Translationele datawetenschap in populatiegerichte zorg

Spruit, M.R.

Prof. Dr. Marco Spruit

Translationele Datawetenschap in Populatiegerichte Zorg

Universiteit Leiden

Bij ons leer je de wereld kennen

Translationele Datawetenschap in Populatiegerichte Zorg

Rede uitgesproken door

Prof. Dr. Marco Spruit

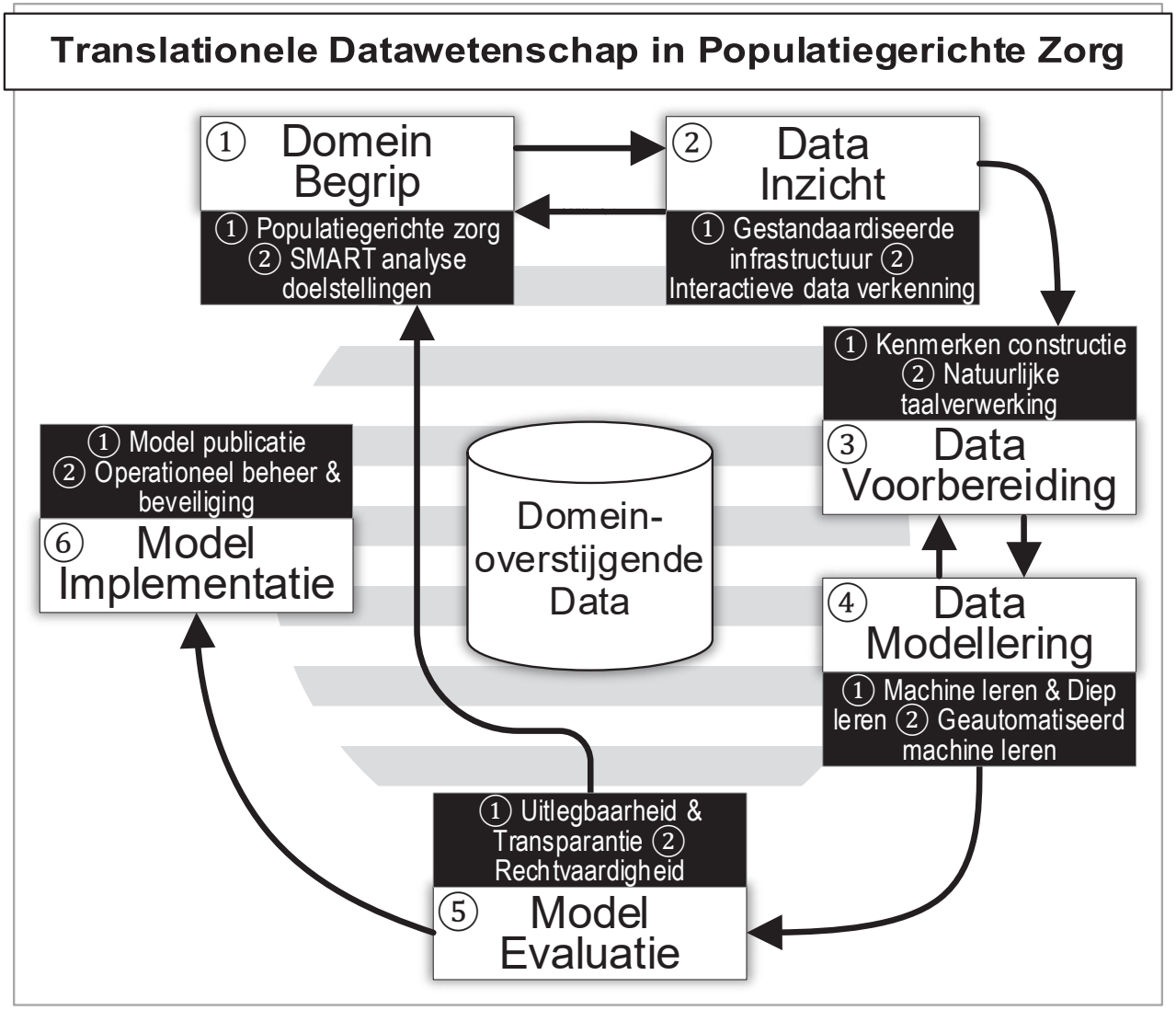
bij de aanvaarding van het ambt van hoogleraar met als leeropdracht

Advanced Data Science in Population Health

aan de Universiteit Leiden

op 1 april 2022

Universiteit Leiden


Translationele Datawetenschap in Populatiegerichte Zorg
1 Domein Begrip
1 Populatiegerichte zorg 2 SMART analyse doelstellingen
2 Data Inzicht
1 Gestandaardiseerde infrastructuur 2 Interactieve data verkenning
1 Kenmerken constructie 2 Natuurlijke taalverwerking
3 Data Voorbereiding
4 Data Modellering
1 Machine leren & Diep leren 2 Geautomatiseerd machine leren
Domein-overstijgende Data
1 Uitlegbaarheid & Transparantie 2 Rechtvaardigheid
5 Model Evaluatie
1 Model publicatie 2 Operationeel beheer & beveiliging
6 Model Implementatie

Mevrouw de Rector Magnificus, leden van de Raad van Bestuur van het Leids Universitair Medisch Centrum, leden van het bestuur van de Faculteit der Wiskunde en Natuurwetenschappen, beste collega's, lieve familie en vrienden, zeer gewaardeerde toehoorders.

Met deze openbare les, getiteld Translationele Datawetenschap in Populatiegerichte Zorg, aanvaard ik officieel mijn benoeming als hoogleraar bij zowel het Leids Universitair Medisch Centrum alsook de Faculteit der Wiskunde en Natuurwetenschappen van de Universiteit Leiden op de leerstoel *"Advanced Data Science in Population Health"*.

Translationele datawetenschap: het vakgebied
In de komende 45 minuten introduceer ik de translationele datawetenschap als zelfstandig vakgebied aan de Universiteit Leiden en in het Nederlandse wetenschapslandschap. Ik zal toelichten *waarom*, *hoe* en *wat*. De brede verhaallijn verloopt van conceptueel wetenschapsbeleid tot weerbarstige praktijkimplementatie.

Dit verhaal begint in 1945, met de zeergeleerde Vannevar Bush. In zijn functie als directeur van het Amerikaanse bureau voor wetenschappelijk onderzoek, schreef hij het zeer invloedrijke artikel met de fraaie titel *"Wetenschap: De eindeloze grens"*. Hiermee ontstond een standaard beleidsclassificatie omtrent de aard van elk wetenschappelijk werk, dat óf fundamenteel óf toegepast van aard zou zijn.[1]

Tot 25 jaar geleden. In 1997 publiceerde hooggeleerde Donald Stokes, indertijd decaan aan Princeton University, een alternatief classificatieinstrument voor wetenschappelijk onderzoek: Pasteur's Kwadrant.[2] Pasteur's Kwadrant is een beleidsmodel in de vorm van een vierkant met twee rijen en twee kolommen, dat wetenschappelijk onderzoek classificeert op basis van twéé dimensies in plaats van slechts één. De verticale as betreft de reeds gebruikelijke "*mate van fundamenteel begrip*". De horizontale as toont de "*mate van praktische gebruiksoverwegingen*". Stokes ontwikkelde Pasteur's Kwadrant om de aloude frictie tussen fundamenteel en toegepast onderzoek te nuanceren.

Wanneer we Pasteur's Kwadrant van wetenschappelijke onderzoeksbenaderingen toepassen op het brede scala van aangrenzende datawetenschapsdisciplines, dan komen de volgende drie deeldisciplines van datawetenschap naar voren. Allereerst, linksboven in het kwadrant, vinden we de zuiver-fundamentele datawetenschap, waarin men voornamelijk nieuwe algoritmes en datamethoden ontwikkelt, om almaar béter datawetenschap te kunnen doen, met almaar betere analyse instrumenten. In de woorden van Bush: *"zonder te denken aan praktische toepassingen"*.[2] Echter, het blijkt dat resultaten uit dit type wetenschappelijk onderzoek veelal niet vanzelf doorsijpelen naar de samenleving. Het overgrote deel van dit type onderzoek eindigt in de vergetelheid...[1] Rechtsonder in dit kwadrant van onderzoeksbenaderingen bevindt zich de zuiver-toegepaste datawetenschap, waarin men een maatschappelijke vraaggestuurde aanpak hanteert, en waarin men beoogt reeds op kortere termijn een concrete bijdrage te kunnen leveren aan innovatie en de aanpak van maatschappelijke vraagstukken.

Tenslotte bevindt zich rechtsboven in Pasteur's Kwadrant het vakgebied van de toepassingsgericht-fundamentele datawetenschap. Oftewel, *translationele* datawetenschap. De translationele datawetenschapsbenadering is enerzijds, net als de zuiver-fundamentele benadering, op zoek naar een beter fundamenteel begrip van de wereld om ons heen. Anderzijds is het, net als de zuiver-toegepaste benadering, maatschappelijk geïnspireerd, vraaggestuurd en oplossingsgericht. De translationele datawetenschap biedt kortom het beste van beide werelden! Translationele datawetenschap is overigens geen nieuwe deeldiscipline binnen het spectrum van datawetenschap disciplines. De eerste academische conferentie die geheel werd gewijd aan precies dit onderwerp, vond vijf jaar geleden plaats in Chicago. Aldaar werd translationele datawetenschap – vrij vertaald – gedefinieerd als: *"[...] een vakgebied dat datawetenschappelijke principes, technieken en technologieën toepast op*

problemen in andere vakgebieden [...] en waarvan de uitkomsten tevens ons fundamentele begrip van de datawetenschap verbreedt of verdiept".[3]

Tenslotte wordt translationeel onderzoek traditiegetrouw onderverdeeld in twee categorieën.[4] Toegepast op het onderwerp van deze rede, richt "T1" translationeel datawetenschappelijk onderzoek zich primair op de effectieve vertaling van nieuwe algoritmes, data-modellen en technieken naar populatiegerichte zorg. Voor de datawetenschap resulteert dit in een beter begrip van de exacte werking van de techniek. Voor de populatiegerichte zorg biedt het inzicht in nieuwe interventiemogelijkheden voor preventie, diagnose en behandeling van ziekten, en voor de verbetering van de gezondheid. "T2" translationeel datawetenschappelijk onderzoek richt zich voornamelijk op het vertalen van de in T1 opgedane kennis naar de dagelijkse praktijk via het zorgvuldig toetsen van nieuwe behandelingen en onderzoekskennis op de juiste patiënten of bevolkingsgroepen.

T1, T2... Translationele geneeskunde als analogie
De meest bekende tak van de translationele wetenschappen is de translationele geneeskunde. Deze wordt kernachtig beschreven met het motto *"van bank tot bed"*. In deze tak van de geneeskunde vertaalt men bijvoorbeeld nieuwe kennis uit het onderzoekslaboratorium over de werking van het COVID-19 virus ("T1") naar een Janssen-vaccin in de dagelijkse zorgpraktijk van huisartsen en de GGD, om de gezondheidsbescherming te bevorderen bij de juiste bevolkingsgroepen ("T2"). Dankzij het T2 onderzoek weten we bijvoorbeeld nu dat zwangere vrouwen dit Janssen vaccin niet zouden moeten ontvangen.

Translationele datawetenschap: het proces
Tot zover de wetenschappelijke theorie. U weet nu *wat* translationele datawetenschap is, en *waar* dit vakgebied aansluit op het wetenschappelijke veld. U begrijpt tevens *waarom* het integreren van fundamenteel en toegepast onderzoek zowel meerwaarde biedt voor de datawetenschap als voor de maatschappij. Nu zal ik kort toelichten *hoe* het standaard proces voor translationele datawetenschap uitgevoerd wordt. Dat doe ik aan de hand van de aloude en reeds langbeproefde wetenschappelijke methode, in vier stappen. Vervolgens licht ik het translationele datawetenschapsproces toe als uitbreiding op deze wetenschappelijke methode.

In Stap 1 van de wetenschappelijke methode formuleert u een nieuwe onderzoeksvraag. Het vereist *domein begrip* van een vakexpert om überhaupt tot een relevante onderzoeksvraag te kunnen komen. In Stap 2 stelt u een bijpassende onderzoeksmethodologie op — zoals bijvoorbeeld een computationeel experiment of een klinische studie — waarmee u deze vraag op betrouwbare en reproduceerbare wijze kan beantwoorden. In Stap 3 verzamelt u alle relevante data die u hiervoor nodig heeft. In Stap 4 analyseert u tenslotte de data uit Stap 3, met behulp van het in Stap 2 opgestelde onderzoeksplan, om een betrouwbaar antwoord op de onderzoeksvraag uit Stap 1 te verkrijgen. Vaak roept de nieuw verkregen kennis nieuwe vervolgvragen op, waardoor u weer teruggaat naar Stap 1, om op basis van het zojuist vergrote *domein begrip* een nieuwe relevante onderzoeksvraag te formuleren. Enzovoort. Dit is in een notendop de wetenschappelijke methode. Het werkt!

Wanneer we nu inzoomen van wetenschappelijk onderzoek in het algemeen naar *data*wetenschappelijk onderzoek in het bijzonder, dan komt er nog één dimensie bij, namelijk die van de *techniek*. Met techniek bedoel ik hier de technologieën en architecturen om de data-analyse tevens te implementeren in software scripts of analytische toepassingen.[5] Daarnaast is de gereedschapskist voor data-analyse uitgebreid met technieken uit zowel het machine leren als de statistiek. Met deze twee uitbreidingen op de wetenschappelijke methode kunnen we het cyclische proces van translationele datawetenschap gaan beschrijven in zes gestandaardiseerde fases.

Datawetenschappers in zowel het bedrijfsleven als de wetenschap gebruiken reeds sinds enkele decennia het zogenaamde "sectoroverschrijdende standaardproces voor data-analyse" (CRISP-DM).[6] Dit proces biedt een meerlagig en cyclisch stappenplan om iedere analyse-taak in iedere fase van het proces op een gestandaardiseerde en reeds beproefde wijze uit te voeren. Dit standaardproces blijkt uitermate goed aan te sluiten op het type wetenschappelijk onderzoek dat we met translationele datawetenschap nastreven. Zo wordt ieder onderzoek geïnitieerd vanuit *"praktische gebruiksoverwegingen"* die via probleemdefinities vertaald worden naar specifieke, meetbare, acceptabele, realistische en tijdgebonden (SMART) data-analyse doelen. Het "*fundamenteel begrip*" van de data-analyse technieken zélf wordt onder meer vergroot dankzij een kwalitatieve *model evaluatie* met domein experts. Zo kan een op zich zeer goed presterend voorspelmodel alsnog worden afgekeurd voor gebruik in de dagelijkse zorgpraktijk, vanwege een gebrek aan transparantie, ethische borging of juridische naleving.

Met deze korte inleiding van het proces voor het uitvoeren van translationele datawetenschap heb ik nu alle puzzelstukken afdoende geïntroduceerd. U weet nu *wat* het is, *waarom* het belangrijk is, en *hoe* je het doet. In het nog resterende deel van deze openbare les zal ik een stap-voor-stap beschrijving geven van de zes fases van het "sectoroverschrijdende standaardproces voor data-analyse", waarbij ik voor iedere fase tenminste twee onderzoeksonderwerpen zal aanstippen, tezamen met korte praktijkvoorbeelden.

Fase 1: Domein Begrip

De eerste van de zes fases in het "sectoroverschrijdende standaardproces voor data-analyse" is het expliciteren van het *domein begrip*. We willen namelijk te werk gaan vanuit *"praktische gebruiksoverwegingen"*. Daarom is het cruciaal om allereerst zorgvuldig de beoogde projectdoelstellingen van de belanghebbenden uit te vragen, en hiervoor tevens succescriteria op te stellen. Op basis daarvan kan de haalbaarheid van de doelstellingen ingeschat worden, gegeven de risico's, beperkingen, beschikbare tijd, geld en menskracht. Daarna vertalen we de doelen naar een technische data-analyse probleemdefinitie, met meetbare prestatie-indicatoren. Kortom, maak een plan!

1.1 Populatiegerichte zorg

Mijn leeropdracht richt zich primair op het toepassingsdomein van de *populatiegerichte zorg* en de data-gedreven sturing ervan. Deze interdisciplinaire benadering is een van de acht strategische focusgebieden van het LUMC. Op de interdisciplinaire Health Campus Den Haag is onze centrale missie om vanuit verschillende wetenschappelijke en klinische invalshoeken, en in samenwerking met regionale partners, bij te dragen aan het langer leven in goede gezondheid voor iedereen. Daarbij staan in onze benadering drie principes centraal: het verkleinen van gezondheidsverschillen, een duurzame benadering en een breed gezondheidsperspectief.[7]

Populatiegerichte gezondheidsondersteuning hanteert een data-gedreven en innovatiegedreven focus op preventie en leefstijl. Uitgangspunt is dat routinematig geregistreerde data over gezondheid, zorg en het sociaal domein hergebruikt worden, voor zover dit op zowel technische als juridisch-acceptabele wijze mogelijk is. Overigens betreft het vaststellen van de juiste doelstellingen veelal een combinatie van zowel kwalitatieve als kwantitatieve databronnen. Door middel van geavanceerde data-analysetechnieken vanuit de datawetenschap en artificiele intelligentie (AI) kunnen we met deze databronnen onder meer risicogroepen identificeren en de juiste zorg op de juiste plek leveren, door medische routine data te koppelen aan gegevens van het sociaal en publiek domein zoals leefomgeving en schuldenproblematiek. Op de interdisciplinaire Health Campus Den Haag werken wij samen in een zogenaamde viervoudige helix van innovatie[8] vanuit ons "Levende Laboratorium voor Populatiegerichte Zorg" in de regio Haaglanden.

1.2 SMART doelstellingen

Mede dankzij de duurzame samenwerking met bedrijven, overheden en burgers in onze viervoudige helix kunnen we veelal met vertrouwen komen tot de juiste informatiebehoeften. Wanneer het echter om enige reden niet goed mogelijk blijkt om de eigen informatiebehoeften te formuleren, dan behoort het tevens tot de mogelijkheden om doelstellingen semiautomatisch te destilleren uit reeds aanwezige organisatiegegevens zoals interne documenten, vergaderingsverslagen en e-mail conversaties.[9] De volgende stap is om deze informatiebehoeften op basis van de beschikbare databronnen te vertalen naar uitvoerbare data-analyse doelstellingen met minimaal te behalen prestatiemetingen.[10] Hiermee is een SMART fundering voor de data-analyse gelegd.

Fase 2: Data Inzicht

De tweede van de zes fases in het "sectoroverschrijdende standaardproces voor data-analyse" betreft *data inzicht*. In deze fase wordt de benodigde data eerst verzameld en beschreven, zoals het aantal beschikbare datapunten en de aanwezige data types. Vanzelfsprekend levert een beschrijving van de data op zich nog niet voldoende data inzicht op; daarvoor is nog een interactieve verkenning nodig van de data zelf, om eigenhandig de verscheidenheid van de voorkomende waarden in de data te observeren. Hieruit volgt tevens een beoordeling van de datakwaliteit, want deze bepaalt de benodigde datavoorbereidende werkzaamheden, alsook de kwaliteit van de analyse-uitkomsten.

2.1 Gestandaardiseerde infrastructuur

Het bijeenbrengen van de benodigde databronnen is meer dan ooit een vak apart. Van oudsher bestaat het dataverzamelings- en beheerproces veelal uit het extraheren en transformeren van operationele data uit databanken, eventueel verrijkt met publiekelijk beschikbare databronnen van bijvoorbeeld het Centraal Bureau voor de Statistiek (CBS). Afhankelijk van de mate van aanwezige structuur in de verzamelde data en de gewenste data inzichten, kan de infrastructuur worden vormgegeven als een data*bank* voor numerieke tabellen, een data*meer* voor teksten en beelden, of een data*stroom* voor voortdurend bijgewerkte berichten.[11]

Echter, in een populatiegerichte zorgcontext waarin verscheidene partners samenwerken in een viervoudige helix, is een *inter*organisationele infrastructuur vereist om voorbij de grenzen van individuele organisaties data te kunnen delen. Routinematig geregistreerde data over gezondheid, zorg en het sociaal domein worden dan op technisch, ethisch en juridisch acceptabele wijze aan elkaar gekoppeld. Een dergelijke data-infrastructuur kan verrijkt inzicht bieden in de gezondheid, kosten en ervaringen van de inwoners van de regio en biedt tevens aanknopingspunten voor populatie-gestuurde gezondheidsondersteuning voor zowel zorgsysteem spelers, organisaties, professionals als inwoners.

ELAN, het Extramuraal LUMC Academisch Netwerk, is zo'n interdisciplinair domein-overstijgend regionaal dataplatform, inclusief bijbehorende beleidsstructuur, dat gezondheidsbeleid en onderzoek ondersteunt. ELAN is gerealiseerd als toegang-op-afstand omgeving bij het CBS. Voor de regio Haaglanden koppelt ELAN reeds gestructureerde medische data, GGZ-informatie, sociaal domein context, en publieke gezondheid van honderdduizenden inwoners. Maar, dankzij de unieke status van het CBS zoals vastgelegd in de CBS-wet uit 2003, kan de ELAN-data verder verrijkt worden met sociaaleconomische gegevens zoals inkomen, opleiding, werkstatus, huishoudsamenstelling en wijk- en buurtgegevens. Op *individueel* niveau. Desalniettemin wordt voldaan aan de privacybepalingen in de Algemene Verordening Gegevensbescherming (AVG).

Federatief leren

Ons vorige kabinet heeft nu precies een jaar geleden besloten om het voorstel voor een nationale gezondheidsdata-infrastructuur te honoreren (Health-RI). Met de investering van 69 miljoen euro uit het Nationaal Groeifonds is men nu begonnen om gezondheidsdata beter toegankelijk te maken voor gezond-

heidsonderzoek en -innovatie. Een belangrijke vraag hier is hoe we de verschillende reeds bestaande architecturen, zoals ELAN, met elkaar kunnen koppelen. Want ook in de toekomst zullen er vele data-infrastructuren naast elkaar blijven bestaan, omdat verschillende analyse-taken nu eenmaal verschillende data-oplossingen vergen. Of vrij vertaald naar hooggeleerde Maslov: *"Om te voorkomen dat wanneer je alleen een hamer hebt, alles op een spijker lijkt"*.

Hoe verhoudt dit nationale plan zich tot ELAN? Concreet ondersteunt ELAN's toegang-op-afstand omgeving bij het CBS geen tekst -en beelddata. Met name is het bekend dat de vele tekstverslagen en klinische notities die artsen en verplegers schrijven over hun patiënten in het Elektronisch Patiëntendossier, waardevolle inzichten bieden over de gezondheid en ervaringen van hun cliënten. Om toch veilig tekstdata te kunnen delen, zijn daarom in recente jaren nieuwe technologieën ontwikkeld: federatief leren en gedistribueerd rekenen.

Inherente discretie

Deze nog jonge vakgebieden van federatief leren en gedistribueerd rekenen bieden een interessante oplossingsrichting voor het interorganisationeel koppelen van ongestructureerde en multimodale data. Hiermee kunnen op een veilige manier patronen geleerd worden van gevoelige data uit meerdere bronnen, zonder deze data te hoeven delen. Bij federatief leren wordt de analysetaak naar de data gebracht in plaats van andersom, zodat de data niet verplaatst of gedeeld hoeft te worden, enkel de software-code voor de analysetaak: *"Computing Visits Data"* (CoViDa).[12] Gedistribueerd rekenen maakt daarentegen gebruik van cryptografie, zodat analyses uitgevoerd worden op een versleutelde versie van de data. Ons fundamentele recht op discretie ('privacy'), zoals ook vastgelegd in de AVG, blijft inherent geborgd. Dit komt het vertrouwen in dergelijke technologie ten goede, en vereenvoudigt het beschikbaar stellen van data, vanwege de inherente garantie van discretie.

Ook recentelijk in zwang gekomen, is het inherent discrete gebruik van *synthetische* data in data-analyses, waarmee gegarandeerd veilig met gevoelige data gewerkt kan worden. Een *synthetische* dataset is een volledig nieuw gegenereerde dataset op basis van de daadwerkelijke brondata, die de oorspronkelijke kenmerken, relaties en statistische patronen van de oorspronkelijke gegevens behoudt. Het gaat anders gezegd om een kloon, een digitale tweeling. Enerzijds resulteert dit in minimale risico's voor en maximaal behoud van discretie, anderzijds is de kwaliteit en levertijd van de synthetische data beter en sneller dan die van het origineel. Een win-win situatie, dus. In het LUMC leid ik op dit moment een initiatief om een synthetische kloon van de unieke ELAN-dataset te construeren, ten behoeve van realistischer en inspirerender data-analyse onderwijs in onder meer de opleidingen Geneeskunde en Populatiegezondheidsbeheer.

2.2 Interactieve data verkenning

Wanneer er eenmaal een geïntegreerde data-infrastructuur is, met de juiste technologieën voor het met vertrouwen toegankelijk maken van data, zoals zojuist geschetst, zijn we nu klaar om naar de juiste data te gaan kijken. De eerste exploratieve bevindingen over de data-analyse doelstellingen helpen om een beter inzicht in de precieze aard van de data te krijgen. Bijvoorbeeld met betrekking tot de datakwaliteit. Hoe compleet is de data, zijn er veel missende waarden, is de spreiding van de mogelijke data-waarden uit balans? Enzovoorts.

Bij een interactieve data-verkenning zijn twee aspecten met name van belang. Allereerst het gebruik van visualisatietechnieken om data-karakteristieken inzichtelijker te maken. Bijvoorbeeld histogrammen om de frequentieverdeling te visualiseren, of een parenplot voor het relateren van paarsgewijze relaties. Uit onderzoek blijkt bovendien dat interactieve grafieken de duiding van de data verhogen.[13] Denk bijvoorbeeld aan mogelijkheden om filters op de getoonde data te specificeren, of door middel van deelgebied-selectie met de computermuis verder in te kunnen zoomen.

Het andere belangrijke aspect bij een interactieve data-verkenning is om deze uit te voeren in multidisciplinair samenwerkende teams met minimaal één datawetenschapper en één domein expert. Uit de praktijk blijkt namelijk dat nieuwe visualisaties door de datawetenschapper inderdaad leiden tot nieuwe hypotheses bij de domein expert, omdat de directe visualisaties de creativiteit stimuleren. De methodologische consequenties hiervan zijn groot, want een dergelijke benadering richt zich op het vinden van nieuwe kennis en onverwachte hypotheses uit data, in plaats van het simpelweg uitvoeren van de zorgvuldig gedefinieerde data-analyse taken uit Fase 1.[14] De strekking van deze bevinding vinden we ook terug in het bekende Afrikaanse spreekwoord: *"Alleen ga je sneller, maar samen komen we verder"*.

Fase 3: Data Voorbereiding

Nu zijn we toegekomen aan de derde van de zes fases in het "sectoroverschrijdende standaardproces voor data-analyse": *data voorbereiding*. Het is in deze fase waar de noeste arbeid wordt verricht. Men schat dat 80 tot 90% van al het uitvoerende datawetenschapswerk plaats vindt in deze data-voorbereidingsfase. Allereerst selecteren we enkel de te analyseren data, want helaas blijkt het marketing mantra *"hoe meer data hoe meer inzicht"* niet op te gaan in de weerbarstige praktijk. *"Hoe meer data hoe meer ruis"* zou een accurater mantra zijn. Daarna is het vaak nodig om de datakwaliteit te verbeteren tot een niveau dat minimaal verondersteld wordt bij de gekozen analysetechnieken. Hoe gaan we om met missende waarden? Bijvoorbeeld door ze uit de dataselectie te houden, of voegen we geschikte standaardwaarden in, of kunnen we ontbrekende data via een model betrouwbaar genoeg schatten? Het hoofddoel in deze fase is om alle benodigde voorbereidingen te treffen om de hiernavolgende data-analyse zo goed mogelijk uit te kunnen voeren.

3.1 Kenmerken constructie

De data voorbereidingsfase is hét moment om nieuwe, onderscheidende kenmerken aan te maken op basis van reeds beschikbare data. Dit kan variëren van het afleiden van de provincienaam aan de hand van een plaatsnaam, tot het automatisch extraheren van de huidige rookstatus van een patiënt op basis van de journaalteksten die de huisarts tijdens consultaties als vrij geformuleerde tekst in het elektronisch patiëntendossier heeft bijgeschreven. Inclusief typefouten, halve zinnen en allerlei afkortingen.

Ook zijn allerlei datatransformaties noodzakelijk, bijvoorbeeld om het datatype van een data-attribuut te converteren, zodat de gewenste data-analysetechniek toegepast kan worden. De rookstatus "Niet-roker" kan worden gerepresenteerd met een 0, de rookstatus "Roker" met een 1, "Gestopte roker" met een 2. Op vergelijkbare wijze kunnen ook data-kenmerken worden samengevoegd, zoals "Voornaam" en "Achternaam" tot "Volledige naam". Enzovoorts. Kortom, u begrijpt inmiddels waarom deze vaak tijdrovende Fase 3 van data voorbereiding ook wel smalend data *goochelen* wordt genoemd!

Data standaardisatie

Hoe dan ook, uiteindelijk beschikken we over de te analyseren data. Opgeschoond, verrijkt, gecorrigeerd. Maar nóg is deze fase niet volledig. Duurzame datawetenschap vereist namelijk dat de data tevens FAIR is: vindbaar ('findable'), toegankelijk ('accessible'), interoperabel ('interoperable') en herbruikbaar ('reusable'). Dit is vanzelfsprekend een nobel streven, maar vergt tevens vooraf extra aandacht voor data standaardisatie aspecten. Hoe zorgen we er bijvoorbeeld voor dat de data enerzijds toegankelijk is voor hergebruik, zonder dat de AVG-wetgeving overtreden wordt? Een deelantwoord hierop kan federatief leren zijn, maar alleen wanneer de partijen dezelfde taal spreken.

Noemenswaardig is hier het meertalige medisch terminologiestelsel van en voor zorgprofessionals (SNOMED). Deze ontologie bevat een enorme verzameling van circa 370.000 medische begrippen en hun synoniemen. Zorgverleners kunnen deze begrippen gebruiken om allerlei zorginformatie eenduidig vast te

leggen, zoals klachten, symptomen, en diagnoses. Gegevens die met deze ontologie zijn vastgelegd, zijn daarom zeer geschikt voor uitwisseling en hergebruik.[15] Daarnaast is voor het digitaal uitwisselen van zorgdata binnen en tussen zorginstellingen ook een "vurige" zorguitwisselingstaal ontwikkeld (FHIR). En is er zelfs een universele uitwisselingstaal voor datamodel definities (ONNX), zodat voorspelmodellen in ieder systeem kunnen worden hergebruikt. Tenslotte is er voor het tevens grootschalig analyseren van de uiteindelijke medische uitkomsten inmiddels ook een heuse "Odyssee" gestart, naar analogie met het epos van Homerus.[16]

Het ontwikkelen van een domein-overstijgende data infrastructuur, zoals ik in Fase 2 reeds toelichtte, vereist daarom een hoge mate van data standaardisatie om op duurzame wijze populatiegerichte zorg en onderzoek te kunnen faciliteren.

3.2 Natuurlijke taalverwerking

Ik benoemde zojuist als een van de data-voorbereidingsstappen de mogelijkheid om de huidige rookstatus van een patiënt automatisch uit een tekstveld binnen het elektronisch patiëntendossier te extraheren. Hier wil ik graag dieper op in gaan. Dit is namelijk een relevant voorbeeld van natuurlijke taalverwerking binnen de datawetenschap. Natuurlijke taalverwerking is overigens *ook* een prominent deelgebied binnen het vakgebied van de Artificiële Intelligentie (AI). In AI is het uiteindelijke doel is om machines menselijk gedrag te laten vertonen. Maar vanuit mijn datawetenschappelijke perspectief zijn natuurlijke taalverwerkingstechnieken vooral interessant om de inhoud van ongestructureerde teksten te begrijpen, met als doel om de reeds beschikbare gestructureerde data verder te verrijken, zodat we tot de meest complete data-analyse inzichten kunnen komen.

Toen ikzelf in het begin van de jaren '90 computationele taalkunde studeerde aan de Universiteit van Amsterdam, bestond de opleiding voornamelijk uit het bestuderen en programmeren van vele taalkundige theorieën, in de vorm van zogenaamde herschrijfregels. De zeer eenvoudige Nederlandstalige zin *"Mijn dochter luistert"* wordt bijvoorbeeld herschreven als een opeenvolging van een naamwoordelijk deel *"Mijn dochter"* plus een werkwoordelijk deel *"luistert"*. Het naamwoordelijke deel *"mijn dochter"* is op zijn beurt weer opgebouwd uit een bezittelijk voornaamwoord en een zelfstandig naamwoord. Deze zinstructuur van herschrijfregels kunnen we tevens visualiseren als een hiërarchische boomstructuur met vertakkingen, waarin de blaadjes de daadwerkelijke woorden van de zin zijn. Dit klinkt allemaal heel intuïtief, logisch en taaltheoretisch onderbouwd. Desondanks bleek gedurende ieder vak opnieuw dat deze ingenieuze grammatica's van herschrijfregels niet voldoende de weerbarstige taal van alledag konden representeren. Heel frustrerend was dat. Onze taal blijkt simpelweg te dynamisch, te onvoorspelbaar om computers op ónze manier te laten begrijpen.

Eén vak gedurende mijn opleiding volgde echter een compleet andere benadering, namelijk die van de stochastische taalmodellen. In plaats van logische herschrijfregels wordt in deze benadering de grammaticaliteit van zinnen berekend aan de hand van kansberekeningen. Het is wiskunde met taal! Geen herschrijfregels voor semantiek, maar hoogdimensionale vectorruimtes, in lijn met de zogenaamde *distributiehypothese*.[17] Deze distributiehypothese stelt dat de betekenis van een woord voortvloeit uit de manier waarop het woord is ingebed of "gedistribueerd" in ons dagelijks taalgebruik. Wanneer een woord geprojecteerd wordt als een vector in een hoogdimensionale ruimte, dan representeren de nabijgelegen vectoren in deze ruimte de specifieke context, en daarmee ook diens betekenis. Deze alternatieve, bijna-magische benadering voor natuurlijke taalverwerking greep mij, en ik studeerde in 1995 dan ook af op een zelflerend, artificieel neuraal netwerk voor het selectief filteren van informatie in tekstdatastromen.[18] Ik ben zeer vereerd dat mijn begeleider van toen, hooggeleerde Scholtes, ook vandaag hier aanwezig is, 27 jaar na dato!

Het afgelopen decennium is deze stochastische benadering van natuurlijke taalverwerking allesoverheersend geworden in zowel de wetenschappelijke theorievorming als in praktijkimplementaties. De vooruitgang die de afgelopen 30 jaar geboekt is, heeft eerlijk gezegd mijn stoutste verwachtingen overtroffen. Desondanks zijn we er nog niet. Computers begrijpen namelijk nog steeds niets van onze natuurlijke taal. We zijn er echter desondanks al wel redelijk in geslaagd om computers eenvoudige gesprekjes met ons te laten voeren in de trant van *"Oké Google... waar is het Academiegebouw?"*.

De laatste jaren is echter het besef ontstaan dat de stochastische benadering van natuurlijke taalverwerking geïntegreerd zou moeten worden met de aloude logische benadering. Om gedistribueerde woordinbeddingen te integreren met symbolische herschrijfregels. Dit is overigens geen nieuw idee. Al sinds de jaren '80 proberen onderzoekers ons eigen menselijk taalvermogen te begrijpen door de subsymbolische, stochastische en de symbolische, logische benaderingen voor natuurlijke taalverwerking met elkaar te verbinden.[19] Ik wil de komende jaren graag bijdragen aan het integreren van deze twee schijnbaar complementaire taalverwerkingsbenaderingen, vanuit mijn toepassingsgerichte doel van verrijkt data inzicht, ten dienste van betekenisvollere data-analyse uitkomsten.

Concreet betekent dit dat ik onder meer interesse heb in natuurlijke taalverwerkingstaken zoals open informatie extractie, onderwerp modellering en taalmarkering detectie. Zo is een belangrijke zorgtoepassing van informatie extractie op de klinische notities in elektronische patiëntendossiers het deïdentificeren van alle naar patiënt herleidbare gegevens, zoals naam, adres, telefoonnummer, geboortedatum, enzovoorts.[20] Dit is niet alleen ethisch zeer wenselijk, maar ook wettelijk verplicht volgens de AVG. Onderwerp modellering is een populaire taalverwerkingstechniek om automatisch latente (verborgen) onderwerpen uit bijvoorbeeld klinische notities te extraheren. Een patiëntenverslag kan bijvoorbeeld 30% over een specifieke aandoening gaan, en 70% over angstgevoelens bij de patiënt.[21]

Tenslotte heeft ook mijn bijzondere aandacht het detecteren van iemands mentale gesteldheid op basis van de wijze waarop deze zijn verhaal doet, door onder meer de woordkeus, zinsbouw en prosodie te analyseren. Iemands taalgebruik reflecteert namelijk iemands identiteit. Mijn aanname is dat er inderdaad taalmarkeringen bestaan, vergelijkbaar met biomarkeringen, die de aanwezigheid van iemands biologische eigenschappen aanduiden. Via taalmarkeringen in iemands taaluitingen, zouden we dan inzicht kunnen krijgen in iemands mentale gesteldheid. Wanneer iemand bijvoorbeeld de eerste persoon enkelvoudsvorm "ik" overmatig hanteert in zijn communicatie, zou dit in combinatie met overmatig woordgebruik rondom thuis en beweging, een taalmarkering kunnen zijn voor autisme. Het zijn dit soort fascinerende onderzoeksvragen die momenteel mijn grote aandacht hebben voor natuurlijke taalverwerkingstechnieken binnen het zorg en welzijnsdomein.[22]

Fase 4: Data Modellering

Na deze drie fases met elk drie deelonderwerpen, zijn we nu eindelijk beland in de vierde van de zes fases in het "sectoroverschrijdende standaardproces voor data-analyse": *data modellering*. Het is in deze fase dat de ware datawetenschapper schittert! De data-analyse doelstellingen werden reeds in Fase 1 opgesteld. Nu is het moment gekomen om allereerst de meest geschikte algoritmes voor iedere doelstelling te selecteren. Dit is nog niet zo eenvoudig aangezien er vele honderden zo niet duizenden unieke algoritmes bestaan.

Het selecteren van de juiste modelleertechniek en algoritme start met het bepalen van het type taak van de gewenste data-analyse. Denk bijvoorbeeld aan het voorspellen van iemands gewicht in kilogrammen (Y) op basis van enkel de lengte in centimeters van die persoon (X). Hier is het doel om een numerieke waarde te voorspellen. Een andere veelvoorkomende taak is classificatie, waarin we de juiste categorie willen voorspellen. Denk dan bijvoorbeeld niet aan iemands gewicht in kilogrammen als uitkomst, maar aan het indelen in vooraf

bepaalde categorieën: ondergewicht, gezond gewicht, of overgewicht.

Behalve regressie en classificatie omvat de datawetenschap ook vele andere analyse-taken zoals clusters vormen, associatieregels opstellen, uitzonderingen detecteren en tijdreeksen voorspellen. Daarnaast kunnen data-analyse taken beschrijvend, voorspellend of voorschrijvend van aard zijn. De datawetenschap speelt zich af rondom een zeer rijke en veelomvattende gereedschapskist!

4.1 Machine leren

In de datawetenschap bestuderen en gebruiken we met name *data modellering* technieken die behoren tot het vakgebied machine leren. Machine leren is kort gezegd een techniek om software en algoritmes zich *autonoom* te laten verbeteren door het analyseren en herkennen van patronen in data. Veelal met als doel om accurate voorspellingen te kunnen doen, zoals het voorspellen van iemands gewicht in kilogrammen op basis van diens lengte, leeftijd en buikomtrek.

Machine leren lijkt op, maar is niet hetzelfde als, statistiek. Daar waar machine leren zich vooral toelegt op het blootleggen van generaliseerbare en accurate voorspellende patronen, probeert men in de statistiek bovenal populatie-brede conclusies uit een steekproef af te leiden.[23] Daarnaast kennen deze twee aanpalende disciplines ook een aanzienlijk verschillende cultuur en taalgebruik. Programmeurs versus methodologen. Vals-positieven versus type I-fouten. Python versus R. Juist daarom willen we in de translationele datawetenschap beide culturen "erkennen en waarderen"[24], teneinde de meest complete antwoorden op onze data-analyse doelstellingen te kunnen krijgen.

Wat betreft de gecombineerde gereedschapskist voor data-modellering, deze bevat inmiddels vele honderden machine leren-technieken en statistische testen. Zo zijn er bijvoorbeeld algoritmes voor technieken die structuur aanbrengen in de data op basis van bijvoorbeeld beslisbomen (C4.5), associatieregels (Apriori), netwerkverbindingen (PageRank), een bos van beslisbomen (Random Forest), of een ensemble van zwakke voorspellers (AdaBoost). Enzovoorts.

Daarnaast hebben de meeste algoritmes specifieke configuratie-instellingen om optimaal te functioneren. Dit aspect krijgt echter veel te weinig aandacht. Vandaar mijn pleidooi voor meta-algoritmisch modelleren, wat als doel heeft om de beste praktijken voor het optimaal gebruiken van algoritmes in de datawetenschap gestandaardiseerd te documenteren.[25] Wanneer een dataset slechts 250 datapunten heeft, met 10% missende data, welk classificatie algoritme kan ik dan het beste selecteren, en hoe bepaal ik vervolgens de bijbehorende optimale instellingen?

Diep leren

Dit brengt mij tot het populairste deelgebied binnen machine leren: *diep* leren. Diep leren is kort gezegd een verbeterde versie van de techniek uit mijn afstudeerproject van 27 jaar geleden. Nog steeds spreekt de analogie van deze neurale netwerken die het gedrag van het menselijk brein simuleren, bij velen tot de verbeelding. Net als ons menselijk brein kunnen diepe neurale netwerken complexe patronen leren van grote hoeveelheden gegevens.

Er zijn vele voorbeelden van artificiële intelligentie en datawetenschap: zelfrijdende auto's, gezichtsherkenning op foto's, automatische tekstvertalingen, slimme spelcomputers, virtuele gesprekspartners... Diep geleerde modellen zijn net als ons menselijk brein robuust van aard en in staat tot grootse prestaties, maar we begrijpen echter nog steeds niet zo goed hoe zowel ons brein als diep leren precies werkt. De prijs van de topprestaties van diep leren technieken is kortom een gebrek aan transparantie.

Daar komt bij dat de allerbest presterende diep geleerde modellen tevens een onwaarschijnlijk grote hoeveelheid para-

meters kennen. Zo benadert het aantal parameters in recente diep geleerde taalmodellen inmiddels het aantal synapsen in ons menselijk brein, namelijk 100 *biljoen*. Men schat dat het de grote technologiebedrijven omstreeks 100 miljoen euro heeft gekost om dit enorme taalmodel te creëren! Zoveel budget heb ik helaas niet toegezegd gekregen van de Universiteit Leiden.

Wat nu? Gelukkig is het voor veel data-analyse taken niet nodig om een geheel nieuw diep leren model te ontwikkelen. Voor veel taken zijn reeds vooraf getrainde modellen beschikbaar, die we kunnen optimaliseren met behulp van transfer leren. Met transfer leren wordt een voorgetraind data-model op maat afgesteld zodat de reeds gemodelleerde kennis in het model behouden blijft, maar dat tegelijkertijd nieuwe data-analyse taken efficiënt en effectief gerepresenteerd kunnen worden.[26] Daarnaast biedt het ook de mogelijkheid om symbolische, grammaticale informatie toe te voegen aan dergelijke diep lerende, probabilistische modellen.

4.2 Geautomatiseerd machine leren

Wellicht dat het u inmiddels duizelt? Mijn vakgebied van de translationele datawetenschap combineert immers natuurlijke taalverwerking, federatief leren, machine leren, diep leren, transfer leren, enzovoorts. Gelukkig is er een technologische oplossing voor resultaat-gedreven translationele datawetenschappers: *geautomatiseerd* machine leren! Zoals de naam al doet vermoeden, neemt geautomatiseerd machine leren zo veel mogelijk het data-analyse proceswerk uit handen. Allereerst kan geautomatiseerd machine leren de vaak zeer tijdrovende *data voorbereiding* in Fase 3 enorm verkorten. Zo kan geheel automatisch de datakwaliteit verbeterd worden door missende waarden te berekenen, door datatypes vooraf te converteren, en kunnen nieuwe, onderscheidende kenmerken van de data vanzelf ontbloot worden. Handmatig data*goochelen* wordt automatische data*magie*.

Het meest tot de verbeelding sprekende aspect van geautomatiseerd machine leren bevindt zich echter in het automatisch bepalen van het beste machine leren-algoritme, inclusief de juiste instellingen ervan, tijdens de data modellering in Fase 4. Gegeven een dataset en de gewenste uitkomstmaat zoals bijvoorbeeld de mate van verklaarde variantie, kan volautomatisch het optimale algoritme en de optimale configuratie worden bepaald. Daarnaast is het nog zaak om de heersende cultuur mee te laten bewegen. Zo blijkt bijvoorbeeld dat Nederlandse ziekenhuisartsen veel interesse hebben in geautomatiseerd machine leren voor hun klinische onderzoek, tenminste... als het maar wel techniek X gebruikt, want (begin citaat:) "anders begrijpen de anonieme beoordelaars het niet en wijzen ze het manuscript af. Niet omdat het slecht is, alleen omdat ze het niet begrijpen" (einde citaat).[27]

Mede hierom heb ik mijzelf met deze aanstelling ten doel gesteld om geautomatiseerd machine leren breder te introduceren binnen de medische wereld om de datawetenschap te democratiseren, en de zorg en gezondheid sector versneld te innoveren. Om de daad bij het woord te voegen, ontwikkelen we op dit moment dan ook een nieuw mastervak getiteld "translationele datawetenschap" voor studenten met een medische achtergrond dat rondom de mogelijkheden van geautomatiseerd machine leren wordt ontworpen. Tenslotte is er geen betere omgeving dan de Universiteit Leiden om dit plan te realiseren, want LIACS-collega's, zoals onder meer hoogleerde Hoos, zijn immers wereldvermaarde experts op het vakgebied van geautomatiseerd machine leren.

Fase 5: Model Evaluatie

We hebben inmiddels de twee laatste fases in het "sectoroverschrijdende standaardproces voor data-analyse" bereikt: *model evaluatie* en *model implementatie*. Gezien de tijd zal ik deze laatste twee fases korter bespreken. Vanzelfsprekend wil dat echter niet zeggen dat ze minder belangrijk zijn!

Neem nu Fase 5 in het "sectoroverschrijdende standaardproces voor data-analyse": *model evaluatie.* Hét moment waar al het voorgaande werk toe geleid heeft. Hebben we een betekenisvol

antwoord op onze vraag gevonden? Nieuw inzicht gekregen? Belangrijk is het te beseffen dat een datawetenschapper alléén hier geen afdoende antwoord op kan geven. Het betekenisvol duiden van de uitkomsten vergt namelijk domein expertise. In deze *model evaluatie* stap beoordelen we bijvoorbeeld de mate waarin het datamodel voldoet aan de doelstellingen uit Fase 1, en aan de vooraf gestelde succescriteria van de belanghebbenden. Daarnaast trachten we de situaties waarin het beste model tekortschiet, inhoudelijk te duiden.

5.1 Uitlegbaarheid

Vanwege het belang van het kunnen duiden van de modeluitkomsten, dat immers een voorwaarde is voor acceptatie door de eindgebruikers, is het geen verrassing dat het beter uitlegbaar maken van modeluitkomsten inmiddels een zeer actief onderzoeksgebied is. Denk bijvoorbeeld aan het voorspellen van het risico op agressie-incidenten bij psychiatrische patiënten aan de hand van de vele dagelijkse notities van zowel de behandelend arts als de verplegers over het reilen en zeilen van hun patiënten. In eerder onderzoek ontwikkelden we met natuurlijke taalverwerking een voorspelmodel dat in principe beter dan de arts zelf kan voorspellen of een patiënt een verhoogd risico op een toekomstig agressie incident heeft.[28] Desondanks wordt dit model nog niet gebruikt in de dagelijkse praktijk, wat vooral komt doordat de beweegredenen van het diep leren-gebaseerde voorspelmodel ondoorzichtig blijven. Dit resulteert, terecht, in een gebrek aan vertrouwen door eindgebruikers zoals artsen. Zij blijven namelijk wel eindverantwoordelijk voor de genomen beslissing, en moeten deze ook kunnen communiceren met de patiënt en collega's.

Interactieve visualisaties kunnen hier uitkomst bieden, net als in de reeds eerder besproken interactieve dataverkenning uit Fase 2. Bijvoorbeeld via dimensie-reductietechnieken, waarmee datapunten vanuit een hoogdimensionale ruimte, met behoud van onderlinge verhoudingen, naar een tweedimensionale visualisatie vertaald worden. Of door de invloed van bepaalde datapunten te verhelderen door ze te verwijderen uit de data, en het verschil vóór en ná te visualiseren. Of door de nabije omgeving van enkele willekeurige datapunten te visualiseren, om zo een representatieve indruk te krijgen. Er zijn zelfs veelbelovende technieken die het complete voorspelmodel invoeren aan een nieuw te genereren onderscheidend voorspelmodel, zodat vervolgens op interactieve wijze de invloed van specifieke stijlkenmerken kan worden vergeleken.[29,30]

5.2 Rechtvaardigheid

Naast de noodzaak voor uitlegbaarheid van een voorspelmodel is er ook nog de noodzaak voor de rechtvaardigheid van de uitkomsten ervan. De afgelopen jaren is namelijk duidelijk geworden dat voorspelmodellen die in de dagelijkse praktijk ingezet worden, veelal aanzienlijke vooroordelen in zich kunnen dragen. Overigens is dit niet per definitie slecht, of soms zelfs onontkoombaar, tenzij de gebruikers ervan zich hier niet bewust van zijn.[31] Denk bijvoorbeeld aan de Toeslagenaffaire, waarin de Belastingdienst achteraf moest toegeven dat onder meer de kenmerken nationaliteit en tweede nationaliteit, als risicoverhogende factoren werd gevoerd aan de Belastingdienst algoritmes, zonder dat de belanghebbenden dit konden weten.

Vanzelfsprekend is ook in de populatiegerichte zorg de rechtvaardigheid van de gebruikte voorspelmodellen een belangrijk aspect. Zo hebben we recentelijk de rechtvaardigheid onderzocht van een voorspelmodel voor het toedienen van kalmeringsmiddelen bij psychiatrische patiënten.[32] Hier bleek dat geslacht een ongewenste invloed had: vrouwen kregen namelijk gedurende de eerste drie dagen méér kalmeringsmiddelen toegediend, enkel op basis van hun geslacht. Echter, met recentelijk ontwikkelde compensatietechnieken zoals de VooroordeelVerwijderaar-methode kunnen we dergelijke ongewenste vooroordelen neutraliseren en het voorspelmodel rechtvaardiger en inclusiever laten werken.

Voor het beantwoorden van dergelijke relevante onderzoeksvragen rondom de uitlegbaarheid en rechtvaardigheid van artificiële intelligentie en datawetenschap in zorg en gezond-

heid, bundelen wij aan de Universiteit Leiden onze multidisciplinaire krachten in een ELSA-laboratorium voor Gezonde Samenleving en Artificiële Intelligentie, waarin onderzoekers in samenwerking de invloed en impact van ethische, juridische en sociale aspecten in samenhang bestuderen. Ook is onlangs de "Leidraad kwaliteit AI in de zorg" gepubliceerd om zorgprofessionals bij te staan in het juist gebruiken van voorspellende AI-gestuurde modellen. Tenslotte heeft onlangs het Europees Parlement de Wet op de Artificiële Intelligentie in behandeling genomen, in een poging tijdig de interne markt te waarborgen, door de voorwaarden te scheppen voor de ontwikkeling en het gebruik van betrouwbare artificiële intelligentie in de Europese Unie.

Fase 6: Model Implementatie

Fase 6, de laatste fase in het "sectoroverschrijdende standaardproces voor data-analyse", is *model implementatie.* Het is dit deelgebied waarin de uitkomsten van het datawetenschappelijk onderzoek, oftewel alle stappen die ik tot nu toe heb besproken, op de juiste wijze in de populatiegerichte zorgpraktijk worden gebracht. Vanzelfsprekend is hier de nodige documentatie vereist die de vele beslissingen in de voorgaande stappen onderbouwen, zodat de modeluitkomsten op voldoende vertrouwen kunnen rekenen. Maar vanuit datawetenschappelijk perspectief zijn we bovenal geïnteresseerd hoe het voorspelmodel in de praktijk gebruikt kan gaan worden. Want... uit grootschalig onderzoek is onlangs gebleken dat wereldwijd slechts 13% van alle ontwikkelde datamodellen daadwerkelijk in de praktijk geïmplementeerd wordt.[33] Denk bijvoorbeeld maar terug aan het goede risicovoorspelmodel voor agressie incidenten bij psychiatrische patiënten dat ik zojuist besprak. Vervolgens, na publicatie van het model, wordt de vraag vanzelfsprekend hoe het model beheerd zal worden.

6.1 Model publicatie

Er zijn vele mogelijke methoden om een datamodel te implementeren voor gebruik in de dagelijkse praktijk. Een elegante en veilige manier is bijvoorbeeld het publiceren van alle karakteristieken van het datamodel in een standaard uitwisselingsformaat. De data waarmee het model gemaakt is, blijft kortom buitenspel. Een andere strategie gaat nog een stap verder; naast het datamodel wordt tevens een op maat gebouwde eindgebruikersapplicatie opgeleverd, zodat het datamodel zelfomvattend toegepast kan worden in iedere praktijk.

Het hoogst haalbare is wellicht om een bruikbaar datamodel direct te publiceren binnen een interorganisationeel informatiesysteem zoals het Elektronisch patiëntendossier. De arts kan het datamodel dan direct binnen zijn standaard werkomgeving gebruiken. De eerlijkheid gebiedt echter te zeggen dat dit maar zeer zelden lukt. Ook mijn eigen implementatie-ervaringen met het STRIPA-systeem voor het beter voorschrijven van medicatie lijken te duiden op een systeemfout in het huidige Nederlandse zorgsysteem.[34]

In de praktijk zien we vooral een geheel andere publicatiebenadering, namelijk om simpelweg de volledige broncode te publiceren die leidde tot het uiteindelijke datamodel, zodat op de doelcomputers het datamodel opnieuw geconstrueerd kan worden. Dit is voor datawetenschappers een zeer laagdrempelige en populaire optie, maar voor eindgebruikers in de zorgpraktijk echter veelal onwerkbaar.

6.2 Operationeel beheer & beveiliging

Hoe dan ook, los van de gevolgde publicatiestrategie voor het datamodel, rest nog de vraag hoe dit model vervolgens de tand des tijds kan doorstaan. Want... stilstand is achteruitgang. De wereld verandert, dus de data die de wereld beschrijft, verandert mee. Kortom, ook een datamodel heeft periodiek onderhoud nodig om relevant te blijven. Veelal levert het gebruik van een voorspelmodel in de dagelijkse praktijk nieuwe inzichten op, die niet voorzien waren gedurende het ontwikkelproces, wat resulteert in nieuwe data-analyse doelstellingen in een volgende iteratie van het "sectoroverschrijdende standaardproces voor data-analyse".

Tenslotte is er vanwege het gebruik van zorgmodellen, die veelal op basis van gevoelige data geconstrueerd zijn, een expliciete rol voor informatiebeveiliging weggelegd als integraal onderdeel van het procesbeheer, in het bijzonder cybersecurity. Enerzijds is het belangrijk om de inherente risico's rondom de voortdurend verstrengelde procescycli van interne modelontwikkeling naar externe praktijkimplementatie te beheersen.[35] Anderzijds dienen proactief periodieke controles uitgevoerd te worden om bijvoorbeeld het risico op heridentificatie van persoonlijke data te minimaliseren vanwege incomplete anonimiseringsprocessen. Immers, in 1997 bleek reeds dat de medische gegevens van gouverneur Weld van Massachusetts in de Verenigde Staten, alsnog via heridentificatietechnieken uit een geanonimiseerde dataset met verzekeringsgegevens getoverd konden worden.

Samenvatting

Samengevat heb ik allereerst mijn vakgebied Translationele Datawetenschap geïntroduceerd en heb ik tevens toegelicht *waarom* het integreren van fundamenteel en toegepast onderzoek zowel meerwaarde biedt voor de datawetenschap als voor de maatschappij, en populatiegerichte zorg in het bijzonder. Daarna heb ik toegelicht *hoe* het standaard proces voor translationele datawetenschap functioneert, als technische uitbreiding op de wetenschappelijke methode. Tenslotte heb ik een beschrijving gegeven van de zes fases van dit "sectoroverschrijdende standaardproces voor data-analyse", door voor iedere van de zes fases enkele onderzoeksgebieden uit te lichten, zodat u ook kunt begrijpen *wat* translationele activiteiten concreet inhouden.

Dankwoord

Dank aan allen die aan de totstandkoming van mijn benoeming hebben bijgedragen. Allereerst wil ik bedanken het College van Bestuur van de Universiteit Leiden, de Raad van Bestuur van het LUMC, en de afdelingshoofden van PHEG en LIACS, voor het in mij gestelde vertrouwen. Hooggeleerden Numans en Plaat, beste Mattijs en Aske, ik ben een gelukkig mens.

Velen hebben door de jaren heen een rol gespeeld in mijn wetenschappelijke vorming. Vanzelfsprekend kan ik hen niet allen hier benoemen. Daarnaast wil ik jullie, alle aanwezigen hier, hartelijk danken voor jullie komst, ook diegenen die nu via de livestream meekijken. Fijn dat jullie er zijn!

Ik wil mij nu graag kort richten tot mijn persoonlijke Top-3 Beïnvloeders. Allereerst mijn studievriend Edwin Brinkhuis. Beste Edwin, dankzij jouw sluwe list kwam een promotieplek op mijn radar. Jij zag dat ik toe was aan een nieuwe, verdiepende uitdaging. Al wist ik het zelf nog niet. Daar ben ik je eeuwig dankbaar voor.

Ten tweede hooggeleerde Barbiers. Beste Sjef, onder jouw ontspannen maar gerichte leiding bij het Meertens Instituut heb ik mijzelf in vier jaar tijd met grote handelingsvrijheid kunnen ontplooien als datawetenschapper. Jij benoemde bovendien als eerste dat mijn latere start als wetenschapper het vervolg van mijn academische carrière wel eens zou kunnen bespoedigen.

Ten derde hooggeleerde Brinkkemper. Beste Sjaak, in die precies twaalf jaar onder jouw hoede aan de Universiteit Utrecht heb ik alle facetten van het wetenschappelijke werk kunnen ervaren. Jij hebt me altijd begeleid om meer uit mezelf te halen, en moedigde mij ook als eerste aan om te solliciteren op een hoogleraarsfunctie. Daarmee gaf je mij jouw vertrouwen, waarvoor mijn grote dank.

Tenslotte, de eerste rij... Jullie zijn van een geheel andere orde. Lieve pa, dank je dat je mij op 16-jarige leeftijd op subtiele wijze op het VWO wist te houden. Lieve Karin, mijn zus, dank je dat je er werkelijk altijd voor mij bent geweest. Liefste Jet, mijn vrouwlief, jouw levensmotto zegt alles: "*The best is JET to come*". Al 18 jaar geef jij mijn leven kleur, van reuring tot chaos, wij zijn op avontuur! Samen met onze allerliefste Fien. Jij bent mijn goedlachse en zingende schat, jullie maken me *kaulo* gelukkig.

Ik heb gezegd.

Referenties

1 Bush, V. (1945). Science: The Endless Frontier. *Transactions of the Kansas Academy of Science (1903-), 48(*3), 231–264.

2 Stokes, D. (1997). *Pasteurs Quadrant: Basic Science and Technological Innovation*. Brookings Institution Press 1997.

3 Baru, C., Blatecky, A. Croson, R., Grossman, R., Howe, B., Machiraju, R., & Zheleva, E. (2017). *Report of the First Translational Data Science (TDS) Workshop*. Illinois, Chicago.

4 Woolf, S. (2008). The meaning of translational research and why it matters. *Jama, 299(*2), 211-213.

5 Spruit, M., & Lytras, M. (2018). Applied Data Science in Patient-centric Healthcare: Adaptive Analytic Systems for Empowering Physicians and Patients. *Telematics and Informatics, 35(*4), Patient Centric Healthcare, 643–653.

6 Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. & Flach, P. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering, 33(*8), 3048–3061.

7 LUMC Campus Den Haag (2022). De Interdisciplinaire Health Campus Den Haag. Visie document 2022-2025.

8 Carayannis, E., & Campbell, D. (2009). 'Mode 3' and 'Quadruple Helix': toward a 21st century fractal innovation ecosystem. *International journal of technology management, 46(*3-4), 201-234.

9 Spruit, M., Kais, M., & Menger, V. (2021). Automated Business Goal Extraction from E-mail Repositories to Bootstrap Business Understanding. *Future Internet, 13(*10), Trends of Data Science and Knowledge Discovery, 243.

10 Spruit, M., Vroon, R., & Batenburg, R. (2014). Towards healthcare business intelligence in long-term care: an explorative case study in the Netherlands. *Computers in Human Behavior, 30*, ICTs for Human Capital, 698–707.

11 Spruit,M., & Sacu,C. (2015). DWCMM: The Data Warehouse Capability Maturity Model. *Journal of Universal Computer Science, 21(*11), 1508-1534.

12 Borger, T., Mosteiro, P., Kaya, H., Rijcken, E., Salah, A., & Scheepers, F & Spruit, M. (2022). Federated Learning for Violence Incident Prediction in a Simulated Cross-institutional Psychiatric Setting. *Expert Systems with Applications*, 116720.

13 Omta, W., Nobel, J. de, Klumperman, J., Egan, D., Spruit, M., & Brinkhuis, M. (2017). Improving Comprehension Efficiency of HCS Data Through Interactive Visualizations. *ASSAY and Drug Development Technologies, 15(*6), 247–256.

14 Menger, V., Spruit, M., Hagoort, K., & Scheepers, F. (2016). Transitioning to a data driven mental health practice: collaborative expert sessions for knowledge and hypothesis finding. *Computational and Mathematical Methods in Medicine,* Article ID 9089321, 11.

15 Lee, D., de Keizer, N., Lau, F., & Cornet, R. (2014). Literature review of SNOMED CT use. *Journal of the American Medical Informatics Association, 21(*e1), e11-e19.

16 Hripcsak, G., Duke, J., Shah, N., Reich, C., Huser, V., Schuemie, M., ... & Ryan, P. (2015). Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Studies in health technology and informatics, 216*, 574. MEDINFO 2015: eHealth-enabled Health.

17 Harris, Z. (1954). Distributional structure. *Word, 10(*2-3), 146-162.

18 Spruit, M. (1995). FILTER prototype. In Scholtes, J. (Ed.), *Artificial neural networks for information retrieval in a libraries context* (pp. 213–251). European Commission, DG XIII-E3.

19 Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28(*1-2), 3-71.

20 Menger, V., Scheepers, F., Wijk, L. van, & Spruit, M. (2018). DEDUCE: A pattern matching method for au-

tomatic de-identification of Dutch medical text. *Telematics and Informatics, 35(*4), Patient Centric Healthcare, 727–736.

21 Mosteiro, P., Rijcken, E., Zervanou, K., Kaymak, U., Scheepers, F., & Spruit, M. (2021). Machine Learning for Violence Risk Assessment Using Dutch Clinical Notes. *Journal of Artificial Intelligence for Medical Sciences, 2(*1–2), 44–54.

22 Spruit, M., Verkleij, S., Schepper, C. de, & Scheepers, F. (2022). Exploring Language Markers of Mental Health in Psychiatric Stories. *Applied Sciences, 12(*4), Current Approaches and Applications in Natural Language Processing, 2179.

23 Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods 15*, 233–234.

24 VSNU, NFU, KNAW, NWO and ZonMw (2019). *Room for everyone's talent: towards a new balance in recognising and rewarding academics.* White paper. The Hague, November 2019.

25 Spruit, M., & Jagesar, R. (2016). *Power to the People! Meta-algorithmic modelling in applied data science.* In Fred, A. et al. (Ed.), Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (pp. 400–406). KDIR 2016, November 11-13, 2016, Porto, Portugal: ScitePress.

26 Sarhan, I., & Spruit, M. (2020). Can We Survive without Labelled Data in NLP? Transfer Learning for Open Information Extraction . *Applied Sciences, 10(*17), Natural Language Processing: Emerging Neural Approaches and Applications, 5758.

27 Ooms, R., & Spruit, M. (2020). Self-Service Data Science in Healthcare with Automated Machine Learning. *Applied Sciences, 10(*9), Medical Artificial Intelligence, 2992.

28 Menger, V., Spruit, M., Est, R. van, Nap, E., & Scheepers, F. (2019). Machine Learning Approach to Inpatient Violence Risk Assessment Using Routinely Collected Clinical Notes in Electronic Health Records. *JAMA Network Open, 2(*7), e196709.

29 Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., ... & Mosseri, I. (2021). Explaining in Style: Training a GAN to explain a classifier in StyleSpace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 693-702).

30 Jin, D., Jin, Z., Hu, Z., Vechtomova, O., Mihalcea, R. (2022). Deep Learning for Text Style Transfer: A Survey. Computational Linguistics, *48(*1), 1–52.

31 Meppelink, J., Langen, J. van, Siebes, A., & Spruit, M. (2020). Beware Thy Bias: Scaling Mobile Phone Data to Measure Traffic Intensities. *Sustainability, 12(*9), Exploring the Impact of AI on Politics and Society , 3631.

32 Mosteiro, P., Kuiper, J., Masthoff, J., Scheepers, F., Spruit, M. (submitted). Bias Discovery in Machine Learning Models for Mental Health.

33 Davenport, T., & Malone, K. (2021). Deployment as a Critical Business Data Science Discipline. *Harvard Data Science Review.* Published on February 10, 2021.

34 Blum, M., Sallevelt, B., Spinewine, A., O'Mahony, D., ..., Spruit, M., Dalleur, O., Knol, W., Trelle, S., Rodondi, N. (2021). Optimizing Therapy to Prevent Avoidable Hospital Admissions in Multimorbid Older Adults (OPERAM): Cluster Randomised Controlled Trial. *BMJ, 374(*n1585).

35 Pieket Weeserik, B., & Spruit, M. (2018). Improving Operational Risk Management using Business Performance Management technologies. *Sustainability, 10(*3), 640.

Prof. Dr. Marco Spruit

2020 –	Hoogleraar Geavanceerde Datawetenschap in Populatiegerichte Zorg, Universiteit Leiden
2019 – 2020	Universitair hoofddocent Toegepaste Datawetenschap, Universiteit Utrecht
2007 – 2018	Universitair docent Informatiekunde, Universiteit Utrecht
2003 – 2007	Onderzoeker in opleiding Computationele Taalkunde, Universiteit van Amsterdam
1997 – 2006	Onafhankelijk productsoftware ontwikkelaar, Wizzer & Insertable Objects
1995 – 2001	Redacteur Personal Computer Magazine, VNU Business Publications B.V.
1995 – 1997	Big Data systeemontwikkelaar, Marine Inlichtingen- en Veiligheidsdienst
1993 – 1995	Applicatie programmeur, ZyLAB Europe B.V.
1989 – 1995	Doctorandus, Alfa-Informatica, Universiteit van Amsterdam
1990 – 1991	Propaedeuse Muziekwetenschap, Universiteit van Amsterdam
1988 – 1989	Propaedeuse Nederlandse Taal- en Letterkunde, Universiteit van Amsterdam

Marco Spruit is Hoogleraar Geavanceerde Datawetenschap in Populatiegerichte Zorg aan de Universiteit Leiden bij zowel het departement Publieke Zorg & Eerstelijnsgeneeskunde (PHEG) aan de Medische Faculteit (LUMC) als het Leiden Instituut voor Informatica (LIACS) aan de Faculteit der Wiskunde & Natuurwetenschappen (FWN). Hij is zowel geïnteresseerd in het vertalen van nieuwe algoritmes naar nieuwe zorgtoepassingen als in het implementeren van nieuwe inzichten uit deze nieuwe toepassingen in de dagelijkse praktijk.

Marco's strategische onderzoeksdoelstelling is het opzetten van een gezaghebbende nationale infrastructuur voor Nederlandse Taalverwerking en Machine Leren om de Datawetenschap te *democratiseren*. Hij richt zich in het bijzonder op het domein Populatiegerichte Zorg & Welzijn in zijn Translationele Datawetenschap Laboratorium.

Marco leidt de onderzoekslijn Translationele Datawetenschap in Bevolkingsgerichte Zorg op de Health Campus Den Haag. Deze onderzoekslijn heeft drie thema's. Ten eerste, in *Data Techniek* onderzoekt hij de verdere consolidatie, standaardisatie en verrijking van de Extramurale LUMC Academische Netwerk (ELAN) data infrastructuur, in lijn met landelijke initiatieven en in samenwerking met zijn PHEG collega's. Ten tweede, in *Data Analyse* onderzoekt hij technieken voor Natuurlijke Taalverwerking en Machine Leren op hun geschiktheid om huidige en nieuwe soorten translationele onderzoeksvragen te beantwoorden, met name vanuit een democratiserend datawetenschapsperspectief en in samenwerking met zijn LIACS collega's. Ten derde, in *e-Health Implementatie* ontwerpt en implementeert Marco interventies voor Datawetenschap door middel van e-Health software oplossingen binnen de regio in nauwe samenwerking met de Campus partners.

Tot 2020 werkte Marco als universitair hoofddocent in de Natuurlijke Taalverwerkingsgroep bij het departement Informatica van de Universiteit Utrecht, waar hij met name tal van Europees gefinancierde studies (OPERAM, SAF21, SMESEC, GEIGER, OPTICA) en nationaal gefinancierde onderzoeksprojecten (STRIMP, COVIDA) uitvoerde. Hij nam deel aan meerdere leiderschapsprogramma's en behaalde wetenschappelijke kwalificaties zoals de Senior Kwalificatie Onderzoek, Senior Kwalificatie Onderwijs, en het *Ius Promovendi*. Van 2007-2018 was hij universitair docent Informatiekunde en was hij onder meer enkele jaren onderwijsmanager van de Informatiekunde en Toegepaste Datawetenschap opleidingen.

Van 2003-2007 werkte Marco als onderzoeker in opleiding bij de Taalvariatie groep van het Meertens Instituut op het snijvlak van syntactische variatie en dialectometrie als taalkundig datawetenschapper. In 2005 ontving hij een Association for Literary and Linguistic Computing aanmoedigingsprijs voor zijn wetenschappelijke werk. Vóór 2003 was hij tien jaar actief in het bedrijfsleven als Natuurlijke Taalverwerking en Big Data softwareontwikkelaar bij onder meer ZyLAB Europe B.V. en de Nederlandse Militaire Inlichtingen- en Veiligheidsdienst. In 1995 voltooide hij de bovenbouwstudie Alfa-Informatica aan de Universiteit van Amsterdam.

Universiteit
Leiden