Inaugurele rede Wessel Kraaij, leerstoel 'applied data analytics', uitgesproken op 24 2 2017

https://www.universiteitleiden.nl/medewerkers/wessel-kraaij
http://liacs.leidenuniv.nl/~kraaijw/

Data van Waarde

Mijnheer de Rector Magnificus, beste collega's, familie en vrienden, zeer gewaardeerde toehoorders,

Ik wil u graag meenemen in mijn verhaal over de waarde van alledaagse data: gegevens, metingen die u vaak ook zelf kunt verzamelen, en hoe met behulp van data analytics, waarde kan worden gecreëerd voor het individu maar ook voor onze maatschappij.

1 Applied Data Analytics

1.1 Het belang van data voor wetenschap, economie en maatschappij

Onze samenleving wordt stap voor stap steeds verder gedigitaliseerd. Algoritmen zijn steeds beter in staat om van grote hoeveelheden data te leren. Ik noem drie recente voorbeelden: het computerprogramma Alphago heeft de wereldkampioen Go verslagen. Met behulp van *deep learning* kan huidkanker even goed herkend worden als door een dermatoloog [1]. Ook op het terrein van automatisch vertalen wordt de kwaliteit van een menselijke vertaling benaderd [2].

Een domein waar de komende jaren ook veel gaat veranderen is onze mobiliteit. De besturing van voertuigen kan geheel automatisch verlopen op basis van real-time verwerking van sensorische informatie, auto's kunnen ook met elkaar communiceren. Door deze automatische besturing zal het mogelijk worden om het aantal files drastisch te verminderen en zuiniger en veiliger te rijden.

Deze technologische doorbraken zijn mogelijk door recente ontwikkelingen in hardware en algoritmiek. Ook zijn we steeds beter in staat om grote hoeveelheden data te benutten om nieuwe verbanden te ontdekken. Dit wordt steeds belangrijker voor de verschillende wetenschapsgebieden. Het zoeken naar patronen in *big data*, wordt dan ook een *game changer* genoemd in de route 'Toegankelijke en Verantwoorde Waarde Creatie uit Big Data' van de Nationale Wetenschaps Agenda [3].

Het moge duidelijk zijn, deze rede gaat over *Data Science* en *Data Analytics*. De titel van mijn betoog is: "Data van Waarde".

In mijn rede wil ik allereerst kort stilstaan bij mijn keuze om mijn onderzoek vooral te verbinden met maatschappelijke waarde. Daarna zal ik een aantal voorbeelden geven van de toepassing van data science onderzoek in het domein gezondheid. Vervolgens geef ik een voorbeeld van een toepassing in het domein openbaar bestuur. Ik sluit af men een aantal onderzoeksvragen.

1.2 Data Science en Data Analytics

Het vakgebied data science is een interdisciplinair onderzoeksveld met een brede scope dat inmiddels op de meeste universiteiten een plaats in het curriculum heeft gekregen. In Leiden is een prachtig universiteitsbreed Data Science programma gestart waarin er intensieve interactie is tussen wetenschappers uit verschillende faculteiten (bijvoorbeeld gedragswetenschappen, letteren of archeologie) en *data scientists* verbonden aan het LIACS [4] en het Mathematisch Instituut.

Data science is vooral een toegepaste wetenschap waar de wetenschappelijke impact wordt gerealiseerd door het toepassen van data science methoden in specifieke domeinen. Er zijn echter ook uitdagingen op het vlak van de data science zelf, bijvoorbeeld op het vlak van algoritmiek, methodologie, data stewardship en ethiek. Data Science bouwt voort op vakgebieden zoals statistiek, data mining, information retrieval en patroonherkenning. Mijn leerstoel 'Applied Data Analytics' is een versterking van het Leidse Data Science programma en richt zich in het bijzonder op het ontwikkelen van algoritmen voor de automatische interpretatie van grote hoeveelheden ongestructureerde informatie. U kunt daarbij denken aan het inhoudelijk analyseren en beschrijven van een onbekende dataset (welke activiteiten, locaties of personen komen veel voor in deze collectie videomateriaal?), het verklaren van observaties (welke factoren bepalen het succes van een interventie?), of voorspellingen (wat is de kans dat een incident op de A13 uitgroeit tot een verkeersinfarct?).

1.3 Data Analytics voor maatschappelijke waarde

Een aantal platformbedrijven (u kunt denken aan Facebook, Google of Uber) heeft grote financiële successen behaald met innovatieve contentdiensten. Essentieel element van het succes is dat daarbij ook veel informatie over gebruikers wordt verzameld [5]. De gratis beschikbaarheid van de diensten van deze bedrijven heeft wel een keerzijde. Er wordt wel gezegd dat in dit geval de gebruikers zelf het product zijn geworden dat wordt verhandeld. Feit is dat de businessmodellen vaak draaien om gepersonaliseerde advertenties.

Ook overheden verzamelen in toenemende mate persoonsgegevens om beleid te optimaliseren. Het combineren van gegevens van personen en processen kan veel nieuwe kennis opleveren en als basis dienen voor een leerproces. Er is echter ook zorg over de veiligheid van persoonsdata, vanwege regelmatige berichten over onbedoelde lekken van persoonsgegevens. Opvoeding, scholing en wetenschappelijk onderzoek met betrekking tot het beheer en de verwerking van data en de daaraan verbonden risico's is daarom noodzakelijk.

Data science kan op veel manieren worden toegepast. Ik kies in mijn onderzoek aan de universiteit Leiden nadrukkelijk voor de ontwikkeling van Data Analytics technieken voor het creëren van maatschappelijke waarde, zoals gezondheid en duurzaamheid. In een aantal domeinen vallen de waarde voor het individu en voor de maatschappij voor een belangrijk

deel samen. Door longitudinale data over een grote populatie te verzamelen kan een veel persoonlijker (en dus beter) gezondheidsadvies worden afgegeven. De kans dat een geadviseerde behandeling resultaat heeft neemt dan toe. Een persoonlijker advies maakt het aan de andere kant ook mogelijk om beter te voorspellen wanneer een behandeling voor een individuele patiënt niet werkt. Onnodige behandelingen kunnen daarmee worden voorkomen met een positief effect op de kwaliteit van leven van patiënten en een afname van kosten in de zorg. Door persoonlijke gezondheidsdata beschikbaar te stellen voor onderzoek draagt een individu bij aan de dataverzameling op populatieniveau die nodig is om behandeladviezen persoonlijker te maken. Er is daarom een wederzijdse afhankelijkheid van het individuele belang en het algemene belang.

Die maatschappelijke waarde kan in de verschillende domeinen uiteraard niet uitsluitend met data science worden gerealiseerd. Integendeel, intensieve samenwerking met wetenschappers uit bijvoorbeeld het gezondheidsdomein of de gedragswetenschappen is een voorwaarde. Echte voortgang wordt geboekt wanneer *data scientists* de wetenschappelijke uitdagingen in het betreffende domein doorgronden en er vanuit de discipline tegelijk ook begrip is voor de nieuwe mogelijkheden van grootschalige dataverwerking. Die interdisciplinaire samenwerking is cruciaal en gelukkig ook steeds meer vanzelfsprekend, getuige het Data Science programma.

In mijn onderzoek richt ik me momenteel vooral op de empowerment van het individu om de eigen gezondheid en kwaliteit van de leefomgeving positief te beïnvloeden. Een tweede aandachtsgebied is het toepassen van Data Analytics voor stedelijke transities waarvoor een gedragsverandering noodzakelijk is, bijvoorbeeld de energietransitie.

Mijn stelling is dat het verzamelen en analyseren van gedetailleerde data van een individu en de omgeving waarin hij leeft en de vergelijking ervan met de gegevens van een zo specifiek mogelijke referentiegroep leidt tot waarde voor zowel het individu als de groep. Een goed voorbeeld van deze aanpak is zelfmonitoring, waar burgers zelf data over hun gezondheid verzamelen en beheren.

Zelfmonitoring kan zorgen voor de empowerment van burgers onder twee voorwaarden. Als eerste is er specifieke domeinkennis nodig om valide interpretaties en adviezen te kunnen geven. Ten tweede is er bijzondere aandacht nodig voor data governance om persoonsgegevens te beschermen.

2 Data analytics en gezondheid

2.1 Betere zorg en preventie door het koppelen en analyseren van nieuwe databronnen

Geachte toehoorders,

Na deze inleidende woorden over de positionering van mijn onderzoek, wil ik nu breder ingaan op het domein gezondheid als toepassing van data science. Allereerst wil ik iets zeggen over de systeembenadering van gezondheid. Daarna zal ik een aantal voorbeelden te geven van toepassingen van data analytics in de context van die systeembenadering.

Het domein gezondheid associëren we vooral met curatieve zorg, maar het gaat bij gezondheidszorg niet alleen om het behandelen van ziekte. Het gaat ook om voorzorg: het voorkómen van ziekte, bijvoorbeeld door het aanleren van een gezonde levensstijl. Het gaat ook om het beter begrijpen waarom mensen ziek worden. Dit kan bijvoorbeeld door voor grote groepen mensen te inventariseren welke factoren invloed hebben op gezondheid en subgroepen te kunnen onderscheiden. Een belangrijke factor waaraan veel onderzoek wordt gedaan is de genetische aanleg maar daarnaast zijn ook leefstijl gerelateerde factoren (zoals bv. voeding, beweging en slaap) en omgevingsfactoren van invloed.

In de gezondheidssector is een belangrijke beweging gaande die ik als kader wil gebruiken voor het bespreken van de voorbeelden. Het gaat om het zogenaamde P4 concept van gezondheid. P4 staat voor predictie, preventie, personalisatie en participatie. Een belangrijke motivatie is de observatie dat de zorg te veel gericht is op behandeling van ziekten en te weinig op *preventie*. Een tweede observatie is dat behandeling en diagnose gebaseerd zijn op populatiegemiddelden. De top tien voorgeschreven medicijnen in de VS werken in het beste geval bij maar een op de vier patiënten en in het slechtste geval bij een op 25 [6]. In sommige gevallen werkt de medicatie zelfs negatief. Doordat steeds meer data verzameld wordt, komt er steeds meer ruimte voor precisiebehandelingen en medicatie, dat is dus de P van *personalisatie*. Die gepersonaliseerde behandeling maakt gebruik van *predictieve* modellen. Deze voorspellen gezondheidsuitkomsten op basis van longitudinale data verzameld over verschillende gezondheidsdimensies.
Participatie betekent het nadrukkelijk betrekken en centraal stellen van het individu bij alle handelingen rond zijn gezondheid. Het P4 concept heeft zijn oorsprong in de systeembiologie, en is dus een systeembenadering, waarin de interactie tussen verschillende factoren in een wiskundig model wordt gevangen. P4 is de afgelopen jaren in de Verenigde Staten onder de aandacht gebracht door Leroy Hood en in Nederland door Jan van der Greef van de Universiteit Leiden en TNO [7], [8]. De vier principes vormen samen een raamwerk om gezondheidskwaliteit te verbeteren. Om P4 te operationaliseren is het nodig om data van waarde te verzamelen en te interpreteren. In een tweetal P's klinkt nadrukkelijk het belang van het individu door, personalisatie en participatie. Een belangrijke toevoeging aan dit raamwerk voor een meer persoonlijke en op levenskwaliteit gerichte gezondheidszorg is het systematisch over een langere tijd meten van de door patiënten zelf ervaren en gerapporteerde gezondheid, centraal element uit de value-based health care van de Amerikaanse econoom Michael Porter[9].

Ik zal nu een aantal voorbeelden van onderzoek benoemen waar individuele burgers en patiënten zelf een actieve rol spelen in het verbeteren van hun gezondheid. Het verzamelen, analyseren en delen van data speelt bij alle projecten een belangrijke rol.

Een eerste voorbeeld van een project gericht op preventie, predictie en personalisatie is het project SWELL, onderdeel van het nationale ICT-onderzoeksprogramma COMMIT/, dat is uitgevoerd tussen 2011 en 2016.

2.2 Voorbeeld: Zelfmanagement van mentale en fysieke gezondheid bij kenniswerkers (personalisatie, predictie, preventie)

SWELL is gericht op het ontwikkelen van data science technieken als basis voor zelfmanagement van mentale en fysieke gezondheid van kenniswerkers.

Uit gezamenlijk onderzoek van CBS en TNO blijkt dat één op de zeven werknemers in Nederland[10] te maken krijgt met burn-out klachten, met een grote impact op de betrokken persoon en zijn omgeving maar ook de betrokken werkgever. Er zijn indicaties dat de nieuwe mogelijkheden om altijd en overal te werken een risicofactor kunnen zijn. Deze nieuwe mogelijkheden dwingen werknemers om bewuste keuzes te maken, zelf structuur aan te brengen in hun werk en pauzes te nemen. Dit heeft positieve kanten, maar niet iedereen kan even goed met deze vrijheid om gaan. De leidende visie van het SWELL-project is het ontwikkelen van een digitale alter ego die met je mee kijkt, en je activiteiten, fitheid en vermoeidheid registreert. Op basis van die registraties kunnen gewoontes met negatieve gevolgen worden gesignaleerd. De laatste stap is het geven van gepersonaliseerde feedback om gedrag aan te passen. De kracht van de aanpak zit vooral in het combineren van verschillende typen sensor informatie, die longitudinaal worden verzameld in de vorm van een *lifelog*. Als in dit digitale dagboek activiteiten, sociale interactie, momenten van concentratie, emotie, lichamelijke en mentale conditie worden vastgelegd, kan dit veel inzicht bieden in het eigen functioneren. Bijvoorbeeld: *Welke situaties leveren positieve emoties op, welke activiteiten kosten veel energie?* Daarvoor is het echter wel noodzakelijk om met behulp van machine learning technieken de heterogene, ruwe sensordata om te zetten in begrijpelijke statusinformatie. U kunt dan denken aan het type activiteit, tijdsduur, plaats, sociale context, en uitkomstmaten zoals lichamelijke en mentale fitheid. In een bredere gezondheidssetting zou je ook kunnen denken aan luchtkwaliteit of voeding.

In het SWELL-project hebben we een gecontroleerde omgeving gecreëerd om te onderzoeken of we in een werksetting een dergelijk *lifelog* kunnen construeren.

In een lab-experiment zijn proefpersonen aan het werk gezet om een middag lang te werken aan verschillende opdrachten, zoals het schrijven van een essay of het voorbereiden van een presentatie. Proefpersonen waren voorzien van opgeplakte sensoren om ECG en huidgeleiding te meten. Daarnaast werden ze gefilmd met een gewone en 3D camera. Die signalen werden met algoritmen vertaald naar een gestandaardiseerde computer leesbare beschrijving van gelaatsexpressie en lichaamshouding.

Stress is een moeilijk meetbaar fenomeen. Het is helaas niet zo dat je zomaar met een polsbandje stress kan meten. Mijn promovenda Saskia Koldijk heeft bekeken of we nieuwe – objectieve – indicatoren kunnen vinden voor stress. Zo is bekend dat het hormoon cortisol een belangrijke maat is voor stress. LUMC-collega Meijer heeft daar onlangs in zijn oratie over gesproken[11]. In ons onderzoek zijn we echter op zoek gegaan naar een alternatief voor de cortisolmetingen, omdat cortisol alleen in het lab kan worden bepaald en de meetmethodiek behoorlijk intrusief is. Saskia heeft in haar experimenten gevonden dat de status van proefpersonen inderdaad significant verschilt tussen de controle condities en de stress condities. De sterkste verschillen zijn te vinden in de houding, gevolgd door de gezichtsuitdrukking. Zoals u wellicht uit eigen ervaring zou verwachten zijn er aanzienlijke

verschillen tussen personen. De studie heeft laten zien dat verhoogde mentale inspanning zich op een beperkt aantal manieren uit in het gezicht. De ene groep heeft weinig expressie; een tweede groep heeft spanning rond de ogen en een ontspannen mond; een derde groep heeft juist wijd open ogen en een gespannen mond[12].

Uiteraard is het belangrijk om de gemeten sensordata ook van context informatie te voorzien, bijvoorbeeld om artefacten te herkennen, maar ook om naar interacties tussen emotie en stress en specifieke activiteiten en contexten te kunnen zoeken.

Daarom heeft mijn promovenda Maya Sappelli gewerkt aan het herkennen van verschillende werkcontexten aan de hand van computer-interactie (zoals toetsaanslagen en muisklikken). Bij context kunt u denken aan de verschillende taken op een dag van een kenniswerker zoals het afhandelen van email, het maken van een rapport of presentatie. Het blijkt dat een aanpak op basis van een neuraal netwerk in staat is om de juiste context te herkennen op grond van de ruwe ongelabelde computer interactie [13]. De techniek om contexten te herkennen en te labelen kan gebruikt worden voor het automatisch segmenteren en indexeren van computeractiviteiten op onderwerpsniveau. De context index maakt het mogelijk om te zoeken naar interacties tussen activiteiten en mentale toestand zoals emoties, vermoeidheid en stress.

Een nog grotere uitdaging is de ontwikkeling van effectieve coaching-strategieën voor zelfmanagement. Immers, de menselijk neiging is om vaste gewoontes niet zo snel op te geven. Betere informatie kan helpen om awareness te vergroten en dat kan een eerste stap zijn in de richting van gedragsverandering. Bij de ontwikkeling van apps om nieuw gedrag aan te leren in SWELL is inspiratie gezocht bij de theorie van o.a. de Canadese psycholoog Albert Bandura door de coachberichten aan te laten sluiten bij het individuele niveau van self-efficacy, fase in veranderingsproces en context. Kern van de zaak is dat de longitudinale set health analytics, contexten en uitkomsten enerzijds voldoende specifiek moeten zijn om aan te sluiten bij een willekeurig individu, maar aan de andere kant voldoende moeten generaliseren om enige voorspellende waarde te hebben.

De registratie van al deze persoonlijke data vraagt uiteraard om speciale aandacht voor veilige opslag. In het SWELL-project is vooralsnog gekozen voor een architectuur waar de persoonlijke data ook echt alleen voor het individu toegankelijk is, al zijn er wel methoden ontwikkeld om op een gebruikersvriendelijke manier geaggregeerde gegevens (bv het aantal stappen per dag, of slaapkwaliteit) te delen met *peers*. Door de geïnterpreteerde meetdata te vergelijken met vooraf ingestelde doelen ontstaan er mogelijkheden voor sturing. Wanneer mensen een moment van actieve reflectie hebben ingebouwd in hun weekritme[14] kan het nadenken over de discrepantie tussen voorgenomen doelen en behaalde resultaten leiden tot gedragsaanpassingen.

Het laten zien van het verband tussen een bepaald type gedrag en de daarmee geassocieerde toekomstige gezondheidsrisico's is slechts één van de manieren om gedrag te beïnvloeden. De Britse gezondheidspsychologen Charles Abraham en Susan Michie hebben enkele tientallen technieken voor gedragsverandering in kaart gebracht[15]. Er is echter nog weinig bekend over de effectiviteit van de technieken voor verschillende persoonlijkheidstypen.

Samengevat, in SWELL hebben we op basis van data analytics nieuwe technieken ontwikkeld om stress en inspanning te meten en op gedrag te coachen. Een goed voorbeeld dus van de toepassing van de P4 elementen predictie, preventie en personalisatie waarin bovendien de cruciale rol van de sociale wetenschappen naar voren komt.

2.3 Vergelijking met populatiedata: het belang van data governance en 'privacy by design'

De strikte benadering waar persoonlijke data ook alleen voor de persoon zelf toegankelijk is, geeft een maximale privacybescherming. Maar de interpretatie van persoonlijke data op het gebied van leefstijl (voeding, bewegen, slaap, computergebruik) kan veel meer aan betekenis winnen als deze kan worden vergeleken met populatiedata. Door die vergelijking kan iemand veel sneller zien of een bepaalde conditie normaal is, denk bijvoorbeeld aan de groei van een baby. Als de groei achterloopt (ten opzichte van kinderen met een vergelijkbare groeilijn) kan besloten worden tot extra voeding.

Idealiter streven we naar het verzamelen van relevante gezondheidsparameters van ieder individu vanaf 10 maanden voor de geboorte tot het einde van het leven[16], omdat op die manier inzicht ontstaat in de referentiepopulatie. Je zou dan specifieke vragen kunnen stellen zoals: 'Wat is de impact van vroeggeboorte op schoolprestaties?' of 'Hoe goed is mijn conditie in vergelijking met mannen van dezelfde leeftijd? Hoeveel slapen zij gemiddeld per nacht, hoeveel sporten zij?' Door het systematisch in kaart brengen van gezondheidsgerelateerde parameters in een persoonlijk gezondheidsdossier en door de individuele data te kunnen vergelijken met vergelijkbare individuen wordt het mogelijk gemaakt om enerzijds de eigen situatie te beoordelen, maar ook om persoonlijke prognostiek af te geven. Daarvoor is het wel noodzakelijk dat er ook historische gegevens aanwezig zijn.

Hoe persoonlijker we de modellen willen maken, hoe meer data we zullen moeten combineren. Er zijn al plannen gemaakt voor een nationale, uiteindelijke misschien zelfs internationaal georganiseerde data-infrastructuur[17] die het mogelijk maakt om te leren over gedistribueerde datasets voor onderzoek, die door wetgeving omtrent persoonsgegevens en overwegingen van security, niet centraal kunnen worden opgeslagen. Maastricht UMC, LUMC en DTL ontwikkelen samen met partners de 'personal health train' infrastructuur om data die op verschillende plekken is opgeslagen toch beschikbaar te maken voor analyse. Semantische interoperabiliteit van data [18] is daarbij een voorwaarde om algoritmen en modellen naar de data te brengen voor een gedistribueerde machine learning aanpak[19]. In het *Prana Data* project[20] worden daarnaast ook pilots uitgevoerd met data-encryptiemethoden[21] die eenvoudige vormen van data analyse ondersteunen, bijvoorbeeld het berekenen van populatie gemiddelden. We zitten nu in een periode waarin technieken nog niet zijn uitontwikkeld en verschillende ideeën worden getest. Er loopt op dit moment een pilotproject in Limburg met de Personal Health Train. In Rotterdam wordt op initiatief van Medical Delta een proef voorbereid met het beheren van eigen gezondheidsdata onder de naam Mijn Data Onze Gezondheid. Het uitvoeren van dit soort pilots helpt om in de praktijk te leren hoe technieken gericht op de reductie van privacy risico's gerelateerd aan data-analyse, het beste kunnen worden ingezet. Aan de ene kant proberen we de risico's voor individuen te minimaliseren door de burger of patiënt controle te geven over wie toegang heeft tot zijn data. Aan de andere kant moet de architectuur ook mogelijkheden geven om

studies uit te voeren of gepersonaliseerde adviezen te genereren op basis van de data van de mensen die daarvoor toestemming hebben gegeven.

Het is mijn overtuiging dat het op een verantwoorde manier toegankelijk maken (uiteraard onder de nodige voorzorgen) van longitudinale gezondheids- en behandelingsinformatie van een grote populatie kan helpen om het lerend vermogen van de gezondheidszorg te versnellen. Idealiter kunnen arts en patiënt samen een besluit nemen gericht op de beste kwaliteit van leven, mede op basis van deze longitudinale referentiegegevens.

Samenvattend: toegang tot populatiedata kan een belangrijke rol spelen voor gepersonaliseerde adviezen, maar goede data governance is noodzakelijk om burgers ook met een gerust hart data beschikbaar te laten stellen.

2.4 Voorbeeld: actieve leefstijl voor rolstoelgebruikers (Predictie, personalisatie)

Een voorbeeld waar intensief gebruik gemaakt zal worden van verschillende referentiepopulaties is het project ‘Van data naar actie’ dat binnenkort van start gaat in een samenwerking tussen de VU Amsterdam, de Hogeschool van Amsterdam, de Universiteit Campinas in Brazilië en de Universiteit Leiden. Dit consortium is een interdisciplinaire samenwerking tussen revalidatieonderzoek, bewegingswetenschap, voedingswetenschap en data science. Doel van dit onderzoek is het leren van een persoonlijk beweeg- en voedingsadvies voor rolstoelgebruikers, door gebruik te maken van de vastgelegde ervaringen van vergelijkbare individuen in een beveiligde database.

Mensen die in een rolstoel terecht komen na een dwarslaesie of amputatie hebben al gauw te maken met de gevolgen van bewegingsarmoede. Hun activiteitenniveau daalt tot 40% van het normale niveau. Dit leidt tot een verhoogd risico op o.a. overgewicht, diabetes en hart en vaatziekten, met een lagere kwaliteit van leven tot gevolg. We willen voor het onderzoek gebruik maken van een bestaand digitaal platform bestaande uit een centrale database, lerende algoritmen en apps voor gepersonaliseerd advies voor oefeningen en voeding. Dit platform is momenteel nog niet bruikbaar voor rolstoelgebruikers, vanwege hun specifieke beperkingen en mogelijkheden. In het project gaan we daarom eerst onderzoek doen naar de determinanten en factoren die fysieke activiteit en gezondheid bevorderen bij mensen in een rolstoel. We zullen daarbij met een systeemaanpak werken en kijken naar fysiologie, voeding, verhouding activiteit-rust, sociale en fysieke context en psychologische factoren. Iedereen is immers anders en er zijn aanwijzingen dat een gepersonaliseerd advies effectiever is dan een algemeen advies[22]. Dit is zeker het geval met rolstoelgebruikers met zeer verschillende medische achtergrond.

Op basis van de nieuwe kennis zullen we het digitale platform aanpassen voor rolstoelgebruikers, om hen te ondersteunen in het aanleren van een actieve leefstijl. Door gebruik te maken van sensoren kan het beweeggedrag worden vastgelegd in combinatie met de registratie van relevante uitkomstfactoren. De bijdrage van mijn groep is primair gericht op de analyse van de grote hoeveelheid sensordata. We gaan ook voorspellende modellen ontwikkelen voor het succes van een specifiek oefenprogramma, met als input een groot

aantal individu gerelateerde factoren. Als laatste, willen we het advies personaliseren op basis van een zelflerend algoritme.

Voor dit aanbevelingsalgoritme willen we gebruik maken van bestaande meetdata van rolstoelgebruikers, maar het systeem ook laten leren van nieuwe gebruikers van het platform. We verwachten de predictieve modellen en aanbevelingen te kunnen verfijnen door vergelijkingen te maken tussen verschillende groepen: rolstoelgebruiker versus niet rolstoelgebruiker, Amsterdam versus Sao Paulo, beginnende sporter versus gevorderde sporter versus elite sporter. We willen uit de data leren welke factoren bijdragen aan de ontwikkeling van individuen qua fysieke activiteit en welke belemmerende factoren er zijn.

Samenvattend: dit project werkt aan de empowerment van rolstoelgebruikers door een zo goed mogelijk persoonlijk advies te geven, gebaseerd op een predictief model, gebaseerd op ervaringen van vergelijkbare rolstoelgebruikers.

2.5 Voorbeeld : Patient-forum-miner (Participatie, patient reported outcomes)

In het laatste voorbeeld uit het gezondheidsdomein, het Patiënt-Forum-Miner project, komt de P van participatie aan bod.

Patiënten worden mondiger en door hun verbeterde kennis en datapositie stellen zij zich actiever op met betrekking tot regie over eigen gezondheid. De afgelopen twee jaar hebben we projecten uitgevoerd met de contactgroep GIST Nederland, die patiënten met een zeldzame vorm van maag-darmkanker verenigt. Doordat GIST zo weinig voor komt, is er relatief weinig over de ziekte en therapie bekend. In die zin is GIST een typische weesziekte. Patiënten uit de hele wereld communiceren met elkaar via een Facebookgroep en een besloten email-lijst. Op initiatief van de contactgroep GIST hebben we het archief van de email en forum communicatiekanalen semantisch geïndexeerd en gefilterd. Naast de berichten van sociale steun worden er namelijk ook ervaringen uitgewisseld, bijvoorbeeld over het verminderen van bijwerkingen van medicatie. Door text analytics en samenvattingstechnieken toe te passen wordt de informatie zowel voor patiënten als kanker-experts doorzoekbaar gemaakt en kunnen er nieuwe hypothesen worden gegenereerd op basis van statistische analyse van de semantisch geïndexeerde berichten. Een voorbeeld is dat patiënten rapporteren dat het helpt om zo puur mogelijke chocolade te eten bij inname van Glivec, om misselijkheid te verminderen. Dit is natuurlijk nog geen causaal verband, en er zou bijvoorbeeld onderzocht moeten worden of er misschien een verborgen nadelig effect is, doordat bijv. opname van Glivec wordt verminderd. Toch is deze ervaringskennis iets wat bij artsen tot voor kort onbekend was. Dit project is een mooi voorbeeld van citizen science, waar patiënten zelf sturing geven aan onderzoek. Ik denk dat een dergelijke aanpak ook wel een bredere impact kan hebben. Er zijn namelijk vele duizenden zeldzame aandoeningen. Naar schatting heeft 7% van de EU-bevolking een weesziekte. Daarnaast zijn er ook al positieve resultaten bekend van de toepassing van text mining technieken op berichten in on-line communities voor het rapporteren van nieuwe bijwerkingen[23]. Door de patiënten zelf te betrekken kan veel aanvullende kennis worden verzameld die de kennis van de clinical trials kan aanvullen. In

toekomstig onderzoek willen we onderzoeken hoe we de kwaliteit van de informatie in de berichten kunnen bepalen en eventueel kunnen verhogen door bijvoorbeeld een koppeling te maken met gecureerde bronnen van medische informatie.

Samenvattend: Door de ervaringen die door patiënten zijn opgeschreven te structureren en doorzoekbaar te maken, wordt de kennispositie van patiënten versterkt en kunnen zij ook beter participeren in beslissingen over zorg en mede sturing geven aan onderzoek.

3 Data Analytics voor beleidsmakers

3.1 Maatschappelijke waarde in de context van stedelijke regio's

Als laatste wil ik schetsen hoe data analytics een rol kunnen spelen in een heel ander domein, namelijk ter ondersteuning van stedelijke beleidsprocessen gericht op een duurzame toekomst. De afgelopen decennia heeft de liberalisering van de wereldhandel en de ontwikkeling van internet gezorgd voor een substantiële verschuiving in termen van werkgelegenheid, arbeidsproductiviteit, logistiek, energiehuishouding en klimaat.

Het is een enorme opgave voor bestuurders om in samenwerking met bedrijven en in relatie tot de belangen van burgers een transformatie naar een duurzame samenleving tot stand te brengen. Nationale overheden en stedelijke regio's realiseren zich steeds meer dat er majeure koerswijzigingen nodig zijn om die transformatie te realiseren. In de metropoolregio Rotterdam Den Haag is er een plan gemaakt door intensieve samenwerking van overheid, markt en kennispartijen. Technologische ontwikkelingen zijn een belangrijke belofte om bijvoorbeeld op het gebied van energie, voeding en mobiliteit duurzamer te gaan leven. Net als bij het domein gezondheid zijn de baten van investeringen niet direct zichtbaar maar liggen in de toekomst. Een publiek private samenwerking kan helpen voorkomen dat vooral op korte termijn belangen wordt gestuurd. In essentie gaat het om het realiseren van een transitie-agenda gericht op duurzame maatschappelijke waarden door samenwerking van verschillende stakeholders

Ik wil onderzoeken of voor deze transitie ook een systeembenadering mogelijk is en of principes uit het P4 raamwerk een nieuwe betekenis kunnen krijgen in deze context. Ook hier willen we de interacties tussen de verschillende factoren en uitkomstmaten kwantificeren op basis van longitudinale big-data. Als we voor de data gebruik maken van metingen per huishouden, ontstaan er mogelijkheden voor aggregatie op wijkniveau, die beleidsmakers beter inzicht geven en de basis vormen voor *evidence based* beleid. Een dergelijke *bottom up* aanpak biedt mogelijkheden om de participatie van burgers te bevorderen en het meetproces transparanter te maken. Als voorbeeld: Het vergelijken van uw maandgemiddelde energiegebruik met de mediaan van de straat kan dan veel inzicht geven voor uzelf, maar dat kan alleen als buurtgenoten ook bijdragen aan de geaggregeerde data. Voor een dergelijke aanpak is transparantie op alle niveaus wat betreft de verwerking en weging van data cruciaal voor acceptatie.

3.2 Data Analytics voor 'real time' beleidsindicatoren

Dit jaar starten we in samenwerking met de faculteit Governance en Global Affairs en het Center for Big Data Statistics van CBS een onderzoek naar het ontwikkelen van beleidsindicatoren op basis van big data bronnen. Door de model gedreven en data gedreven disciplines samen te brengen [24], verwachten we tot een meer robuuste methodiek voor beleidsaanbevelingen te komen.

Het monitoren van indicatoren op verschillende beleidsterreinen zoals werkgelegenheid, mobiliteit, gezondheid en veiligheid behoort tot de standaardpraktijk van openbaar bestuur. Immers, beleid kan alleen worden gemaakt op basis van betrouwbare gegevens. Traditioneel worden veel van dergelijke gegevens verzameld via vragenlijsten. Deze aanpak kent nadelen: de meest recente cijfers lopen vaak flink achter op de actualiteit waardoor kwantitatieve analyses van interventies pas relatief laat beschikbaar komen. Het is daarom moeilijk om een echte feedback loop te maken voor beleidsprocessen. Bovendien is er altijd het risico op bias in steekproeven.

In een big data aanpak wordt in principe alle beschikbare informatie meegenomen in de analyse en kan er veel frequenter worden geactualiseerd. Een voorbeeld van een dergelijke aanpak is het via text mining technieken analyseren van open internetbronnen waaronder sociale media. Daarbij gaat het om technieken zoals entiteit-herkenning, sentiment mining, event herkenning. Websites van bedrijven kunnen bijvoorbeeld worden geanalyseerd op kenmerken van vacature-aankondigingen. Andere voorbeelden van beschikbare big data bronnen zijn verkeerslusdata of *floating cardata* voor verkeersstromen. Met deze bronnen is al ervaring opgedaan die ingezet kan worden om toepassingen voor de grootstedelijke praktijk te ontwikkelen. Ook hier geldt weer dat data governance en privacy by design belangrijke componenten zijn in het onderzoek.

Ons plan is om de komende jaren in samenwerking met Center for Big Data Statistics (CBS, met o.a. UL en TNO) een methodiek te ontwikkelen om beleidsindicatoren voor verschillende beleidsterreinen te realiseren voor stedelijke regio's om te beginnen met Den Haag en Metropool Regio Rotterdam Den Haag. Vervolgvragen zijn gericht op i) de extractie van nieuwe typen informatie door koppeling met traditionele gegevensverzamelingen van CBS en Den Haag; ii) het toepassen van een systeembenadering om de interacties tussen factoren te kwantificeren en het effect van interventies zo goed mogelijk te kunnen voorspellen en visualiseren; iii) generalisering van de methodiek voor toepassing in andere stedelijke regio's in de EU. Dit onderzoek zal ook nauw worden gekoppeld aan het onderwijs van de nieuwe master ICT in Business toegespitst op de publieke sector. Dit onderwijs is erop gericht om toekomstige beleidsadviseurs en ICT-experts werkzaam in de sector een beeld te geven van de mogelijkheden van het gebruik van data analytics voor beleid. Daarnaast wordt aandacht gegeven aan het juridisch en ethisch kader van de methoden.

Samenvattend: Er liggen grote kansen om big data te analyseren als basis voor evidence based beleid en te werken aan een lerend (eco)systeem. Er is onderzoek nodig hoe de nieuwe bronnen kunnen worden gekoppeld aan de gangbare methodiek voor beleidsvoorbereiding en evaluatie.

4 Uitdagingen binnen domein Data Science

Na het beschrijven van een aantal maatschappelijke uitdagingen waar naar mijn idee een data gedreven aanpak kan leiden tot betere keuzes en prioriteiten, wil ik een aantal uitdagingen voor *data science* onderzoek noemen waaraan ik in multidisciplinair verband ga werken in Leiden, ook in relatie met mijn andere werkgever TNO.

Als eerste wil ik een systematiek ontwikkelen voor het verzamelen en langdurig bewaren van data, voor gevalideerde en robuuste indicatoren voor gezondheid en voor de waardensystemen van duurzame stedelijke regio's. Het gaat daar vooral om het verzamelen van de juiste data en in sommige gevallen om het combineren van verschillende modaliteiten. Het is belangrijk om bij die dataverzameling ook altijd contextinformatie mee te nemen. Een tweede uitdaging is het ontwikkelen van een architectuur om analyses en aggregaties uit te voeren die de rechten van de datasubjecten respecteert en hen controle geeft over wie toegang heeft wanneer het persoonsdata betreft. Voorbeelden daarvan zijn health data cooperatives of de personal health train. Een derde uitdaging is het definiëren en valideren van een similarity functie voor lifelogs, de longitudinale gezondheidsdata, hoe kunnen we vanuit de ruwe data zien welke personen qua verloop in gezondheidsparameters op elkaar lijken? De vierde uitdaging is het leren van voorspellende modellen vanuit longitudinale geobserveerde data en de daaraan gekoppelde onzekerheid. Een laatste belangrijke interdisciplinaire uitdaging is de toepassing van value based analytics voor het berekenen van de effecten van verschillende scenario's. Ik verwacht dat combinatie van data over domeinen heen en de optimalisatie over het ensemble van waarde indicatoren, in beleidsmatige termen: integraal beleid, een grote potentie heeft, ten opzichte van de praktijk van optimalisatie per beleidsterrein.

5 Afsluiting

Dames en heren, ik heb in mijn rede uitgelegd dat data, dataverwerkingstechnieken, data science in zichzelf neutraal zijn. In deze 21e eeuw hebben internetbedrijven gezorgd voor een transformatie door grote hoeveelheden data te verzamelen en daarmee gepersonaliseerde diensten aan te bieden. In mijn onderzoek richt ik me op het versterken van de positie van het individu, door data in te zetten voor maatschappelijke waarden, die ten goede komen aan de gehele maatschappij. Ik heb uitgelegd dat het systematisch verzamelen van data over bijvoorbeeld gedrag, omgeving en indicatoren voor waarden zoals gezondheid, duurzaamheid, veiligheid en kwaliteit leefomgeving kan leiden tot nieuwe inzichten. De inzichten kunnen ontstaan door het combineren van data uit verschillende domeinen en op verschillende aggregatieniveaus. Ik heb de impact van deze aanpak geïllustreerd aan de hand van een aantal voorbeelden. Voorwaarde voor een succesvolle op maatschappelijke waarde gerichte data analytics aanpak is wel dat de persoonlijke levenssfeer effectief wordt beschermd door privacy en data eigenaarschap op te nemen in het ontwerpproces.

6 Dankwoord

Ik ben aan het eind gekomen van mijn rede. Ik wil graag een aantal mensen bedanken die in het bijzonder hebben bijgedragen aan het tot stand komen van deze leerstoel.

Als eerste wil ik het College van bestuur en het faculteitsbestuur bedanken voor het in mij gestelde vertrouwen.

Ik wil uiteraard ook een aantal LIACS collega's bedanken:

Hooggeleerde Kok, beste Joost, door jouw inzet in 2015, - je was toen nog wetenschappelijk directeur van het LIACS- is het proces om deze leerstoel te realiseren bijzonder prettig en vlot verlopen. Al vanaf ons eerste gesprek in Leiden had ik het gevoel dat er veel potentie zat in een mogelijke samenwerking. We hebben intussen al een aantal mooie resultaten gerealiseerd en ik denk dat het daar niet bij zal blijven. Ik zie ook mooie kansen voor het uitbouwen van het Data Science programma.

Hooggeleerde Plaat, beste Aske, mijn benoeming werd een feit in februari 2016. Jij was toen net begonnen als de nieuwe wetenschappelijk directeur. Ik wil je bedanken voor het feit dat je altijd tijd hebt willen vrijmaken om me wegwijs te maken in de nieuwe organisatie. Ik wil je ook bedanken voor je inspanningen die ervoor gezorgd hebben dat dit instituut een zo prettige werksfeer heeft, wat een belangrijke bouwsteen is voor excellent onderzoek en onderwijs.

Bij LIACS werk ik inmiddels samen met een flink aantal collega's die ik helaas niet allemaal individueel kan bedanken. Twee collega's die ik wel wil noemen zijn Jaap van den Herik en Cor Veenman. Hooggeleerde van den Herik, beste Jaap, door jouw activiteiten vanuit het LCDS komen er mooie samenwerkingen tot stand. Ik hoop daar ook mijn steentje aan te kunnen bijdragen. Beste Cor, dank voor je inzet om samen een nieuw master vak op te zetten. Door jouw ervaring kunnen we een breed scala aan praktijk cases aan de studenten voorleggen.

Ik wil graag ook mijn andere werkgever TNO bedanken. Hooggeleerde Keurentjes, Hooggeleerde Werkhoven, aanwezig namens de Raad van Bestuur en de directie Technical Sciences van TNO. Beste Jos, beste Peter, ik wil jullie bedanken omdat ik altijd de ruimte heb gekregen om mijn academische carrière op te bouwen. Ik hoop dat de combinatie van het fundamentele onderzoek aan het LIACS en het toegepaste onderzoek bij TNO zal leiden tot meer maatschappelijke waarde.

Ik heb in mijn rede gesproken over het SWELL project waaraan ik leiding heb mogen geven binnen het COMMIT/ programma. Hooggeleerde Smeulders, Hooggeleerde Lagendijk, beste Arnold, beste Inald. Bedankt dat jullie mij hebben betrokken bij het opzetten van de COMMIT/ community in Nederland. Ik heb er veel aan te danken.

Er zijn vandaag ook TNO-collega's aanwezig en veel bekende gezichten uit de netwerken big data en gezondheid. Ik waardeer het bijzonder dat jullie vandaag aanwezig zijn en ik zie uit naar het continueren van de samenwerking.

Ik wil ook mijn studenten en Aio's bedanken voor hun aanwezigheid. Het is een rijke ervaring om met jullie te kunnen werken en jullie te mogen begeleiden.

Beste vrienden en familie, fijn dat jullie er zijn! Het geeft me een bijzonder gevoel om deze rede uit te spreken in deze stad met een historische band met de wetenschap maar ook met de familie Kraaij.

Beste vader en moeder, ik heb zoveel aan jullie te danken. Ik hoop dat jullie ook genieten van deze dag.

Lieve Lyne, Ruben en Gaël, wat hebben we toch een mooi leven samen! Dank voor jullie steun en inspiratie.

Ik heb gezegd.

7 Referenties

[1] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[2] Y. Wu *et al.*, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *ArXiv160908144 Cs*, Sep. 2016.

[3] "Nationale Wetenschaps Agenda | vragen – verbindingen – vergezichten." [Online]. Available: http://www.wetenschapsagenda.nl/. [Accessed: 04-Feb-2017].

[4] "- LIACS - Leiden Institute of Advanced Computer Science." [Online]. Available: http://liacs.leidenuniv.nl/. [Accessed: 04-Feb-2017].

[5] A. Roosendaal, "Digital Personae and Profiles in Law: Protecting Individuals' Rights in Online Contexts," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2313576, Aug. 2013.

[6] N. J. Schork, "Personalized medicine: Time for one-person trials," *Nat. News*, vol. 520, no. 7549, p. 609, Apr. 2015.

[7] L. Hood, "Systems Biology and P4 Medicine: Past, Present, and Future," *Rambam Maimonides Med. J.*, vol. 4, no. 2, Apr. 2013.

[8] J. van der Greef, T. Hankemeier, and R. N. McBurney, "Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials?," *Pharmacogenomics*, vol. 7, no. 7, pp. 1087–1094, Oct. 2006.

[9] M. E. Porter, "What Is Value in Health Care?," *N. Engl. J. Med.*, vol. 363, no. 26, pp. 2477–2481, Dec. 2010.

[10] CBS, "CBS en TNO: Een op de zeven werknemers heeft burn-outklachten." [Online]. Available: https://www.cbs.nl/nl-nl/nieuws/2015/47/cbs-en-tno-een-op-de-zeven-werknemers-heeft-burn-outklachten. [Accessed: 05-Feb-2017].

[11] O. C. Meijer, *Cortisol van kop tot teen: over "goed en kwaad" van een stresshormoon*. Leiden: Universiteit Leiden, 2016.

[12] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Trans. Affect. Comput.*, vol. PP, no. 99, pp. 1–1, 2016.

[13] M. Sappelli, S. Verberne, and W. Kraaij, "Adapting the Interactive Activation Model for Context Recognition and Identification," *ACM Trans Interact Intell Syst*, vol. 6, no. 3, p. 22:1–22:30, Sep. 2016.

[14] D. A. Schön, *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*, vol. xvii. San Francisco, CA, US: Jossey-Bass, 1987.

[15] C. Abraham and S. Michie, "A taxonomy of behavior change techniques used in interventions," *Health Psychol. Off. J. Div. Health Psychol. Am. Psychol. Assoc.*, vol. 27, no. 3, pp. 379–387, May 2008.

[16] E. J. Topol, "Individualized medicine from prewomb to tomb," *Cell*, vol. 157, no. 1, pp. 241–253, Mar. 2014.

[17] "Health-RI," *Dutch Techcentre for Life Sciences*. [Online]. Available: http://www.dtls.nl/health-ri/. [Accessed: 05-Feb-2017].

[18] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, p. 160018, Mar. 2016.

[19] A. Damiani *et al.*, "Distributed Learning to Protect Privacy in Multi-centric Clinical Studies," in *The 15th Conference on Artificial Intelligence in Medicine*, J. H. Holmes, R. Bellazzi, L. Sacchi, and N. Peek, Eds. Pavia, Italy: Springer, 2015, pp. 65–75.

[20] "PRANA-DATA." [Online]. Available: https://pranadata.nl/. [Accessed: 06-Feb-2017].
[21] Z. Erkin, T. Veugen, T. Toft, and R. L. Lagendijk, "Generating private recommendations efficiently using homomorphic encryption and data packing," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 3, pp. 1053–1066, 2012.
[22] P. Krebs, J. O. Prochaska, and J. S. Rossi, "DEFINING WHAT WORKS IN TAILORING: A META-ANALYSIS OF COMPUTERTAILORED INTERVENTIONS FOR HEALTH BEHAVIOR CHANGE," *Prev. Med.*, vol. 51, no. 3–4, p. 214, Oct. 2010.
[23] A. Sarker *et al.*, "Utilizing social media data for pharmacovigilance: A review," *J. Biomed. Inform.*, vol. 54, pp. 202–212, Apr. 2015.
[24] M. Janssen and G. Kuk, "Big and Open Linked Data (BOLD) in research, policy, and practice," *J. Organ. Comput. Electron. Commer.*, vol. 26, no. 1–2, pp. 3–13, Apr. 2016.