
Computer Science	De studie van theorieën, experimenten en ontwikkelingen, die de basis vormen voor het ontwerpen en gebruiken van computer.
Computer Vision	Met behulp van een computer beelden interpreteren die met een camera zijn vastgelegd.
Cross-Industry Standard Process for Data Mining (CRISP-DM)	De meest gebruikte fasering voor Data Mining.
Data Mining	Het proces om betekenisvolle correlaties, patronen en trends te ontdekken middels het doorzoeken van grote datasets en daarbij gebruik te maken van patroonherkenning, statistiek en wiskunde.
Data Engineering, of Wrangling, of Munging	Een combinatie van data verzamelen, opschonen, transformeren en geschikt maken voor volgende activiteiten (zoals Machine Learning). Onderdeel van Data Science.
Data Science	Het proces waarin data omgezet wordt in waardevolle inzichten, beslissingen en producten en waarbij meerdere disciplines waaronder statistiek en wiskunde, Machine Learning, algoritmen en visualisatie, communicatie en inzicht in ethische, juridische en economische problematiek, samenkomen.
Deep Learning	Nieuwe(re) generatie Machine Learning die meer richting Kunstmatige Intelligentie gaat.
Kenniskring	De lector(en) werken samen met een kenniskring bestaande uit docenten en externe experts.
Kunstmatige Intelligentie	De intelligentie waarmee machines, software en apparaten zelfstandig problemen oplossen. Zij imiteren hierbij het denkvermogen van een mens.
Lector(aat)	Een lector is een expert in een bepaald vakgebied. Lectoren zetten hun kennis en ervaring in voor onderzoek en innovatie binnen het onderwijs en de beroepspraktijk. In wisselwerking met de praktijk ontwikkelt het lectoraat kennis die praktisch toepasbaar is.
Machine Learning	Onderdeel van Data Science. Algoritmes die leren om taken uit te voeren (zoals voorspellen of beslissingen nemen) op basis van observaties (data).
Natural Language Processing	Het toepassen van computertechnieken voor de analyse en synthese van natuurlijke taal en spraak.
Predictive Analytics	Een invulling van Data Mining met de nadruk op voorspelling (in plaats van beschrijving, classificatie of clustering).
Predictive Maintenance	De toepassing van Predictive Analytics om te voorspellen wanneer onderhoudsactiviteiten nodig zullen zijn.
Process Mining	Technieken om uit specifieke data die vastligt in ICT-systemen (zogenaamde event logs of workflow logs) procesmodellen af te leiden.
Sensor	Een apparaat dat iets kan waarnemen.
Smart	Term die gebruikt wordt om aan te duiden dat in een bepaalde context (zoals een stad of een haven) iedereen, en alles (digitaal) verbonden is.
Supervised Learning	Subset van Machine Learning waarbij bekend is waar men naar zoekt. Een van de variabelen in de train dataset is de target, of afhankelijke variabele. Daar is een label aan gegeven. De overige zijn de onafhankelijke variabelen.
Unsupervised Learning	Subset van Machine Learning waarbij nog niet bekend is waar men naar zoekt. Geen van de variabelen in de train dataset is de target, of afhankelijke variabele.

DON'T BELIEVE THE HYPE?

In de film Spectre (2015) ontvangt James Bond smart blood. Daardoor kan de geheime dienst MI6, Bond overal ter wereld volgen. Hoe realistisch is dat? (Wagstaff, 2015). Feit is dat men de term smart steeds vaker gebruikt om aan te duiden dat iedereen, en alles (digitaal) verbonden is. En dat daarbij beslissingen worden genomen op basis van data. Zo ontstaan smart homes, in smart cities, met smart communities, die gebruik maken van smart traffic et cetera. Vanzelfsprekend genereren de hiervoor verbonden apparaten, real-time enorme hoeveelheden data. We spreken zelfs van datafication, ofwel *"taking all aspects of life and turning them into data"* (Cukier & Mayer-Schoenberger, 2013). En we belanden door deze digitale transformatie in een nieuwe wereld: de nieuwe informatiesamenleving (Klous & Wielaard, 2014).

Voor die groeiende hoeveelheid data gebruikt men vaak de term Big Data. Maar wat Big Data precies is, laat zich niet eenvoudig beschrijven. *"Datasets die te groot in omvang zijn om met algemeen gebruikelijke software tools te verzamelen, verwerken en beheren, en dat proces binnen een acceptabele tijd uit te kunnen voeren"*. Die definitie geven (Snijders, Matzat, & Reips, 2013), als startpunt. Over kansen en risico's van Big Data bestaat veel discussie. De tweet van Europees Commissaris Neelie Kroes, toen verantwoordelijk voor de portefeuille Digitale Agenda sprak boekdelen: *"Big Data is at the heart of solving most of our unsolved problems – from cancer to climate change. We must build trust in it"* (Kroes, 2014). Een grote claim, ook wel de Big Data belofte genoemd. De bedoeling van dataverwerking is natuurlijk om waarde te creëren (door data op te werken naar waardevolle informatie). Die waardecreatie beperkt zich niet tot commerciële doelen. Het oplossen van maatschappelijke problemen op het gebied van mobiliteit, veiligheid en duurzaamheid en het bedenken van toepassingen die het leven veraangenamen, zijn net zo relevant, zo niet relevanter.

Echter, *"Big Data is een hypeterm. Dat data[sets] groot zijn, is een leeg begrip. Door de exponentiële groei van data om ons heen raakt het steeds meer ons dagelijks leven. We zoeken een term om dat een plaats te geven. Maar, daarachter zit een gevestigd vakgebied met tradities. Statistiek is al honderden jaren oud, maar toch een van de basisingrediënten van Big Data"*. Aldus prof. dr. ir. Wil van der Aalst in (Lonkhuyzen, 2017). Er zijn dus meerdere ingrediënten (nodig) om verantwoord met data om te gaan en er waarde mee te creëren. In deze rede, ontrafel en duid ik de (Big) Data hype en laat zien dat het vakgebied Data Science hier een sleutelrol in heeft. Eerst introduceer ik de benodigde concepten en wat daarvan nut en noodzaak zijn. Daaruit wordt afgeleid wat Data Science is. Vervolgens schets ik hoe Data Science binnen de context van een hogeschool praktisch kan worden toegepast. Tenslotte droom ik van een toekomst met de inbedding van Data Science niet alleen in de regionale context, maar ook (inter)nationaal. Dan zult u zien dat we hier niet te maken hebben met de negatieve connotaties van het begrip hype, zoals *"een door veel mensen gedeelde belangstelling of bezigheid die na korte tijd voorbij is"* of een *"modegril, die bijna altijd overwaait"*. Maar, dat het gaat om een *"zichzelf versterkend mechanisme [...] dat uitgroeit tot een werkelijk belangwekkend verschijnsel"*[2]. Do believe the hype!

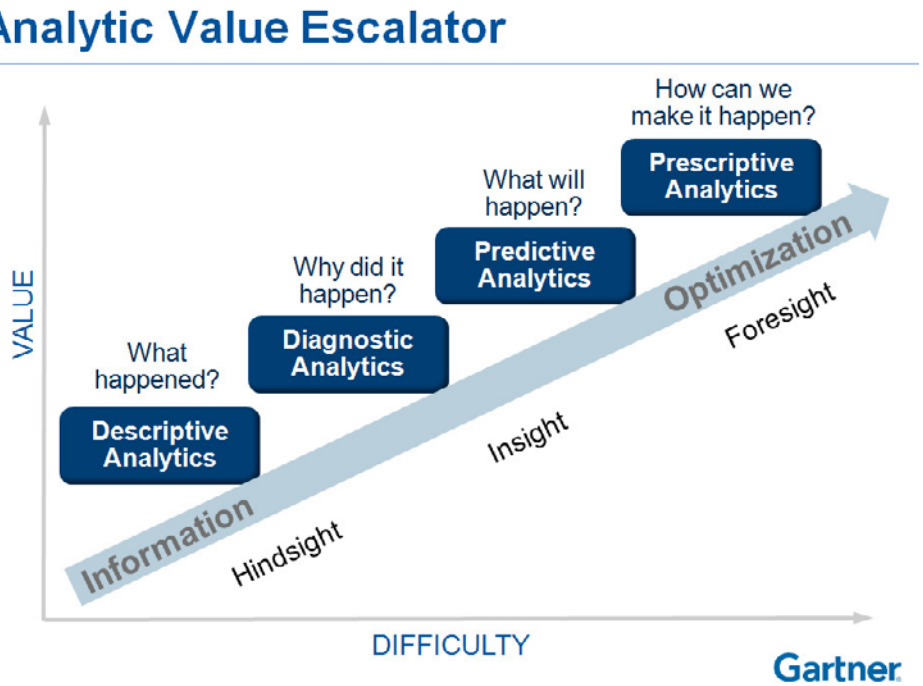


[2] Hier heb ik stukken geciteerd uit diverse definities die te vinden zijn voor het begrip hype.

NU STARTEN MET DATA

Om te achterhalen of en hoe waarde uit Big Data gegenereerd kan worden, is het nodig een aantal vragen te beantwoorden: wat is de context van de data, wie zijn die belanghebbenden, wat levert voor hen waardecreatie, wat is de daar bijhorende informatiebehoefte en is daar al data voor beschikbaar, of moet die data nog worden gegenereerd of verzameld?

Waardecreatie is onlosmakelijk verbonden met analyse. Immers, door gebruik te maken van data wil men zinvolle beslissingen kunnen nemen. Die staan aan de basis van waarde creëren. Een bruikbare typering van mogelijke analyses staat in Figuur 1. Deze typering is gebaseerd op wat Gartner meldt in haar IT Glossary. Zie bijvoorbeeld (Gartner, 2017) voor een beschrijving van Predictive Analytics.



Figuur 1 Analytic Value Escalator, een typering van mogelijke analyses (Analytics).

AI en zelfrijdende auto's

In september 2017 publiceerde NRC een artikel met titel "De grote datarace: een rijbewijs voor de robotauto". Het behandelt enkele typische AI vraagstukken, in dit geval bij zelfrijdende auto's. Een van de vraagstukken is: wanneer ingrijpen? Dat kan alleen onderzocht worden indien inzicht bekend is. Wanneer is inzicht bekend? Wat te doen als een situatie niet herkend wordt? Het artikel vermeldt dat veel onderzoek nodig is om de computer te laten begrijpen wat er in het verkeer gebeurt. Maar wat moet de zelfrijdende auto nog leren om zijn rijbewijs te halen? De robotauto herkent nu al het soort weg, andere auto's, fietsers en voetgangers, stoplichten en verkeersborden. De auto ziet echter geen verschil tussen een opwaaiende krant of een rotsblok, of een overstekende hond. Zodra een obstakel niet herkend wordt zal de auto uit voorzorg stoppen, tot ergernis van het andere verkeer. De oplossing is met nog meer data trainen en nog meer objecten leren herkennen. Echter, objecten herkennen is één ding, voorspellen wat ze gaan doen is een stuk complexer. De auto moet inschatten wat andere verkeersdeelnemers van plan zijn om een veilige route te kunnen kiezen zonder gevaar te veroorzaken.

De computer moet bijvoorbeeld leren lichaamshoudingen van andere verkeersdeelnemers te interpreteren en te vertalen in gedragsmodellen. Een complicerende factor is dat verkeersdeelnemers zich niet altijd aan de regels houden. En dan is er nog het probleem van de 'semantiek'. Bij het nemen van de juiste beslissing heeft de computer er baat bij om relevante informatie te hebben over een gegeven situatie. Bijvoorbeeld bij het naderen van een kruispunt. Wat zijn typische gedrag – en bewegingspatronen op die kruising, komen hier voornamelijk voetgangers of fietsers etc.? (Hijink, 2017)

Nut en noodzaak

Figuur 1 is overgenomen uit het rapport van (Seeters, 2017), waarin een kort verslag staat van een bezoek aan de Gartner Data & Analytics summit 2017. Daar maakte men bekend dat van de 2000 bevraagde bedrijven 74% Descriptive Analytics bedrijft, 34% Diagnostic Analytics, 11% Predictive Analytics en slechts 1% Prescriptive Analytics. Vanzelfsprekend is het kunnen voorspellen wat er gaat gebeuren, of hoe iets te beïnvloeden is, zeer waardevol. Hieruit is meteen evident dat het noodzakelijk is om in dat gebied de mogelijkheden te gaan ontdekken. Ik gebruik dit als opstap om nut en noodzaak van Data Science, waarover later meer, te benadrukken. Als ondersteuning geldt een grote hoeveelheid publicaties, waarvan een selectie is opgenomen in Bijlage 1.

Machine Learning voor image processing

Een Machine Learning (in dit geval een Neuraal Netwerk) algoritme is getraind met (portret)foto's waarvan het geslacht en de leeftijdscategorie van de persoon bekend waren. Vervolgens is dat getrainde algoritme gebruikt om geslacht en leeftijdscategorie voor de persoon op een nieuwe portretfoto te voorspellen. De extra moeilijkheid is dat het algoritme op die foto eerst moet herkennen waar het gezicht is. In Figuur 2 is een voorbeeld te zien waarbij het algoritme automatisch de met de laptop gemaakte foto als input neemt, en de voorspelling retourneert. In dit geval met een correcte voorspelling van het geslacht (m) en de leeftijdscategorie (het lichte gebied in het plaatje, ofwel 21-40 jaar). Het algoritme is robuust. In die zin dat een foto van dezelfde persoon met of zonder bril en met of zonder lichaamsbeharing bijvoorbeeld, hetzelfde resultaat oplevert. Dit algoritme is getraind met relatief weinig voorbeelden. De range van de leeftijdscategorieën kan smaller gezet worden naarmate er met meer relevante voorbeelden is getraind.

De bomen en het bos

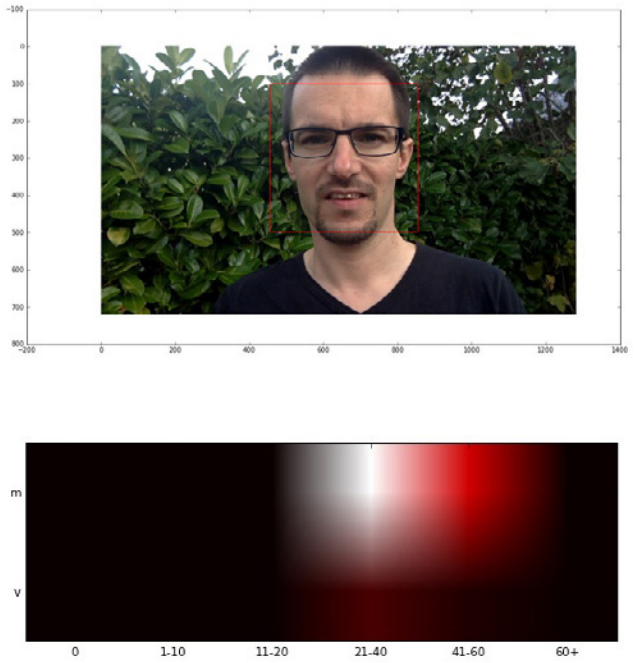
Bij aan de slag gaan met data, en daarin een weg vinden blijkt dat er een scala aan gebruikte concepten en terminologie bestaat. We duiken er wat dieper in om door de bomen het bos te kunnen blijven zien. Hoe zit dat? De genoemde IT Glossary van Gartner geeft voor Predictive Analytics als beschrijving: *"any approach to data mining with [...] an emphasis on prediction (rather than description, classification or clustering)"*[3]. Voorspellen kan alleen als er voorbeelden zijn. Hoe meer voorbeelden, hoe beter dat werkt.

De termen Machine Learning en Artificiële Intelligentie (AI) worden (hierbij) vaak door elkaar gebruikt. Het doel van AI is om computers te leren wat mensen op dit moment nog beter doen. Intelligentie betekent ook zelfstandig beslissingen nemen en vervolgacties inzetten. Dat vereist inzicht en het vermogen tot generaliseren en aanpassen. AI behelst dan ook een combinatie van elementen uit onder meer computer vision, spraakherkenning, taalbegrip, redeneren, plannen, navigeren en manipuleren. Machine Learning is essentieel bij elk van die elementen. Een aspect uit computer vision bijvoorbeeld is het herkennen van (delen) van beelden. Dat kan door een algoritme in te zetten dat een hele reeks gelabelde beelden krijgt voorgeschoteld en leert om de labels goed te voorspellen. Als dat goed gaat kan dit getrainde algoritme ingezet worden om nieuwe beelden, zonder label, te herkennen. Machine Learning is dus een subgebied van Artificiële Intelligentie (AI). Geen data, dan is er niks te leren. Big Data, dan is er veel te leren (zie Kader: AI en zelfrijdende auto's). Dat verklaart meteen de hernieuwde opkomst van Machine Learning (Domingos, 2015) (Shanahan, 2015). Zie ook Figuur 3. Het volgende hoofdstuk gaat wat dieper in op Machine Learning.

Als bovenstaande ook nog eens Big Data betreft, dan creëert dat een intrinsiek complex probleem. Immers, naast de al eerder gegeven Big Data definitie van Schnijders et al., laat Big Data zich ook beschrijven door een aantal V's. Veel gebruikte V's zijn:

- Volume: het betreft enorme hoeveelheden data (zie Bijlage 2 voor een indicatie van wat veel is;
- Variety: we zijn gewend om te werken met gestructureerde data, maar geschat is dat circa 90% van de huidige data ongestructureerd is;

[3] En data mining dan? De IT Glossary meldt daarover: *"The process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques"*. In het volgende hoofdstuk blijkt dat een groot deel van data mining onderdeel is van het Data Science proces.



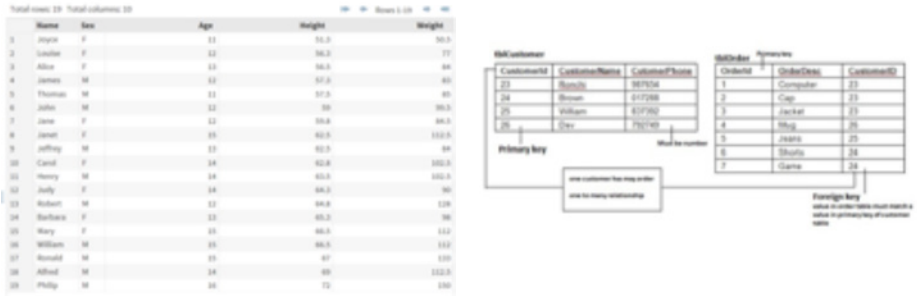
Figuur 2 Voorbeeld van beeldherkenning (prototype ontwikkeld door Gert Jacobusse).

- Velocity: steeds vaker wordt data (semi) real-time geproduceerd (denk aan sensoren) en dat is een uitdaging voor dataverwerking en – opslag;
- Veracity: als er zoveel data, zo snel en in diverse formaten wordt geproduceerd dan kan de betrouwbaarheid van die data in het geding komen.

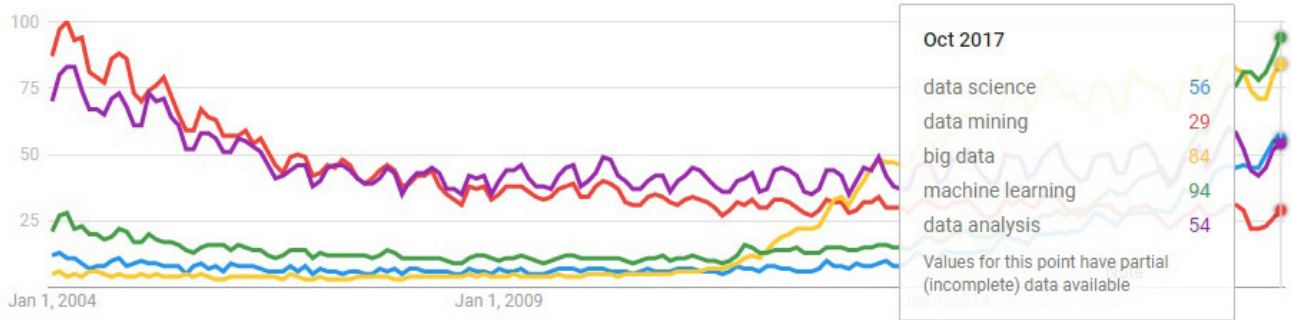
(Cao, 2017) betoogt daarom dat bij zulke intrinsiek complexe problemen een meer wetenschappelijke benadering van de omgang met data vereist is: Data Science.

Gestructureerde versus ongestructureerde data

We zijn gewend om te werken met gestructureerde data. Doorgaans betekent dit dat de data netjes geordend is in rijen en kolommen en/of middels relaties. Spreadsheets, comma separated value (csv) bestanden en tabellen in een relationele database zijn voorbeelden van gestructureerde data.



Qua vorm vrijwel direct geschikt om visualisaties van te maken, verkennende data analyse uit te voeren of complexere analyse op los te laten. Maar tegenwoordig wordt steeds meer ongestructureerde data geproduceerd. Wat te denken van foto's, filmpjes, documenten in allerlei bestandsformaten en volledige websites? In dat geval is een aantal bewerkingsstappen nodig om er gestructureerde data van te maken. Zo bestaat een fotobestand uit meerdere, gestapelde matrices. Onder meer voor de kleuren. Per kleur bevat zo'n matrix de intensiteit per pixel. In dit geval bestaat de bewerking er uit om die gestapelde matrices om te zetten naar een enkele matrix, of tabel. Evenzo voor een document. Dat is eigenlijk een verzameling van woorden en eventueel beeldmateriaal. Die woorden zijn geordend in een zin, die weer in een paragraaf et cetera. Een eenvoudige manier om van zo'n document gestructureerde data te maken is de aanwezige unieke woorden rangschikken en tellen hoe vaak elk in het document voorkomt.



Figuur 3 Frequentie van gebruikte zoektermen voor het omgaan met data, zoals bepaald met Google Trends [4].

[4] Google meldt ter interpretatie van de grafieken: *"Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. Likewise a score of 0 means the term was less than 1% as popular as the peak".*

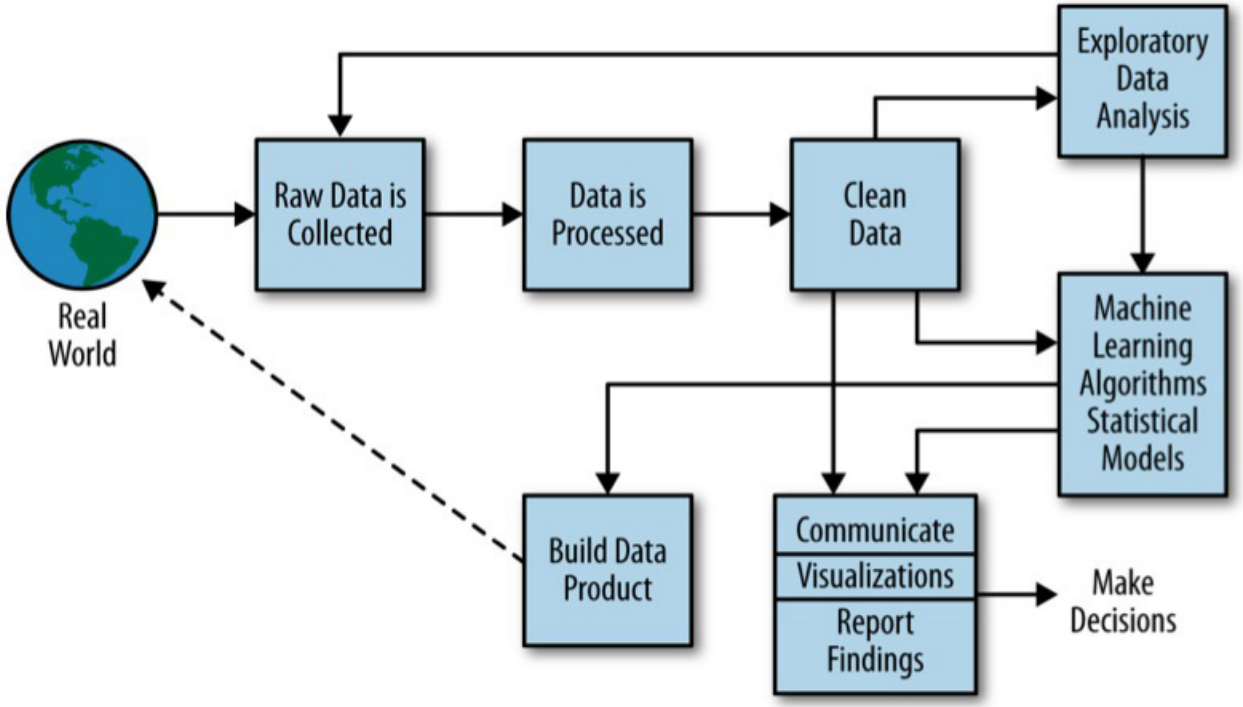
DATA SCIENCE

(Cao, 2017) geeft aan waarom en wanneer we Data Science toepassen, maar hij stelt zich ook een belangrijke vraag, namelijk: *"what makes Data Science, a science?"*.
Dit hoofdstuk gaat gedetailleerder in op Data Science als vakgebied. In het volgende hoofdstuk komt aan de orde hoe deze science een plek krijgt in het hbo-onderwijs en –onderzoek.

Proces

Procesmatig handelen in onze nieuwe informatiesamenleving, om te leren hoe we data in beslissingen kunnen omzetten vraagt om een Data Science-aanpak. (Provost & Fawcett, 2013) beschouwen het verbeteren van beslissingen nemen als het ultieme doel van Data Science (Data-Driven-Decision-making).
Een heel praktische manier om Data Science te benaderen biedt (Foreman, 2013): *"Data science is the transformation of data, using mathematics and statistics, into valuable insights, decisions and products"*[5]. (O'neil & Schutt, 2014) benadrukken dat hiervoor naast statistiek en wiskunde, machine learning en algoritmen en visualisatie, ook andere disciplines zoals communicatie en inzicht in ethische, juridische en economische problematiek vereist zijn. Bijlage 3 is een aantrekkelijke illustratie van wat Data Science aan disciplines verenigt, en aan vaardigheden vereist.

Goed te zien is hoe dit proces zich onderscheidt van een meer 'traditionele', geïsoleerde statistische aanpak.
Het Data Science proces levert een Data Product. Dat wordt opgenomen in de Real World, die de data aanleverde. Gebruikers interacteren met het Data Product en genereren nieuwe data, die weer belandt in de loop enzovoorts[6]. Het is dus niet puur de hoeveelheid data die interessant is (of voor uitdagingen zorgt). Het gaat om de data zelf die (vaak real-time) bouwstenen kan vormen voor dataproducten. De kunst is te leren hoe data omgezet kunnen worden in beslissingen[7].

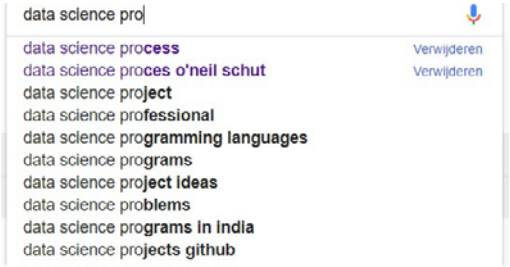


Figuur 4 Het Data Science proces, zoals beschreven in (O'neil & Schutt, 2014).

[5] In zijn boek laat hij zien hoe zien hoe dit te bereiken is met de inzet van spreadsheets (zoals MS-EXCEL), zelfs voor de toepassing van Machine Learning algoritmes.
[6] Het Data Science proces kan gezien worden als een uitbreiding van CRISP-DM, het procesmodel dat geldt als de standaard voor Data Mining (Shearer, 2000). Dit model is uitgebreid beschreven en er zijn diverse templates voor beschikbaar.
[7] De disciplines statistiek en wiskunde, machine learning en algoritmen en visualisatie zijn relatief eenvoudig te herkennen in het proces van Figuur 4. Communicatie ook. De ethische, juridische en economische problematiek komen typisch aan de orde bij de 'business understanding' (en het daaruit volgende verzamelen en verwerken van misschien wel privacy of anderszins gevoelige data) en de implementatie van het uiteindelijke dataproduct (de relatie met de Real World).

Voorbeelden van dataproducten

Bij het intypen van een zoekterm in de Google zoekmachine geeft deze automatisch mogelijk volgende zoektermen. Hoe kan dat? De meest eenvoudige manier is om alle gestelde zoekvragen op te slaan. In het geval van Google zijn dat er natuurlijk enorm veel. Met een zogenaamde n-gram analyse kan voor een bepaald woord, de waarschijnlijkheid worden berekend dat een bepaald woord daar op volgt. Stel er zijn 3-grammen gemaakt en daarin komen de woorden 'data' en 'science' na elkaar voor, dan is er een waarschijnlijkheid bekend voor daarop volgende woorden.
Bij het intypen van 'data science' in de zoekmachine geeft deze het woord terug met de hoogste waarschijnlijkheid, enzovoorts. Een bruikbaar dataproduct (zie Figuur 5). Het is ook mogelijk om hiervoor hele documenten te gebruiken (blogs, tweets, nieuwsberichten, wetenschappelijke artikelen et cetera). Als er dan ook nog relaties en logica worden toegevoegd is het mogelijk chatbots te ontwikkelen. Een chatbot antwoordt op gestelde vragen en is in staat een (simpele) conversatie te voeren. Mitsuku en Facebook M zijn hiervan voorbeelden. Chatbots vinden momenteel hun weg in de klantenservice. Nog een voorbeeld van een dataproduct is een spamfilter. Daarachter zit een algoritme dat geleerd heeft om spamberichten te onderscheiden van gewenste berichten.



Figuur 5 De Google zoekmachine denkt met ons mee.

De science in Data Science

Hoe waarde is te creëren vergt een onderzoeksmatige aanpak. Laten we dat concreet maken: onderzoek doen is een vraag stellen, data verzamelen die nodig is om de vraag te beantwoorden, methoden selecteren om daarvoor analyses uit te voeren, een antwoord formuleren en evalueren in hoeverre dat correct is. Dit is een cyclisch proces. Het Data Science proces geeft hier perfect invulling aan. Duidelijk te zien is ook dat dit proces het eerder aangehaalde Predictive Analytics uitstekend faciliteert. Uitdagingen zitten op diverse plaatsen. Bijvoorbeeld op het gebied van vraagarticulatie. In het Data Science proces van Figuur 4 speelt die zich af vóór de Raw 'Data is collected' activiteit (in CRISP-DM, het standaard procesmodel voor Data Mining (Shearer, 2000) wordt die fase Business Understanding genoemd). Waar is de stakeholder nu echt mee gebaat? Data verzamelen, bewerken en opschonen is doorgaans geen sinecure. Vaak wordt gemeld dat deze activiteiten zo'n 80% van het werk beslaan. En dan is er natuurlijk de kritische interpretatie van resultaten, op waarheid, nauwkeurigheid en praktische relevantie.

Data Science in de praktijk

Dit Data Science proces overkoepelt hetgeen tot nu toe gesteld is over Big Data en bevestigt wat daar óók over geschreven is: ja, het creëert een potentiële revolutie in meten (grof gesteld: het nemen van steekproeven is niet nodig, neem gewoon de hele populatie), ja, het is een invalshoek (filosofie) die belicht hoe beslissingen in de toekomst genomen zullen of moeten worden en ja, het brengt een bundel nieuwe technologieën met zich mee. Voor wat betreft het laatste is het goed een en ander in perspectief te plaatsen. Hadoop wordt veel genoemd als nieuwe technologie[8]. Waar dat inmiddels een heel ecosysteem van technologieën is, begon het als platform om grote hoeveelheden data teverzamelen (op het Hadoop File System, HDFS) en snel (namelijk parallel, op grote hoeveelhedencomputers) te verwerken (met onder meer MapReduce algoritmes). Geleidelijk aan kwamen daar technologieën bij zoals Hbase, Hive en Pig om data in te delen en te bevragen. Deze processen worden soms samengevat met de term Data Engineering. Tot dan toe bedekte dit ecosysteem met name de eerste activiteiten in het Data Science proces (tot aan Exploratory Data Analysis in Figuur 4). Inmiddels zijn er met Mahout ook Machine Learning libraries toegevoegd en breidt het systeem steeds verder uit. De huidige literatuur rondom Data Science legt veel nadruk op Exploratory Data Analysis en Machine Learning. Er zijn publicaties beschikbaar die honderden Data Science tools en pakketten beschrijven en categoriseren. Sommige daarvan zijn toepasbaar op het hele proces. Anderen focussen op een bepaalde activiteit, zoals visualiseren. Om niet te verdrinken in die diversiteit is het zinvol een (initiële) keuze aan tool en pakketten te maken. Dat is natuurlijk sterk afhankelijk van de thematiek en context van het onderzoek. Voor het lectoraat Data Science passen de meest genoemde generieke tools, namelijk Python en R, of hun meer op visueel programmeren gerichte tegenhangers Rapidminer of Orange, specifiekere pakketten zoals Talend

(Data Engineering) en Tableau (visualisatie). Steeds meer in opkomst zijn de zogenaamde cognitive services. Google, Amazon, Microsoft en anderen bieden allemaal on-line diensten om met name Machine Learning taken uit te voeren. Daar zitten natuurlijk wat haken en ogen aan. Die worden later besproken.

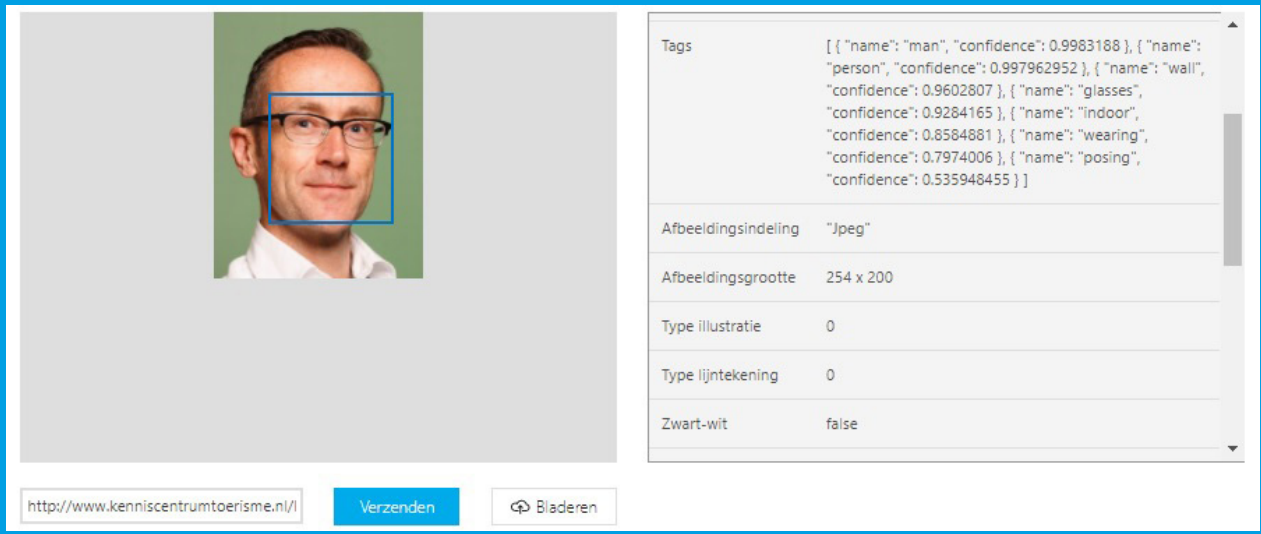
UNSUPERVISED LEARNING	SUPERVISED LEARNING
Niet helemaal bekend/onbekend waar men naar zoekt. Geen van de variabelen is target variabele.	Bekend waar men naar zoekt. Een van de variabelen is de target variabele. Daar is een label aan gegeven.
• Clusteranalyse • Netwerkanalyse • Market/basket analyse	• Classificatie, waarbij de target variabele opgedeeld is in klassen of categorieën (zoals ["echt", "vals"], [0-10, 11-20, 21-30]) • Regressie, waarbij de target variabele een continue waarde kan aannemen (bijvoorbeeld een reëel getal)

Cognitive services

Via https://azure.microsoft.com/nl-nl/services/cognitive-services/directory/ bereiken we de services van Microsoft. Met voor elk van de categorieën Vision, Spraak, Taal, Kennis en Search diverse services. Die zijn online meteen te gebruiken. Het is ook mogelijk de services in eigen software te nemen. Voor een online beschikbare foto van mijzelf, die aangeboden wordt aan de Computer Vision service is het resultaat verrassend. In de Beschrijving (die is niet getoond in Figuur 6, maar bevat nog meer aanwijzingen dan in Tags staat[9]) geeft de service niet alleen aan 'glasses', maar ook 'smiling'. En naast de getoonde Tags vindt de service ook

Categorieën [{ "name": "people_portrait", "score": 0.96875 }]
Gezichten [{ "age": 50, "gender": "Male", "faceRectangle": { "top": 76, "left": 51, "width": 121, "height": 121 } }]

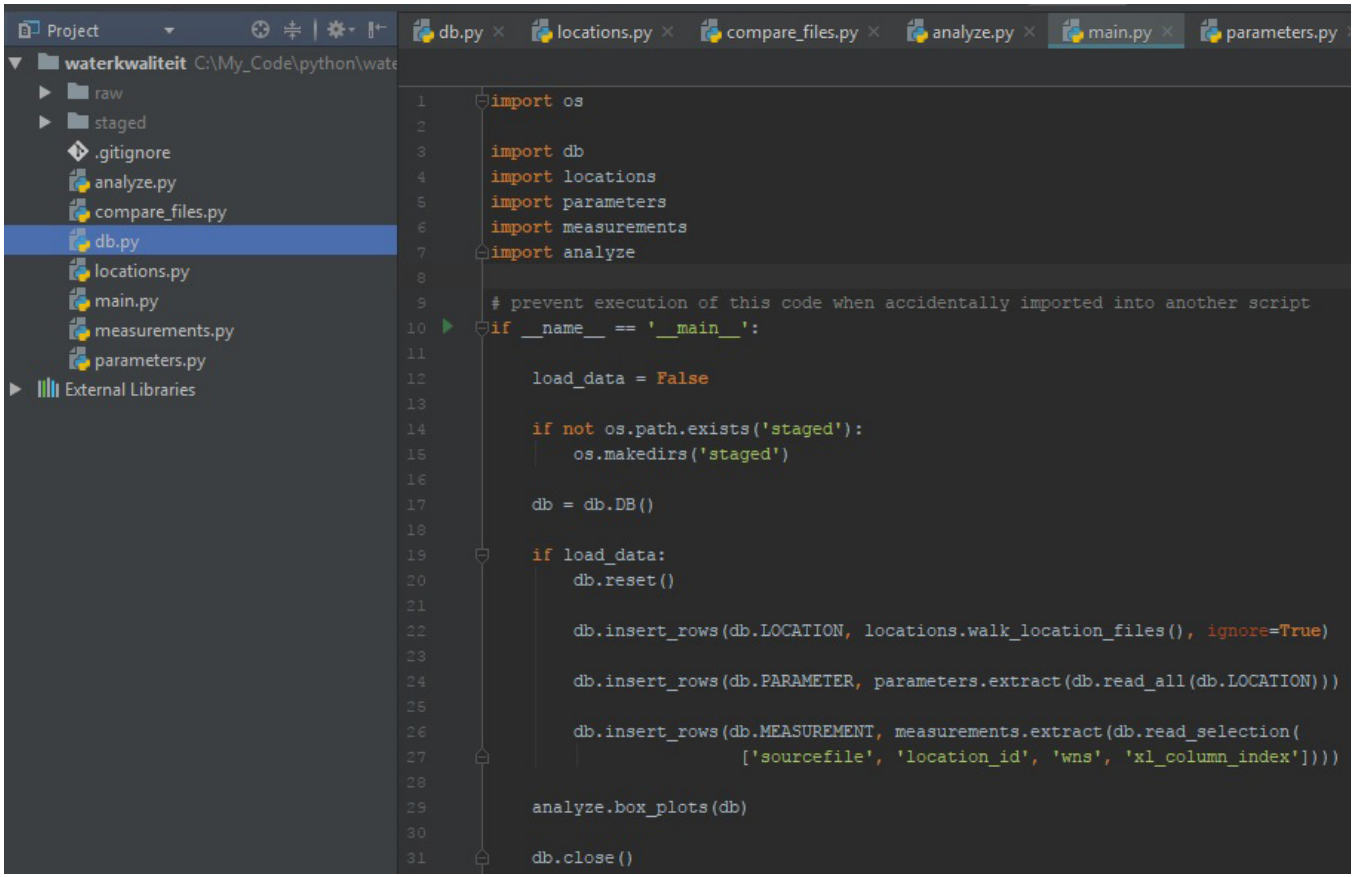
Inderdaad mijn leeftijd, op die specifieke foto.

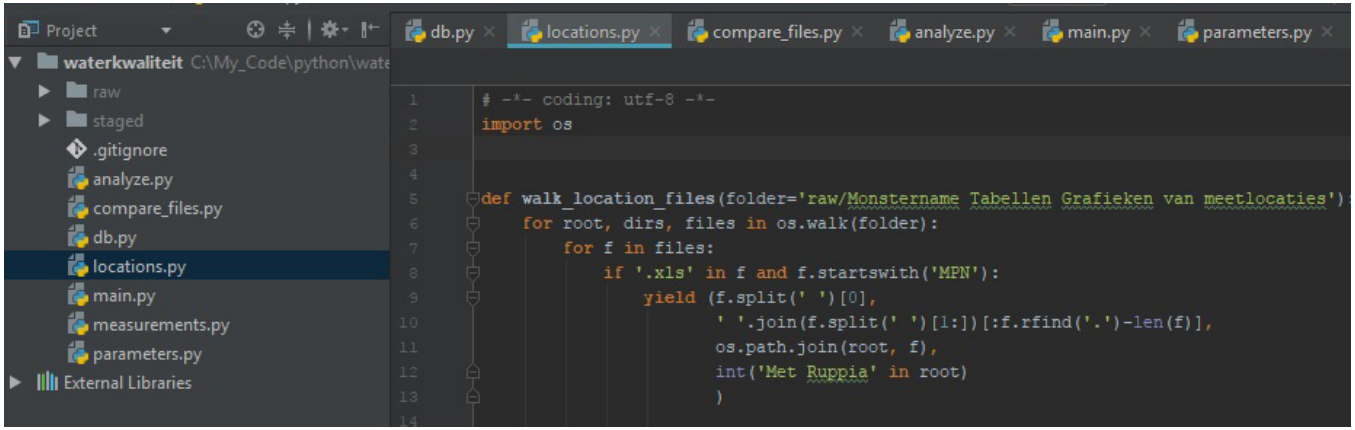


Figuur 6 Online foto aangeboden aan de Computer Vison API.

[8] (The Apache Software Foundation, 2017) toont een recent overzicht van het Hadoop ecosysteem.

[9] In die Beschrijving staat bijvoorbeeld ook 'Suit'. Nu heb ik geen kostuum aan op die foto. Sterker, de classificatie is alleen gemaakt op basis van mijn gezicht (zie het vierkant om mijn gezicht dat het algoritme eerst heeft gemaakt). Hoe komt dat algoritme dan aan de beschrijving 'Suit'? Waarschijnlijk zaten er in de training set foto's van mannen, wellicht van rond de 50, met een bril, die in portretmodus de camera in keken et cetera, die ook een label 'Suit' hadden.

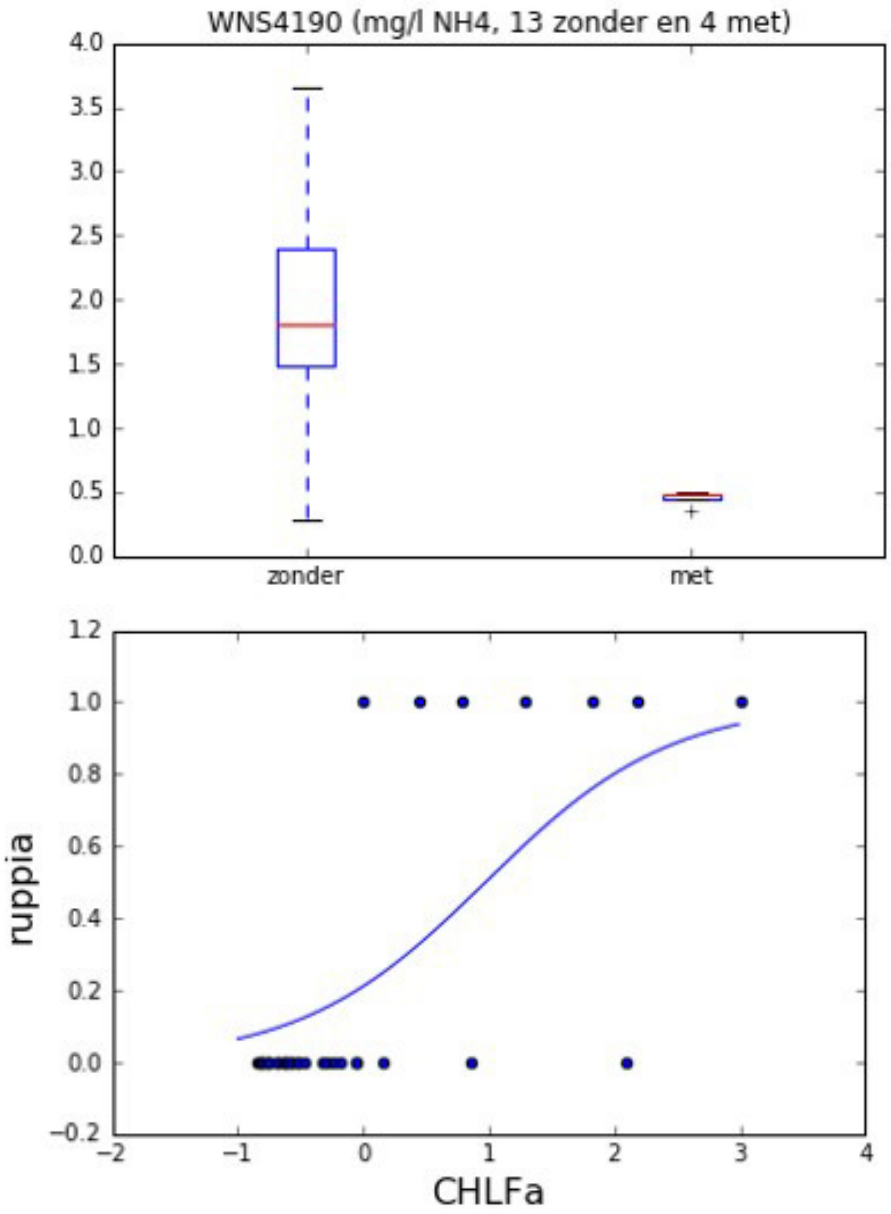




Figuur 7 Deel van een Python script dat op elegante wijze in één keer alle directories doorloopt, bestanden verzamelt, uitleest en analyseert, en vervolgens samenvoegt tot één bestand. Dit prototype is ontwikkeld door Daan de Waard.

Voorbeeld van Data Wrangling

Waterschap Scheldestromen wilde achterhalen waarom een de ruppia plant in bepaalde vooroevers wel groeide en in andere niet. Daarvoor was heel veel data beschikbaar, in 102 bestanden, verdeeld over acht directories. Onder meer over chemische elementen en verbindingen in het water. Die data hadden veel aandacht nodig: er waren diverse lege kolommen, de betekenis van variabelen was niet altijd duidelijk, de range van de waardes voor verschillende variabelen liep erg uiteen, waardes waren soms continu weergegeven (bijvoorbeeld 0,01) maar ook in (semi-)discrete vorm (zoals >0,1) et cetera. Om de data bijeen te brengen, bewerken en opschonen is een Python script geschreven. Een staging database is hierbij ook opgezet. Een mooi voorbeeld van de database en programming skills van de Data Scientist.



Figuur 8 Resultaat van Logistische Regressie op het bestand dat gecreëerd is voor Waterschap Scheldestromen. Deze analyse is uitgevoerd door Gert Jacobusse.

Voorbeeld van toepassing statistisch model/Machine Learning
Met de resulterende dataset (zie Figuur 8) is met verkennende data analyse achterhaald welke variabelen de meeste invloed hadden op het voorkomen van ruppia. Vervolgens is daarmee via logistische regressie een model gemaakt dat voorspelt of ruppia, gegeven de omstandigheden zal voorkomen of niet. Een mooi voorbeeld van de statistiek en machine learning skills van de Data Scientist.

TOEPASSING

Nut en noodzaak van Data Science zijn duidelijk. Dat we er mensen in moeten opleiden ook. Hoe doen we dat binnen de HZ University of Applied Sciences? Op de HZ wordt veel praktijkgericht onderzoek uitgevoerd. Bij dit type onderzoek staat de vraag vanuit de beroepspraktijk centraal. Vraagstukken worden onderzocht door lectoren, docentonderzoekers, professionals uit de praktijk én studenten. Verworven kennis uit het onderzoek levert nieuwe inzichten op, wat kan leiden tot innovaties en zelfs nieuwe producten voor het werkveld. Ofwel, de drie O's (de zogenaamde triple helix) onderwijs, onderzoek en ondernemers zijn onlosmakelijk met elkaar verbonden. Hier hoort een vierde O bij, namelijk overheid. Die heeft net zo goed behoefte aan innovatie en nieuwe producten, maar speelt ook andere rollen. Met betrekking tot Data Science zijn daarbij onder meer relevant: de leverancier van Open Data en de bewaker van privacy.

Integratie in het onderwijs

De opleiding HBO-ICT van de HZ omarmt het concept Student – en Procesgericht Onderwijs (SPO). Daarbij werken studenten aan hun competenties in projecten, gevormd rond een authentieke beroepssituatie en casuïstiek ingebracht door het werkveld of de lectoraten. Het lectoraat Data Science brengt per semester minimaal drie projecten in. Studenten werken daarin aan hun specifieke ICT-skills of, indien gekozen aan Data Science vaardigheden. Daarnaast kunnen studenten zich verdiepen in de combinatie onderzoek en Data Science in de minor Research & Innovation: Data Science. Voorbeelden van projecten en casuïstiek ingebracht door het lectoraat in samenwerking met het werkveld zijn onder meer:

- PROfessional Framework for Innovation in Tourism (PROFIT). In dit Interreg project fungeert het Kenniscentrum Kusttoerisme als interne opdrachtgever. Het vraagstuk betreft het gedrag van toeristen in een deltagebied. Welke inzichten in dat gedrag kunnen leiden tot innovaties voor lokale ondernemers? Aan de basis hiervan staat de slimme combinatie en analyse van systeemdata en Open data (zie ook Figuur 9 en het bijbehorende kader).
- Rijkswaterstaat Zee en Delta (RWS) in combinatie met de landelijke dienst Centrale Informatievoorziening (CIV). RWS wil het onderhoud aan haar vitale assets optimaliseren. Dit moet natuurlijk niet achteraf gebeuren, ook niet te vroeg maar precies op tijd. Om hier inzicht in te krijgen zijn enorme databestanden met metingen van onder meer allerhande sensoren beschikbaar. Als opstap naar grootschaliger onderzoek naar deze problematiek is gestart met een vooronderzoek naar de oorzaken van vaartuigverliesuren als gevolg van suboptimaal gepland onderhoud.
- Stichting Werkt Voor Ouderen Zorg (WVO). Op locatie Ter Reede, een WVO woonzorgcentrum, wonen veel cliënten die een vorm van dementie hebben. Diverse verzorgenden noteren hun bevindingen dagelijks in een bijbehorend cliëntdossier. Om inzicht te krijgen in de status of het welbevinden van een cliënt is daarom veel leeswerk nodig. Als oplossing is hiervoor een flexibele analysetool ontwikkeld. Die genereert, interactief een samenvatting (in een maximum aantal regels, als top-n van representatieve zinnen, op keywords et cetera). Deze tool en kennis is later gebruikt in het PROFIT-project, zie Figuur 9, en het bijbehorende kader.
- Zeeland Seaports. Als havenbeheerder beschikt deze organisatie over grote hoeveelheden data, onder meer AIS data. Die geeft informatie over locaties van schepen, vaarrichting en –snelheid, welke communicatie plaatsvindt et cetera. Zeeland Seaports is geïnteresseerd in mogelijkheden om met deze data veiligheid en efficiëntie in en rondom havens te verhogen.
- Delta Academy. Onderzoekers van deze academie hebben legio vragen. Een waaraan studenten recentelijk werkten was het achterhalen welke factoren de groei van mosselen, op diverse percelen in Zeeland, beïnvloedden.

Duurzame Dynamische Delta

De HZ heeft een relatief unieke locatie, namelijk midden in een deltagebied waar van alles gebeurt. De vraagstukken uit de beroepspraktijk die door de HZ worden opgepakt, hebben een link met deze delta. Het overkoepelende thema is dan ook 'Duurzame Dynamische Delta'. Dit zorgt voor veel variatie! Zo wordt er bijvoorbeeld onderzoek gedaan naar kustverdediging en kusttoerisme.

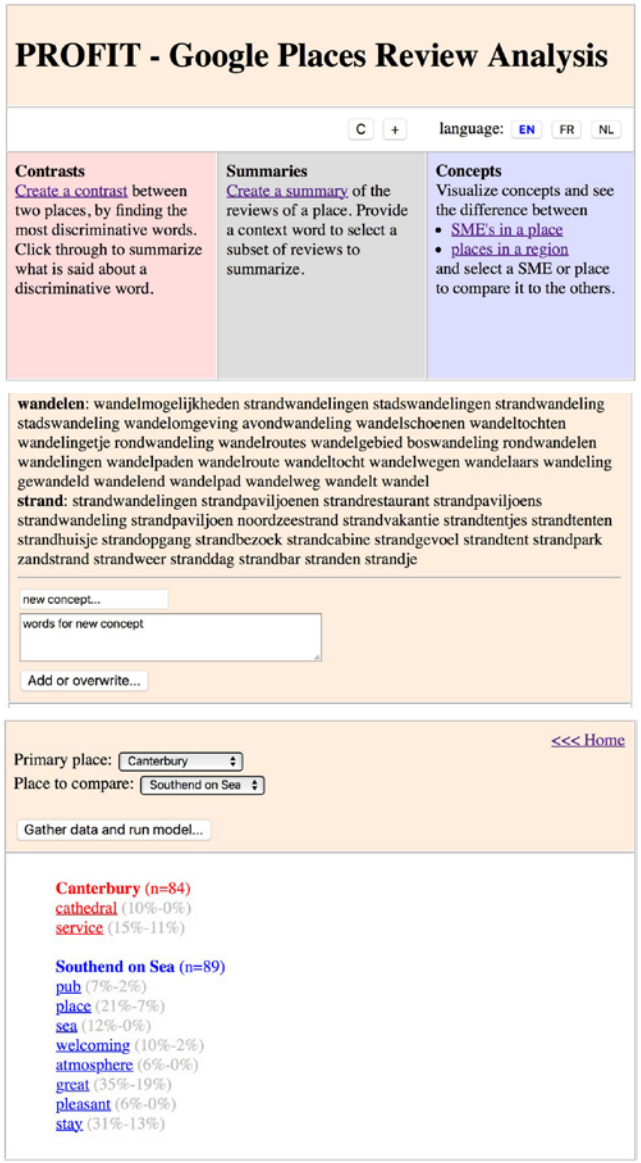
Net zoals in diverse andere contexten worden processen in deltagebieden in rap tempo gedigitaliseerd. In het specifieke geval van het deltagebied van en rondom Zeeland gaat het om processen in met name watermanagement, procesindustrie en de toeristische industrie. Dat hierbij steeds meer digitale data geproduceerd wordt is evident. Dat hieruit voor diverse belanghebbenden waarde te creëren is, ligt voor de hand. Hoe? Dat is de uitdaging waar we als data scientists, samen met die belanghebbenden aangaan.

Predictive Analytics in de context

Het deltagebied vormt een combinatie van diverse ecosystemen. Zoals die van de flora en fauna in en op het water. Bovendien is er druk scheepvaartverkeer op de aanwezige zeeën, rivieren en binnenwateren.
Het deltagebied is immers een interessant logistiek gebied met uitstekende verbindingen naar achterlanden. Dat maakt ook dat de procesindustrie er goed floreert. Evengoed is het een gebied waar diverse vormen van duurzame(re) energieopwekking steeds meer hun plek krijgen. Daarnaast trekt een deltagebied veel toeristen. Al deze ecosystemen hebben invloed op onder meer veiligheid, veerkracht en aantrekkingskracht van het gebied. En dan hebben we de klimatologische omstandigheden er nog niet bij betrokken. Het zou fantastisch zijn als we het geheel van deze ecosystemen en hun invloed op elkaar begrijpen. Daarvoor is het nodig om inzicht te hebben in de afzonderlijke systemen. Predictive Analytics speelt hierin een belangrijke rol. Samen met, zoals eerder opgemerkt,

Een app voor topic modeling en automatische samenvattingen

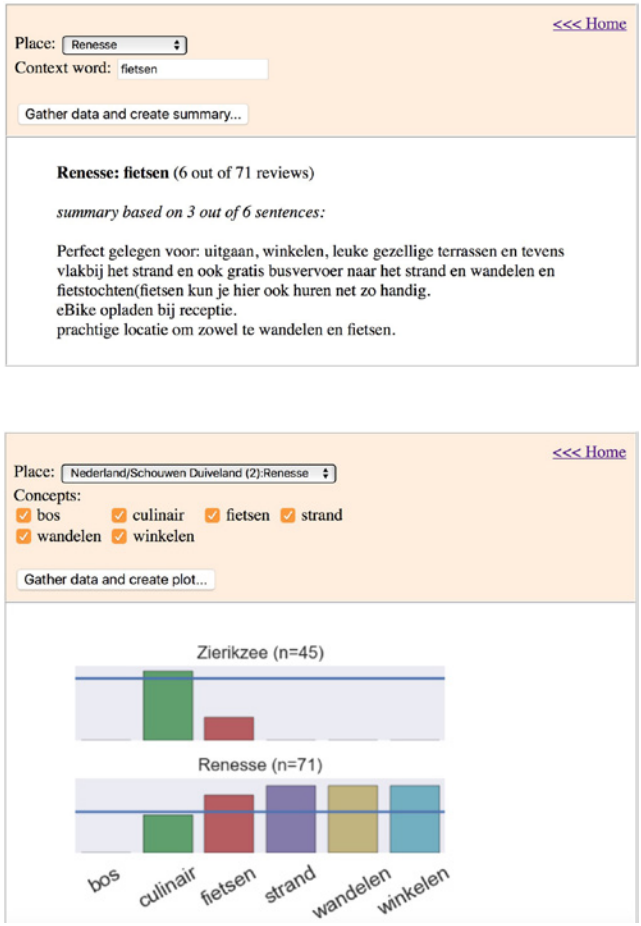
MKB-ondernemers in de vrijetijdseconomie willen graag concrete handvatten om te kunnen innoveren: bruikbare kennis over hun klanten en praktische adviezen die zij direct kunnen gebruiken om hun bedrijfsvoering te verbeteren. In het internationale project PROFIT werken HZ University of Applied Sciences en Economische Impuls Zeeland hieraan samen met een scala aan partners (HZ, 2016). Uitgebreidere informatie 'PROfessional Framework for Innovation in Tourism' staat op (Kenniscentrum Kusttoerisme, 2016). Om klantkennis te vergaren zijn inmiddels meer dan 10.000 databestanden verzameld. Figuur 9 geeft de resultaten voor een specifieke analyse met openbare data, namelijk Google reviews. Voor ondernemers uit PROFIT zijn Google reviews verzameld. Met een aantal specifieke Natural Language Processing algoritmes zijn hier topics uit bepaald en zijn automatische samenvattingen gegenereerd. Dit alles is interactief uitvoerbaar en te visualiseren met een specifieke webapp. Hierbij is nog niet gelet op look and feel maar vooral op het analyseresultaat. Hierbij worden een aantal skills van de Data Scientist gecombineerd namelijk domeinkennis, communicatie en visualisatie maar ook programming.



de expertise van aanpalende disciplines en belanghebbenden. De kracht van samenwerking kan niet vaak genoeg worden benadrukt.

Naar Predictive Maintenance

Een aantal vraagstukken uit ecosystemen in de deltagebieden heeft een gemeenschappelijk karakter: er zijn assets en op een enig moment is er een interventie nodig. Bijvoorbeeld omdat zo'n asset onderhoud nodig heeft. Dan is de vraag: wat is de kans dat iets kapot gaat of niet meer functioneert. In de procesindustrie gaat dat vaak om pompen en flenzen. In en op het water betreft het zogenaamde vitale assets waaronder dijken, sluizen en bruggen. Vanwege diverse redenen (economisch, duurzaamheid, veiligheid, imago et cetera) is het wenselijk om op tijd te signaleren dat een asset gaat falen. Dan kan namelijk, just-in-time onderhoud gepland worden. In dit geval spreken we dan over een speciale vorm van Predictive Analytics, namelijk Predictive Maintenance.



Figuur 9 Topic modeling en generatie van automatische samenvattingen op basis van Google reviews voor verblijfsrecreatie in Zeeland. Deze analyse is uitgevoerd door Gert Jacobusse.

Predictive Maintenance
Een presentatie op het Big Data Event van Kennis- en Innovatiecentrum Maintenance Procesindustrie (Ki<|MPi) / CAMPIONE op 20 oktober 2016 bracht het lectoraat Data Science in beeld voor wat betreft Predictive Maintenance (Procesindustrie, 2016). In samenwerking met DOW Benelux worden datasets bestudeerd met daarin meetwaardes van honderden sensoren en aanvullende periodieke metingen.

Vitale Infrastructuur
In het project Vitale Infrastructuur in een Veerkrachtige Delta neemt het lectoraat Data Science in combinatie met de opleiding HBO-ICT de rol op zich om ontwikkelde modellen ontsluitbaar te maken en beschikbaar te stellen (Vitale infrastructuur in de veerkrachtige delta, 2017).

VALKUILEN

We zagen dat waardecreatie met Data Science vanuit verschillende invalshoeken is te benaderen zoals economische winst, duurzaamheid en veiligheid.
Het is essentieel om kritisch naar de resultaten te kijken: doen we de goede dingen en doen we de dingen goed? Ook hier zijn verschillende invalshoeken mogelijk. Allereerst staan we stil bij de statistische/wiskundige invalshoek. Daarna worden enkele juridische aspecten beschreven. Tenslotte, is er de ethische kant van het verhaal. En hoewel er aparte paragrafen aan worden geweid, overlappen de invalshoeken elkaar en liggen ze in elkaars verlengde.

Goochelen met resultaten

De disciplines statistiek en wiskunde vormen een wezenlijk onderdeel van Data Science. Er zijn legio bronnen die aangeven hoe die statistiek en wiskunde ingezet en geïnterpreteerd zouden moeten worden. In de praktijk pakt dat soms anders uit, bewust of onbewust. Daniel Levitin zet dat in zijn werken helder uiteen. Hij stelt: *"Informatie is tegenwoordig bijna onmiddellijk beschikbaar, maar het wordt steeds moeilijker om uit te maken wat waar is en wat niet, om de verschillende beweringen die we voorgeschoteld krijgen te beoordelen, en te bepalen of ze foute informatie, pseudofeiten, verdraaiingen van de werkelijkheid of regelrechte leugens bevatten"*. In (Levitin, 2016) ontrafelt hij twee categorieën waardoor misleiding kan ontstaan: verkeerd gebruikte statistieken en grafieken, en verkeerd gebruikte argumenten. Met legio voorbeelden. Over ecologische fouten bijvoorbeeld (conclusies trekken over individuen op basis van een groepsgemiddelde), of juist over overgeneralisatie (conclusies trekken over een groep op basis van kennis over een paar uitzonderingen). Natuurlijk komen gegoochel met assen in grafische weergaves, of andere misleidende illustraties van resultaten terug.
En het zogenaamd vinden van causale verbanden uit correlaties, zonder dat er gecontroleerde experimenten plaatsvonden. Het verzamelen en gebruiken van relevante data is cruciaal. Machine learning algoritmes gaan immers gewoon aan de slag met de data die aangeboden wordt. En leren wat er in die dataset aan informatie aanwezig is. Dat kan tot bijzondere resultaten leiden, blijkt uit voorbeelden die her en der te lezen zijn. Zo wilde een headhunter een profiel opstellen van geschikte kandidaten voor een bepaald type, zware bestuursfunctie. Uit het model dat met machine learning was opgesteld bleek dat vrouwen ongeschikt waren voor deze functie. Nader onderzoek leerde dat de trainingset die aangeboden was om het model op te stellen data bevatte van slechts enkele vrouwelijke bestuurders. Hadden er niet meer vrouwelijke bestuurders in de dataset moeten zitten? Was die data er wel? Vragen die raken aan het aspect representativiteit van de dataset. Maar is dat relevant in dit geval? De zoektocht betrof het profiel van een goede bestuurder. Niet of dat een man of vrouw zou moeten zijn. Ofwel, in dit geval moet de variabele geslacht helemaal niet meegenomen worden in het model. Een ander voorbeeld. Chatbots zijn computerprogramma's waarmee een persoon een conversatie kan voeren (of die onderling communiceren). Chatbots leren te communiceren door ze heel veel teksten (en eventueel vraag – antwoord combinaties) aan te bieden. Tweets bijvoorbeeld. Tijdens de meest recente Amerikaanse presidentsverkiezingen werd een verkiezingsbot geïntroduceerd die getraind was en werd met tweets. Toen een bepaalde groepering dat doorhad, plaatste die een grote hoeveelheid racistische tweets. Met als gevolg dat de bot op den duur racistische antwoorden gaf op allerhande vragen.

In dit kader is het goed om nog eens te refereren aan de cognitive services die genoemd worden in de paragraaf Data Science in de praktijk.

Privacy

Wat er inmiddels kan met data en Data Science leeft op gespannen voet met wat er daadwerkelijk mag. Zeker als het gaat om het verzamelen en bewerken van persoonsgegevens. Dat naam en adres onder persoonsgegevens vallen weet iedereen wel. Maar bijvoorbeeld ook een bankrekeningnummer en IP-adres horen daarbij. Een beknopte samenvatting van wat verstaan wordt onder persoonsgegevens geeft (MKBServicedesk, 2017). (Chew-Meij, BigDatamaandag deel 3: Persoonsgegevens en gegevensverwerking, 2014) gaat verder in op daadwerkelijke verwerking van persoonsgegevens. De Autoriteit Persoonsgegevens houdt toezicht op de naleving van de wettelijke regels voor bescherming van persoonsgegevens, ofwel de Wet bescherming persoonsgegevens (Wbp). In mei 2018 wordt de Wbp vervangen door de Nieuwe Europese Wet, ook wel bekend als de General Data Protection Regulation (GDPR) (Autoriteit persoonsgegevens, 2017).

Om te voorkomen dat daadwerkelijk persoonsgegevens verwerkt worden kiest men vaak voor anonimiseren van die gegevens. Bijvoorbeeld door een naam te vervangen door een willekeurige reeks karakters. Dat hieraan risico's kleven is te lezen in (Chew-Meij, Bigdatamaandag Deel 8, het Anonimiseren van Databases, 2014) en daarop volgende stukken. Vaak is namelijk uit geanonimiseerde data indirect af te leiden welke persoon het betreft. Dat we voorzichtig moeten zijn met persoonsgegevens bleek al uit het onderzoek van Matthijs Koot, *"Geanonimiseerde gegevens zijn*

minder anoniem dan gedacht [...]. Koot onderzocht van 2,7 miljoen Nederlanders gegevens uit de gemeentelijke basisadministratie (GBA) over geboortedatum, postcode en geslacht. Daaruit bleek onder meer dat 67 procent van de personen uniek identificeerbaar is op de combinatie van geboortedatum en de vier cijfers van de postcode" (Koot, 2012).

Een aanrader voor wie zich hierin wil verdiepen is (Martijn & Tokmetzis, 2016), een boek met talloze voorbeelden over het verzamelen en verwerken van persoonsgegeven en inbreuken op privacy, dat niet voor niets de subtitel "Over het Levensbelang van Privacy" meekreeg[11].

Ethiek

Stel nu dat er voldoende relevante data verzameld is, dat de Wbp niet overtreden wordt en dat op de juiste manier een algoritme is getraind en een model is gemaakt. Wat vinden we dan van het gebruik daarvan? Cathy O'Neil is een van de onderzoekers die daar diep induikt. In haar boek "Weapons of Math Destruction" (O'neil, 2016) haalt ze talloze voorbeelden aan van Big Data / Data Science toepassingen die in haar ogen ongelijkheid in de wereld verhogen en de democratie aantasten. Ze benadrukt hiervoor een aantal vicieuze cirkels. Een van haar voorbeelden: arme mensen wonen in een omgeving waar meer criminaliteit voorkomt. Classificatiemodellen zouden deze omgeving, op basis van gebruikte data dan als risicogebied kunnen aanduiden. Daarom gaat de politie daar meer surveilleren en vinden meer arrestaties plaats. Die data gaat weer het model in met als gevolg dat de ranking als risicogebied omhoog gaat. Met gevolgen voor alle bewoners: met betrekking tot kans op werk, verhoging van verzekeringspremies et cetera. Eenzelfde vicieuze cirkel treed op in het geval van bijvoorbeeld kredietwaardigheid of het aanbieden van leningen in zo'n omgeving.

Kortom, het is goed om wat er met Big Data / Data Science gebeurt nauwkeurig langs een ethische meetlat te leggen. O'Neil gaat er met de botte bijl in en ziet veel 'slechte' voorbeelden van Big Data / Data Science toepassingen en soms doet ze boude uitspraken als zouden de algoritmes ('weapons of math destruction') de 'schuldigen' zijn.
Het goede is dat ze hierdoor een grote, en zinvolle discussie losmaakt over de ethiek achter dit vakgebied.
Zie bijvoorbeeld het debat onder (O'Neil, 2017).

Vastleggen persoonsgegevens

Een voorbeeld is de reclamezuil met VidiReports technologie. Die maakt een analyse van de mensen die het reclamebord bekijken. VidiReports detecteert automatisch het gezicht van iedere persoon die voorbij de zuil loopt, en stelt vast wie de reclame bekijkt. Van die mensen bepaalt de software in een oogwenk het geslacht en de leeftijdscategorie. Zelfs het humeur kan afgeleid worden. Interessant voor commerciële toepassingen. De exploitant van de zuilen, Exterion, garandeert dat geen data wordt opgeslagen en dat verzamelde data niet tot een persoon is te herleiden. Toch plaatste de Autoriteit Persoonsgegevens (AP) vraagtekens bij de rechtmatigheid van de reclamezuilen. Zij stelt dat als iemand gefilmd wordt, per definitie zijn of haar persoonsgegevens vastgelegd worden. En als dat uit commerciële overwegingen gebeurt, mag dat niet zomaar. Toch durfde de AP niet definitief te stellen dat Exterion in overtreding was (Sondermeijer, 2017).

[11] Privacy is een complex begrip. (Koops, et al., 2016) geeft bijvoorbeeld een overzicht van verschillende typen privacy. Er is veel wetenschappelijke literatuur over verschenen onder meer van Daniel Solove, die diverse boeken schreef over de relatie privacy en informatiesystemen/databases.

TOEKOMST

Tot nu toe kwamen aan de orde: nut en noodzaak van Data Science, toepassing en focus van Data Science in een deltagebied en bekende valkuilen. De combinatie hiervan geeft de juiste munitie om een rijk en verantwoord toekomstbeeld te schetsen. Waar liggen de kansen en uitdagingen en wat zijn potentiële onderzoeksvragen?

Data Science Lab

In eerdere hoofdstukken is de samenwerking van HZ met Rijkswaterstaat beschreven. Die organisatie maakt momenteel een transitie door naar een sterk data gedreven organisatie. Met behoorlijke nadruk op Predictive Maintenance getuige publicaties zoals (Rijkswaterstaat, 2016)[12] en (Rijkswaterstaat, 100% Voorspelbaar Onderhoud? Vitale assets, Proces Optimalisatie, een Data Gedreven RWS!, 2017)[13], en natuurlijk de nominatie voor de Computable Awards 2017 in de categorie Digital Transformation van het Jaar: RWS Datalab (Rijkswaterstaat, Computable Awards 2017: RWS Datalab, 2017). Het Datalab in Delft speelt voor wat betreft de hiervoor benodigde dataverzameling en –analyse een prominente rol.
Veel projecten betreffen het landelijke wegennet. Met onder meer patroonherkenning op gedrag in de schipholtunnel, risicokwantificering van wegen en ongevallenvoorspelling, ofwel plekken herkennen waar mogelijk ongelukken gaan gebeuren.

In het deltagebied is het van belang soortgelijke exercities te kunnen uitvoeren voor het vaarwegennet. Een recent Rijkswaterstaat voorbeeld om van te leren is het duiden van atypisch gedrag in een sluizencomplex in Tiel. Data van sensoren rondom en in het complex zijn gecombineerd met aanvullende gegevens zoals weerdata en waterstanden. Daarmee zijn normen vastgelegd en bij geconstateerde nadering of overschrijding van die norm kan er worden ingegrepen. En, kan eventueel atypisch gedrag nader onderzocht worden.

De HZ, University College Roosevelt/University of Utrecht en ROC Scalda werken samen aan de oprichting van een Joint Research Center Zeeland (JRCZ). Deze experimentele onderwijs - en onderzoeksfaciliteit, van circa 4500m2 vanuit waar op drie onderwijsniveaus (mbo, hbo en wo) wordt gewerkt aan de human Capital agenda van Zeeland, wordt gebouwd op de locatie Groene Woud te Middelburg. Naast onderwijs - en onderzoekslaboratoria op de thema's Ecologie, Chemie, Fysica en Engineering wordt ook een state-of-the-art Data Science Lab ingericht. In Bijlage 4 staat de oorspronkelijke notitie die beschrijft waarom dit Data Science Lab nodig is, welke mogelijkheden het biedt voor de regio en hoe er gestart kan worden. Het lab is een prachtige ontmoetingsplek voor Data Science onderzoek en onderwijs in een deltacontext en sluit daarmee uitstekend aan op nationale agenda's zoals de Kennis- en Innovatie Agenda Deltatechnologie 2018-2021 (Topsector Water, 2017).

Techniek

Hoewel basisingrediënten van Data Science al langer voorhanden zijn groeit de interesse en toepassing van het vakgebied als geheel, exponentieel. Toepassing van proven technology biedt nog ruim kansen voor toegepast onderzoek, getuige uitspraken als "De meeste bedrijven hebben niet zo'n helder inzicht in hun eigen activiteiten. Een groot deel van de mogelijkheden van Big Data op korte termijn is dan ook niet gelegen in complexe kunstmatige intelligentie, maar in het veel aardsere tellen, bewaken en zien van dingen met een grotere precisie en helderheid " (Lohr, 2015).

Sneller rekenen

Maar, de ontwikkelingen in de technieken staan niet stil. De mogelijkheid om steeds meer data, in plaats van nauwkeurig verzamelde kleinere steekproeven te onderzoeken werd groter naarmate meer rekenkracht beschikbaar kwam. Dat kan door algoritmes parallel en of gedistribueerd te laten werken op een veelvoud aan processoren. De laatste jaren zijn er ook ontwikkelingen gaande waarbij niet meer op de central processing unit (CPU) van de computer(s) gerekend wordt maar op de graphical porcessing unit (GPU). Of, er wordt in-memory gerekend. En deze ontwikkelingen staan niet stil.

Nieuwe algoritmes

Ook op het gebied van algoritmes is veel mogelijk. Zogenaamde neurale netwerken worden al lang gebruikt. Deze algoritmes leren op dezelfde manier als onze eigen hersenen leren. Met zo'n algoritme wordt de structuur van de hersenen, waarbij neuronen signalen aan elkaar doorsturen, gesimuleerd. Hoe meer (lagen met) neuronen, hoe complexer de dingen die zo'n netwerk kan leren. Nu we over steeds meer rekenkracht beschikken, kunnen we grotere neurale netwerken (met heel veel lagen met neuronen) bouwen, die inzetbaar zijn voor erg complexe problemen. Die netwerken heten daarom Deep Neural Networks. Zie ook Bijlage 5. Evenzo ontstaat de laatste tijd veel interesse in Process Mining. Data Mining was al langer bekend. Bij Process Mining probeert men van event logs, de onderliggende processen af te leiden. Event logs worden bij veel processen geproduceerd en bevatten in het simpelste geval een (case) id, de naam van een event (een activiteit of gebeurtenis) en een timestamp (een combinatie van een datum en tijd). Met algoritmes die de frequentie van activiteiten, de volgorde et cetera combineren kan bepaald

[12] Gerelateerd aan het Interreg project Be Good.
[13] Gerelateerd aan het Campione Fieldlab.

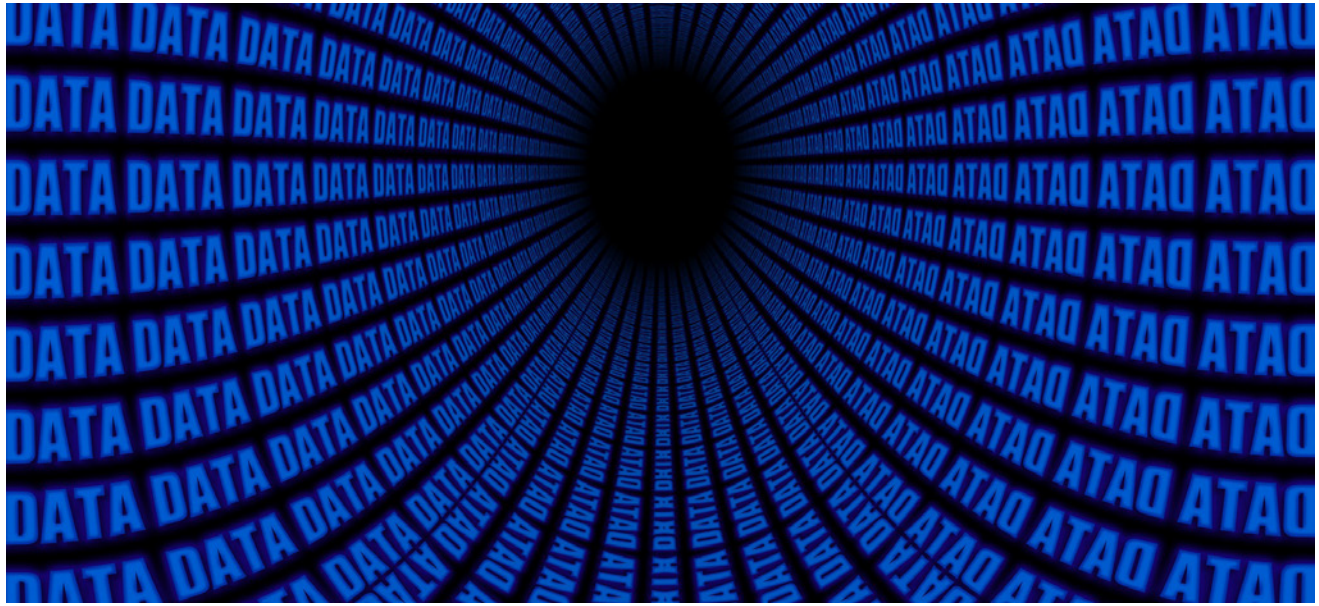
worden wat het eigenlijke proces was. Voorbeelden van activiteiten die gelogd worden zijn er legio: navigatie op een web site, mailgedrag en activiteiten voor klantenservice. Het interessante is dat in Maintenance, in zijn algemeenheid ook heel veel gelogd wordt. Er is dus veel aan gelegen om de kansen van Process Mining in Predictive Maintenance te gaan ontdekken. In de context van aangehaalde casuïstiek kunnen hierbij onderzoeksvragen opgesteld worden over welke nieuwe informatie dit soort algoritmes boven water krijgt.

Nieuwe visualisaties

Een plaatje zegt meer dan duizend woorden geldt ook in Data Science. Een eerste inspectie van data behelst vrijwel altijd visualisatie. Evenzo kan het eindproduct van een Data Science cyclus een visualisatie zijn (een interactief dashboard, een graph die de invloed van meerdere factoren in beeld brengt et cetera). Het ligt voor de hand om bij visualisatie vooruit te kijken naar de combinatie van Data Science met andere upcoming technologies, zoals Augmented en Virtual Reality.

Context

Mooie onderzoeksvragen zijn er ook op het gebied van Predictive Maintenance. Om te kunnen voorspellen wanneer onderhoud nodig is moet bekend zijn wanneer een 'asset' kapot is, of faalt. Stel, het betreft een pomp. Wat definieert wanneer die kapot is, ofwel gefaald heeft? Het spelen met deze definitie kan mogelijke gevolgen hebben voor de kwaliteit van de voorspelling. Een andere onderzoeksvraag ligt op het gebied van generalisatie. Kunnen we Predictive Maintenance als concept inzetten in contexten waar de context niet de traditionele assets betreft? Bijvoorbeeld in de toeristische industrie. Als toeristen niet meer naar een bepaalde locatie komen kan dat geïnterpreteerd worden als falen. Dat willen we voorspellen. Kunnen we daarvoor inzichten uit Predictive Maintenance gebruiken?



DO BELIEVE THE HYPE!

Gaan computers en algoritmes ons nu overnemen? Een veelgehoorde vraag, en wellicht angst. Nee, althans, nog lange tijd niet, blijkt uit Gijsbert Werners essay (Werner, 2017). Natuurlijk, er worden zeer veel vorderingen gemaakt in zelfrijdende auto's, robots als personal assistant et cetera. En het lukt algoritmes vaak zeer goed om te classificeren. Zo zijn er algoritmes getraind die de juiste auteur bij een roman aangeven. Of, de schilder van een bepaald schilderij. Maar op minimaal één belangrijk aspect lopen ze achter: creativiteit. Kan zo'n getraind algoritme ook zelfstandig iets produceren bijvoorbeeld? In de muziek komt men een eind. Zo bestaat er al software die gegeven een bepaalde akkoordsequentie een goed klinkende solo produceert. Harmonisch en melodisch goed. Soms zelfs in de stijl van een bepaalde artiest. Machine Learning kan prima omgaan met teksten. Automatisch samenvattingen genereren bijvoorbeeld. Interpretatie blijft echter mensenwerk. Zo ontwierp Folgert Karsdorp Asibot, een algoritme dat getraind is met romanteksten om op basis daarvan een suggestie te doen voor een eigen romantekst. Een suggestie, want er moet nog een echte schrijver aan te pas komen die regie over het proces houdt (NOS, 2017). Dit is in overeenstemming met diverse uitspraken op dit gebied zoals *"Juist omdat je de interpretatie [...] niet kunt automatiseren. Die is in elke bedrijfssituatie anders. Het blijft mensenwerk"* (Lonkhuyzen, 2017) en *"Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide"* (O'neil, 2016).

In deze rede heb ik achtereenvolgens behandeld het nut en de noodzaak van Data Science, de toepassing en focus van Data Science in een deltagebied, bekende valkuilen, toekomstige ontwikkelingen, daaruit vloeiende kansen en onderzoeksvragen. Genoeg te doen. Nu aan de slag. Ik heb betoogd dat Data Science geen *"modegril [is], die bijna altijd overwaait"* maar dat het gaat om een "een werkelijk belangwekkend verschijnsel". Do believe the hype!

BRONNENLIJST

- *Autoriteit persoonsgegevens.* (2017). Algemene verordening gegevensbescherming. Opgehaald van autoriteitpersoonsgegevens.nl: https://www.autoriteitpersoonsgegevens.nl/nl/onderwerpen/europese-privacywetgeving/algemene-verordening-gegevensbescherming
- *Beckers,* M. (1997). Chemometrics in the Conformational Analysis of Biomacromolecules; Exploring the Possibilities. Dordrecht: Kluwer Academic Publishers.
- *Bicorner.* (2015, 06 11). INFOGRAPHIC: The surprising things you don't know about Big Data. Opgehaald van bicorner.com: https://bicorner.com/2015/06/11/infographic-the-surprising-things-you-dont-know-about-big-data/
- *Buydens,* L., Reijmers, T., Beckers, M., & Wehrens, R. (1999). Molecular data-mining: a challenge for chemometrics. Chemometrics and Intelligent Laboratory Systems, 121-133.
- *Cao, L.* (2017). Data Science: Challenges and Directions. Communications of the ACM, 60(8), 59-68.
- *Chew-Meij,* L. (2014, 07 21). BigDatamaandag deel 3: Persoonsgegevens en gegevensverwerking. Opgehaald van ictrecht.nl: https://ictrecht.nl/2014/07/21/bigdatamaandag-deel-3-persoonsgegevens-en-gegevensverwerking/
- *Chew-Meij,* L. (2014, 08 25). Bigdatamaandag Deel 8, het Anonimiseren van Databases. Opgehaald van https://ictrecht.nl/privacy/: https://ictrecht.nl/2014/08/25/bigdatamaandag-deel-8-het-anonimiseren-van-databases/
- *Cukier, K.*, & Mayer-Schoenberger, V. (2013, 5/6). The Rise of Big Data. Foreign Affairs. Opgehaald van https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data
- *Delta, D. D. (sd).* Actieplan voor ICT-uitdagingen in Nederland. Opgehaald van dutchdigitaldelta.nl: https://www.dutchdigitaldelta.nl/actieplan
- *Denktank, N. (2014, 12 8).* Big Data in zicht. Opgehaald van nationale-denktank.nl: http://nationale-denktank.nl/wp-content/uploads/2015/03/Eindrapport_Big_Data_in_zicht-Nationale_DenkTank_2014.pdf
- *Domingos, P. (2015).* The Master Algorithm. Basic Books.
- *European Commission. (2017, 10 02).* Big Data and Data Science. Opgehaald van ec.europa.eu: https://ec.europa.eu/eurostat/cros/content/big-data-and-data-science_en
- *Foreman, J. (2013).* Data Smart: Using Data Science to Transform Information into Insight. Indianapolis, Indiana: John Wiley & Sons, Inc.
- *Gartner.* (2017, 07 28). Hype Cycle for Data Science and Machine Learning. Opgehaald van gartner.com: https://www.gartner.com/doc/3772081/hype-cycle-data-science-machine
- *Gartner.* (2017). Predictive Analytics. Opgehaald van Gartner.com: http://www.gartner.com/it-glossary/predictive-analytics
- *Hijink, M.* (2017, 09 02/03). De grote datarace: een rijbewijs voor de robotauto. Opgehaald van nrc.nl: https://www.nrc.nl/nieuws/2017/09/01/waardoor-de-zelfrijdende-auto-zakt-voor-zijn-rijexamen-12777304-a1571900
- *HZ.* (2016). PROFIT. Opgehaald van hz.nl: https://hz.nl/projecten/profit
- *Informatiehuis Marien.* (2017, 03 21). Symposium Digishape. Opgehaald van informatiehuismarien.nl: http://www.informatiehuismarien.nl/nieuws/alle-nieuwsberichten/2017/symposium-digishape/
- *Kenniscentrum Kusttoerisme.* (2016). PROFIT, klantkennis voor mkb-ondernemers. Opgehaald van profit.kenniscentrumtoerisme.nl: http://profit.kenniscentrumtoerisme.nl/
- *Klous, S., & Wielaard, N.* (2014). Wij zijn Big Data. Amsterdam. Business Contact.
- *Koops, B.-J.*, Clayton Newell, B., Timan, T., Škorvánek, I., Chokrevski, T., & Galic, M. (2016, 03 24). A Typology of Privacy. University of Pennsylvania Journal of International Law, 483-575. Opgehaald van http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2754043
- Koot, R. (2012, 05 22). Persoonsgegevens Gba Gemakkelijk Identificeerbaar. Opgehaald van Measuring and Predicting Anonymity: https://blog.cyberwar.nl/2012/05/measuring-and-predicting-anonymity-phd-thesis/
- *Kroes, N.* (2014, maart 10). Twitter.com. Opgehaald van https://twitter.com/NeelieKroesEU/status/442985759277346816
- *Levitin, D.* (2016). A Field Guide to Lies, Critical Thinking in the Information Age. Dutton.
- *Lohr, S.* (2015). The Revolution Transforming Decision Making, Consumer Behavior and Almost Everything Else. Maven Publishing B.V.
- *Lonkhuyzen, L.* (2017, 08 08). Nederland heeft een lakse houding. Opgehaald van nrc.nl: https://www.nrc.nl/nieuws/2017/08/08/nederland-heeft-een-lakse-houding-11178248-a1569201
- *Martijn, M., & Tokmetzis, D.* (2016). Je Hebt Wel Iets te Verbergen . Druk Koninklijke Woehrmann .
- *MKBServicedesk. (2017, 06 12).* Het verwerken en omgaan met persoonsgegevens. Opgehaald van mkbservicedesk.nl: https://www.mkbservicedesk.nl/88/het-verwerken-omgaan-met-persoonsgegevens.htm

- *Nationale Wetenschapsagenda.* (2015, 11 27). Nationale Wetenschapsagenda Nederlands. Opgehaald van wetenschapsagenda.nl: https://wetenschapsagenda.nl/publicatie/nationale-wetenschapsagenda-nederlands/
- *Ng, A.* (2017, 09). Introduction to Deep Learning. Opgehaald van coursera.org: https://www.coursera.org/learn/neural-networks-deep-learning/lecture/praGm/why-is-deep-learning-taking-off
- *NOS.* (2017, 06 16). Dit is de robot die je van je writer's block afhelpt. Opgehaald van nos.nl: https://nos.nl/op3/artikel/2178567-dit-is-de-robot-die-je-van-je-writer-s-block-afhelpt.html
- *O'neil, C., & Schutt, R.* (2014). Doing Data Scienc. Sebatopol:O'Reilly Media, Inc.
- *O'neil, C.* (2016). Weapons of Math Destruction. Crown.
- *O'Neil, C.* (2017, 04). The Era of Blind Faith in Big Data Must End. Opgehaald van ted.com: https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end/discussion
- *Procesindustrie, K. e.* (2016, 10 20). Mainenance Procesindustrie. Opgehaald van kicmpi.com: http://kicmpi.com/nieuws.html
- *Provost, F., & Fawcett, T.* (2013). Data Science for Business. Sebastopol: O'Reilly Media, Inc.
- *Radboud University.* (2017). Analytical Chemistry: Chemometrics. Opgehaald van Radboud Universiteit: http://www.ru.nl/science/analyticalchemistry/
- *Rijkswaterstaat.* (2016, 07). Vitale assets. Elektriciteitsverbruik als graadmeter voor conditiegestuurd onderhoud. Opgehaald van www.h2owaternetwerk.nl: https://www.h2owaternetwerk.nl/images/knw/170622_RWS_Vitale_assets.pdf
- *Rijkswaterstaat.* (2017). 100% Voorspelbaar Onderhoud? Vitale assets, Proces Optimalisatie, een Data Gedreven RWS! Opgehaald van geonovum.nl: https://www.geonovum.nl/sites/default/files/1-100%25%20voorspelbaarheid.pdf
- *Rijkswaterstaat.* (2017, 09). Computable Awards 2017: RWS Datalab (Rijkswaterstaat). Opgehaald van youtube.com: https://www.youtube.com/watch?v=bLs3KsP3MOI
- *Seeters, R.* (2017, 05 22). Wat nemen we mee van de Gartner Summit? Opgehaald van inergy.nl: https://inergy.nl/wat-nemen-we-mee-van-de-gartner-summit/
- *Shanahan, M.* (2015). The Technological Singularity. Massaschusetts Institute of Technology.
- *Shearer, C.* (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5, 13-22.
- *Snijders, C., Matzat, U., & Reips, U.* (2013). Big Data: Big Gaps of Knowledge in the Field of Internet Science. International Journal of Internet Science, 7(1), 1-5.
- *Sondermeijer, V.* (2017, 09 05). Plotseling heeft de reclamezuil ogen gekregen. Opgehaald van nrc.nl: https://www.nrc.nl/nieuws/2017/09/05/plotseling-heeft-de-reclamezuil-ogen-gekregen-12849487-a1572341
- *The Apache Software Foundation.* (2017, 08 04). Welcome to Apache™ Hadoop®! Opgehaald van hadoop.apache.org: https://hadoop.apache.org
- *Topsector Water.* (2017, 09). Kennis en Innovatieagenda Deltatechnologie 2018-2021. Opgehaald van tkideltatechnologie.nl: https://www.tkideltatechnologie.nl/wp-content/uploads/2017/10/Kennis-en-Innovatieagenda-2018-2021-Definitief.pdf
- *Vereniging Hogescholen.* (2016, 08). Onderzoek met Impact. Stategische onderzoeksagenda hbo 2016 - 2020. Opgehaald van vereniginghogescholen.nl: http://www.vereniginghogescholen.nl/system/knowledge_base/attachments/files/000/000/601/original/Onderzoek_met_Impact_%28website%29.pdf?1471948142
- *Vitale infrastructuur in de veerkrachtige delta.* (2017). Opgehaald van hz.nl: https://hz.nl/projecten/vitale-infrastructuur-in-de-veerkrachtige-delta
- *Wagstaff, J.* (2015, 10 28). Smart Technology? It's In Our Blood. Opgehaald van channelweb.co.uk: http://www.channelweb.co.uk/crn-uk/opinion/2432561/smart-technology-its-in-our-blood
- *Werner, G.* (2017, 10 14-15). De menselijke geest uniek? Dat had u gedacht. NRC, pp. O&D4-5.
- *Wold, S.* (1995). Chemometrics; what do we mean with it, and what do we want from it? Chemometrics and Intelligent Laboratory Systems, 30(1), 109-115.
- *Zawadzki , K.* (2014, 08 30). Data science skill-set explained. Opgehaald van marketingdistillery.com: http://www.marketingdistillery.com/2014/08/30/data-science-skill-set-explained/

DANKWOORD

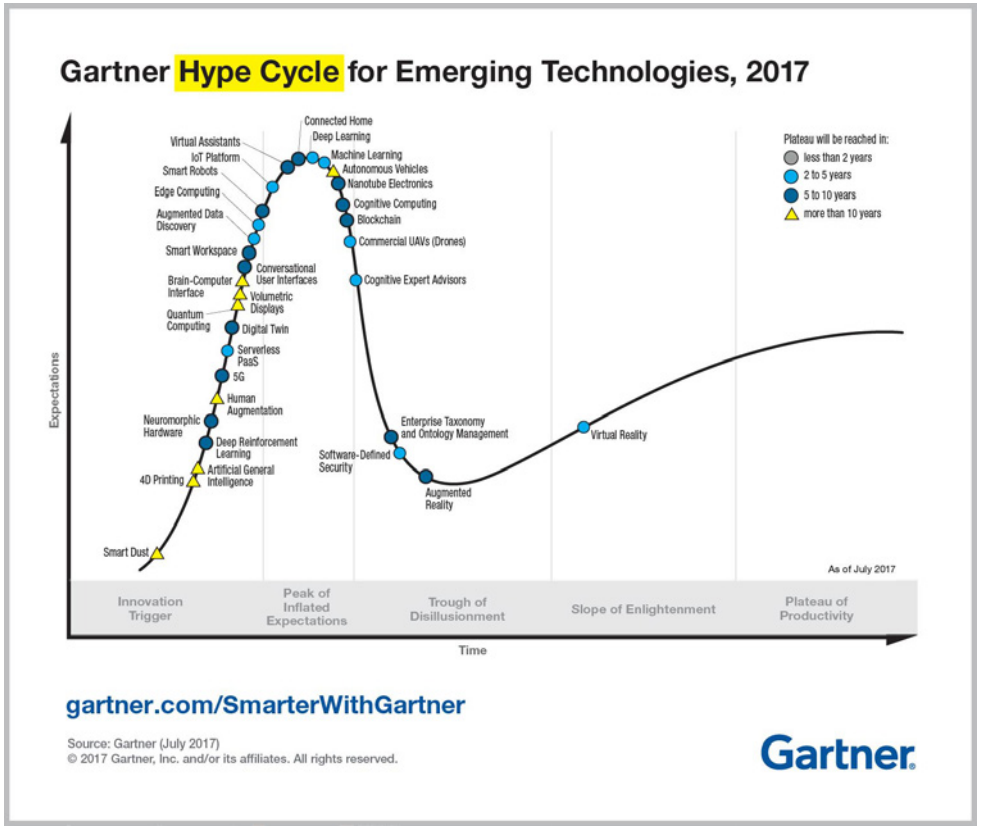
Een dankwoord. Dat is natuurlijk op zijn plaats, want ik ben dankbaar voor de kansen die ik kreeg. Maar, ook lastig. Omdat ik niet iedereen bij naam kan noemen, en niemand te kort wil doen. Hier gaan we.
Willem den Ouden en Hans de Bruin boden de opening naar een lectoraat Data Science. Zij signaleerden relevante vraagstukken en vroegen me heel direct of ik daar mijn kennis en kunde op wilde inzetten. Bert Schollema en Frank Bordui creëerden vervolgens de ruimte om daarmee binnen de Academie voor Technologie en Innovatie aan de slag te kunnen. Natuurlijk kan dat alleen in samenspraak met de opleiding waar ik aan verbonden ben. Met Jorick Vos was het daarover altijd aangenaam sparren. Ook door het docententeam voelde ik me altijd gesteund. Een speciaal bedankje is hierbij op zijn plaats voor Gert Jacobusse, Mathieu Starink, Daan de Waard en Jolene Cijsouw, die samen mijn kenniskring vorm(d)en. Adri de Buck en John Dane van het College van Bestuur gaven me het vertrouwen om het lectoraat te starten. Fijn om daarbij op te kunnen trekken met (inmiddels) collegalectoren binnen de AvTI Willem Böttger, Jacob van Berkel en Dorien Derksen. Ik ben erg blij dat ik deel uit kon maken van de kenniskring van Gerard Schouten en Erik van Tol en van het lectoraat Big Data van Fontys Hogescholen. Daar heb ik veel van geleerd. Ook van de discussies en informele gesprekken met Edwin Torn Broers.
Ik voelde me meteen thuis bij het overleg en andere activiteiten van de lectoren en onderzoekscoördinatoren van de Delta Academie. Het is hard werken voor het PROFIT project. Maar, dat vergeet je snel als je dat kunt doen met de fijne collega's van het Kenniscentrum Kusttoerisme.

CV

Mischa behaalde bij de Hogere Laboratorium Opleiding in Goes de Bachelor Analytische Chemie. Hij sloot vervolgens een verkort doctoraal opleiding aan de Radboud Universiteit in Nijmegen af als doctorandus in de Analytische Chemie / Chemometrie. In 1994 trad hij bij die universiteit in dienst als onderzoeker. Daar ontwikkelde en doceerde hij het lesprogramma "Statistiek voor chemici" en promoveerde in 1997 op het proefschrift "Chemometrics in the conformational analysis of biomacromolecules: exploring the possibilities". Daarna ging hij bij Kluwer Academic Publsihers aan de slag als specialist electronic publishing. In 2000 maakte Mischa de overstap naar Recreanet en werd IT-consultant voor met name de toeristische industrie. De overgang van Recreanet in Maxxton betekende ook een overgang in functie: manager R & D. En steeds ging het om dezelfde fascinatie: informatie maken uit data. In 2005 ontstonden de contacten met de HZ. Mischa werd gevraagd om gastlessen te geven bij de net opgezette opleiding ICT. Toen daar een vacature ontstond maakte hij de stap naar het hoger onderwijs. Mischa was in veel rollen en functies actief zoals docent, ontwikkelaar van het ICT-curriculum en de leerlijn data analyse, onderzoeker / projectleider en opleidingscoördinator. Hij was actief lid van het consortium voor hogere ICT-opleidingen, HBO-i, sprak op landelijke bijeenkomsten zoals het NIOC en maakte een studiereis naar Silicon Valley. Medio 2016 startte een periode van zes maanden kwartier maken voor het lectoraat Data Science. In die periode was hij lid van de kenniskring van het lectoraat Big Data van lector Gerard Schouten (Fontys Hogescholen), verzorgde lezingen / workshops (al dan niet als keynote speaker) bij onder meer het Big Data Event van Ki<MPi / Campione, de Toeristische ontmoetingsdag, het Trendcongres toerisme en het symposium Bouwen aan een veerkrachtige Scheldedelta. In maart 2017 werd Mischa aangesteld als lector Data Science aan de HZ. Tot overige activiteiten waar hij bij betrokken is of was horen het Center of Expertise Water & Energy, het Deltaplatform, de werkgroep Praktijkgericht onderzoek ICT. Bovendien gaat hij aan de slag als trekker van de werkgroep Data Science Lab ten behoeve van het Joint Research Center. Schrijven loopt als een rode draad door zijn nevenactiviteiten. Hij was redacteur van Chemistry Bytes en naast wetenschappelijke publicaties schreef hij tientallen artikelen voor onder meer het Chemisch Magazine en Laboratorium Magazine. Daarnaast werkte Mischa tussen 2007 en 2014 als freelance muziekjournalist voor onder andere de Provinciale Zeeuwse Courant en Jazzenzo.nl. En zelf muziek maken mag tot de grootste hobby worden gerekend, als gitarist in diverse bands.

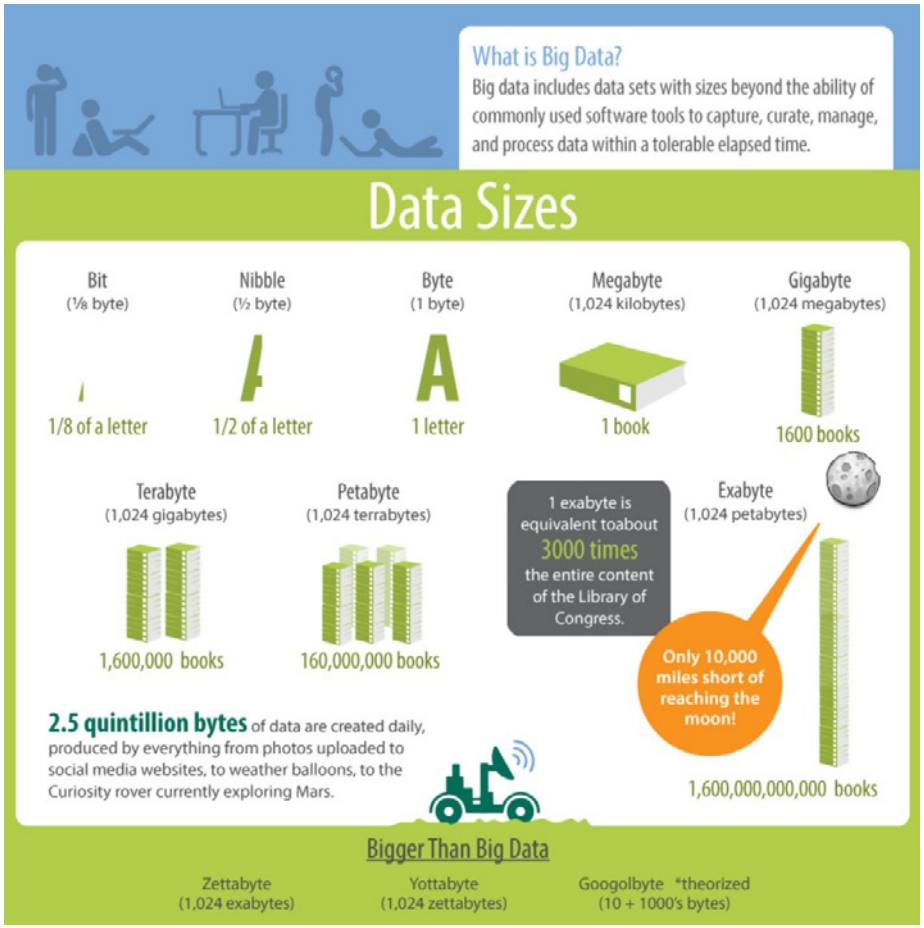
BIJLAGE 1
WAAROM AAN DE SLAG MET DATA SCIENCE?

Elk jaar brengt Gartner, een toonaangevend onderzoeks- en adviesbureau in de informatietechnologie-sector, de hype cycle voor emerging technologies uit. Die cycles geven houvast bij het selecteren van technologieën met innovatiepotentieel. Zonder in detail te treden over de interpretatie van de cycle is in de versie van 2017 te zien dat zowel Deep Learning als Machine Learning naar verwachting hun topniveau bereiken binnen 2 tot 5 jaar. Ofwel, binnen 2 tot 5 jaar als de technologie algemeen geaccepteerd is. Machine Learning (en daarmee Deep Learning) is een essentieel onderdeel van het Data Science process. Gartner hanteert voor Data Science en Machine Learning een eigen Hype Cycle (Gartner, Hype Cycle for Data Science and Machine Learning, 2017). Onder meer Predictive en Prescriptive Analytics bevinden zich op de peak, terwijl Text Analytics en Video/Image Analytics zich op de slope of enlightenment bevinden. Daarnaast zijn Big Data en Data Science hot in alle grote (jnternationale) frameworks, zoals Horizon 2020 (European Commission, 2017). Op nationaal niveau berekende de Nationale Denktank dat er in 2018 8.000 Data Scientists te kort zijn in Nederland (Denktank, 2014). Opleidingen kunnen nog niet aan de vraag voldoen. Ook de Human Capital Agenda ICT (Dutch Digital Delta) benadrukt het belang van Big Date en Data Science. In het actieplan voor ICT-uitdagingen in Nederland stelt Dutch Digital Delta dat voor de sleuteltechnologie ICT de Kennis- en Innovatieagenda (KIA) is vernieuwd. Twee van de vijf onderwerpen met een cross-sectorale impact die daarin centraal staan zijn Big Data en artificiële intelligentie (Delta, n.d.). Tenslotte bevat de Strategische onderzoeksagenda hbo 2016 – 2020 10 thema's waarbij praktijkgericht onderzoek van hogescholen verweven is met de vragen en routes van de Nationale Wetenschapsagenda (NWA). Het gaat met name om thema 4, Slimme technologie en Materialen die NWA route 24 behelst, namelijk Toegankelijke en verantwoorde waarde creatie uit Big Data (Vereniging Hogescholen, 2016) (Nationale Wetenschapsagenda, 2015).



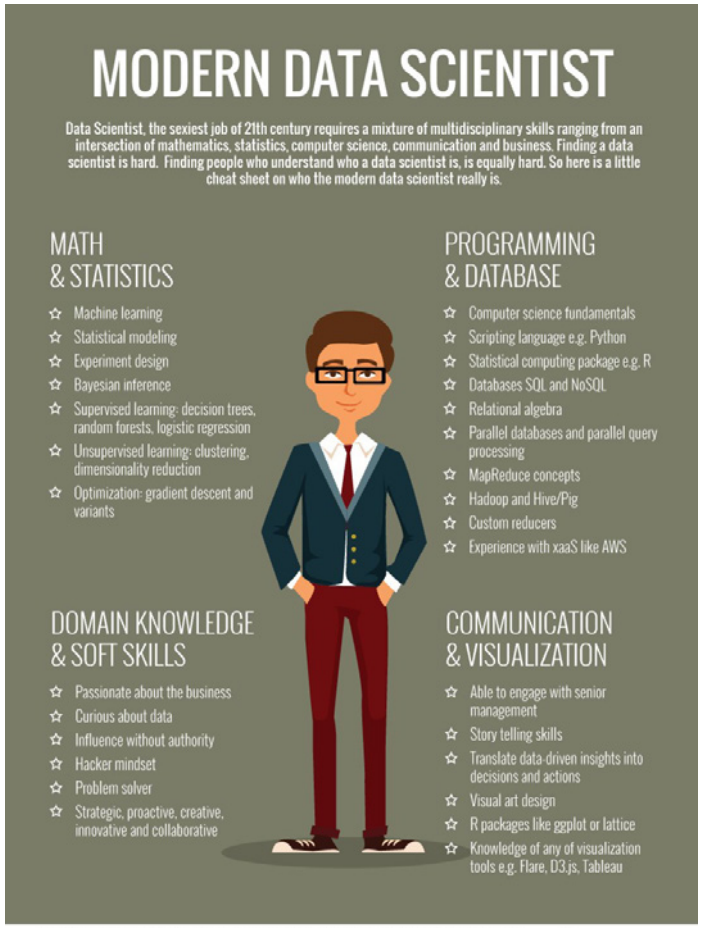
BIJLAGE 2
WAT IS VEEL DATA?

Overgenomen van (Bicorner, 2015), dat naast deze Infographic nog veel meer informatie over Big Data visueel maakt.



BIJLAGE 3
DATA SCIENCE DISCIPLINES EN VAARDIGHEDEN

Overgenomen van (Zawadzki , 2014). Er bestaat ook een vrouwelijke versie van deze set. Gelukkig staan daar precies dezelfde disciplines en vaardigheden op.



BIJLAGE 4
DATA SCIENCE LAB ZUID

Visie Notitie JRCZ - Data Science Lab Zuid
Datum :Juli 2017
Auteur : Dr. Ir. Robert Trouwborst (HZ)

De HZ, University College Roosevelt/University of Utrecht en ROC Scalda werken samen aan de oprichting van een Joint Research Center Zeeland (JRCZ). Deze experimentele onderwijs - en onderzoeksfaciliteit, van circa 4500m2 vanuit waar op drie onderwijsniveaus (mbo, hbo en wo) wordt gewerkt aan de human Capital agenda van Zeeland, wordt gebouwd op de locatie Groene Woud te Middelburg. Naast onderwijs - en onderzoekslaboratoria op de thema's Ecologie, Chemie, Fysica en Engineering wordt ook een state-of-the-art Data Science Lab ingericht.

Data is voor alle partners in de regio van onschatbare waarde. Het gaat zowel om verantwoordelijke organisaties in het beheer van de delta (overheden, nutsbedrijven), economische trekkers (Zeeland Seaports, Impuls Zeeland), de beheer - en veiligheidsketen (Rijkswaterstaat, waterschap, veiligheidsregio, gemeenten), als om kennisinstellingen, die werken aan innovatief onderzoek en het opleiden van nieuwe professionals (HZ, UCR, Scalda en partnerinstellingen).

De opgave is om een veelheid en diversiteit aan data om te kunnen te zetten in waardevolle producten en diensten die bijdragen aan een veilige en economisch aantrekkelijke delta. In vraagstukken op het gebied van waterveiligheid, ruimtelijke adaptatie, biodiversiteit en ecologische kwaliteit zijn grote hoeveelheden data beschikbaar. Hoewel de relevante databronnen bekend zijn, vormen zij voor iedere partij afzonderlijk een soms ondoorgrondelijk web. Dit kwam recent ook naar voren in overleg tussen veiligheidsregio, waterschap, Rijkswaterstaat, gemeenten en HZ over een informatierotonde en bijbehorende toepassingsmogelijkheden (HZ Vlissingen, 14 maart 2017). Ook tijdens het onlangs gehouden Digishape symposium (Informatiehuis Marien, 2017) werd helder dat een goede samenwerking nodig is om vanuit een diversiteit aan organisaties, talenten, belangen, mogelijkheden en opgaven de vruchten van de dataficering van de maatschappij te kunnen plukken.

Het Data Science Lab Zuid wil met de partijen vanuit de regio samenwerken aan de opbouw van nieuwe dataexpertise, de ontwikkeling van nieuwe data instrumenten en innovatieve analysestrategieën, analysetechnieken, dataontsluiting en processen (Big Data, datakwaliteit en veiligheid, Machine Learning etc.).

Beeld toekomst
De verwachting voor de komende decennia is dat de hoeveelheid en complexiteit van data exponentieel toeneemt. De uitdaging waar het netwerk voor staat is om deze complexiteit werkbaar te maken. Een belangrijke voorwaarde hiervoor is het realiseren van een ruimte waar vraag en aanbod elkaar niet alleen kunnen ontmoeten, maar waar via transdisciplinair onderzoek kan worden gewerkt aan de ontwikkeling van nieuwe toepassingen, diensten en producten. Vanuit lopend onderzoek op het gebied van deltatechnologie zijn er vele mogelijkheden voor nieuwe toepassingen door het gebruik van data science technieken.

Het lectoraat Data Science en de daaraan gelieerde opleiding HBO-ICT (Software Engineering) van de HZ beschikt over de kennis en kunde om data te ontsluiten, duiden en visualiseren. Een gezamenlijk Data Science Lab maakt het mogelijk vraag en aanbod in de regio samen te brengen. Het vormt de kernvoorziening in de ontwikkeling van een nieuwe 'ecosysteem' in de delta waarin bedrijfsleven, overheden en kennisinstellingen meerwaarde zien in het uitwisselen van data en gezamenlijk ontwikkelen van nieuwe praktische toepassingen. Waar geoefend en getest kan worden met tools voor de controle op datakwaliteit, real time data visualisatie in combinatie met real time datamanagement. Waar studenten in hackatons werken aan een praktijkcasus en samen met dataspecialisten vanuit een diversiteit aan databronnen zoeken naar nieuwe antwoorden en oplossingen.

Praktijkvoorbeelden
In de praktijk valt bijvoorbeeld te denken aan een project waarin studenten HBO-ICT van een 2D-waterbeeld een 3D-visualisatie maken, op basis waarvan studenten Water Management en Civiele Techniek nieuwe adaptatiemaatregelen kunnen ontwikkelen. Nieuwe technieken zoals Hololens bieden de mogelijkheid te laten zien wat er in een transformatorhuisje (als onderdeel vitale infrastructuur) gebeurt bij wateroverlast. Er ontstaan ook nieuwe vormen

van ontwerpen in deltagebieden, waarin bijvoorbeeld ecologische oplossingen, sociale veiligheid en effecten op gezondheid gecombineerd kunnen worden door inwoners te betrekken bij een realistisch 3D-ontwerp. Het ontsluiten van deltagebieden via 3D (of 4D, inclusief tijd) toepassingen zal ook een enorme impact hebben voor het duurzaam ontwerp en beheer van deze gebieden, bijvoorbeeld bij implementatie van de Omgevingswet.

We beogen om zo complexiteit werkbaar te maken, wat moet leiden tot een nieuw aanbod aan trainingen en opleidingen voor de (toekomstige) professionals op diverse niveaus.

Praktijkgericht en dus wendbaar/agile

Een gezamenlijk Data Science Lab heeft als insteek om snel (lean and agile) instrumenten te kunnen bouwen om snel resultaten te kunnen laten zien en innovaties te implementeren. Het draagt daarmee niet alleen bij aan betere product- en dienstverlening van bestaande deltatechnologie, maar heeft ook een aanjaagfunctie voor startups en ondernemersgeest in de regio.

BIJLAGE 5
WANNEER WERKEN DEEP NEURAL NETWORKS GOED?

Dit plaatje is gebaseerd op de uitleg die Andrew Ng geeft in de video "Why is Deep Learning taking off?", van de course "Introduction to Deep Learning" van Coursera (Ng, 2017).