Prof.dr.ir. Wessel Kraaij

Data of Value

Universiteit
Leiden
The Netherlands

Discover the world at Leiden University

Data of Value*

Inaugural lecture by

Prof.dr.ir. Wessel Kraaij

on the acceptance of his position as professor of

Applied Data Analytics

at the Universiteit Leiden

on Friday February 24, 2017.

Universiteit
Leiden
The Netherlands

* This is the English translation of the lecture that was originally given in Dutch.

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

I would like to share with you my thoughts about the value of everyday data: information and measurements that you can often even collect yourself and how data analytics can be used to create value for the individual, but also for society.

1 Applied Data Analytics

1.1 The importance of data for science, economy and society

Our society is gradually being digitalized, step by step. Algorithms are becoming increasingly able to learn from large amounts of data. Three recent examples: the Alphago computer program has defeated the world champion of the game Go (Silver et al. 2016). With the help of 'deep learning', skin cancer can be recognized with the same level of accuracy as by a dermatologist (Esteva et al. 2017). Finally, in the field of translation, the quality of automatic translation now approaches that of a translation by a human (Wu et al. 2016).

A domain that will change rapidly in the coming years is our mobility. Vehicles can be completely automatically controlled, based on real-time processing of sensory information. Cars will also be able to communicate with each other. This automatic control will make it possible to drastically reduce traffic congestion and to drive more efficiently and safely.

These technological breakthroughs are possible due to recent developments in hardware and algorithms. We are also increasingly able to utilize large amounts of data to discover new associations. This is becoming increasingly important for different areas of science. The search for patterns in big data, therefore, is considered a game changer in the "Accessible and Responsible Value Creation from Big Data" roadmap of the Dutch *Nationale Wetenschaps Agenda*.[1]

It should have become clear by now that this lecture is about Data Science and Data Analytics. The title of my lecture is: "Data of Value".

In my lecture, I would like to begin by briefly stating my motivation in choosing to connect my research with societal value. Then I will give some examples of the application of data science research in the field of health. Next, I will give an example of an application in the Public Policy domain. I will conclude with a series of research challenges.

1.2 Data Science and Data Analysis

The field of data science is an interdisciplinary research field with a broad scope that has now become a part of the curriculum of most universities. In Leiden, a wonderful university-wide Data Science program has been launched in which there is intensive interaction between scientists from different faculties (e.g. behavioural sciences, literature or archaeology) and data scientists appointed at the LIACS[2] and the Mathematical Institute.

Data science is primarily an applied science where the scientific impact is realized by applying data science methods in specific domains. However, there are also challenges in the field of data science itself, for example in the field of algorithms, methodology, data stewardship and ethics. Data Science builds on fields such as statistics, data mining, information retrieval and pattern recognition. My chair 'Applied Data Analytics' reinforces the Leiden Data Science program and focuses on developing algorithms for the automatic interpretation of large amounts of unstructured information. You could think of analysing and describing the content of an unknown dataset (what activities, locations or persons are common in this collection of video material?); explaining observations (which factors determine the success of an intervention?); or predictions (what is the probability that an incident on the A13 motorway will become a traffic jam?).

1.3 Data Analytics for societal value

A number of platform companies (for example, Facebook, Google or Uber) have achieved huge financial successes with innovative content services. An essential element of the success is that much information about users is being collected (Roosendaal 2013). The free availability of the services of these companies has a downside. It is said that in this case the users themselves have become the product that is being traded. In fact, business models are often based on personalized ads (Ford, Kraft and Tewari 2003).

Also, governments increasingly gather personal data to optimize policies. Combining data from persons and processes can provide a lot of new knowledge and serve as a basis for a learning process. However, there are also concerns about the security of personal data due to regular reports of accidental leaks of personal data.[3] Education, training and scientific research regarding the management and processing of data and the associated risks is therefore necessary.

Data science can be applied in many ways. In my research at Leiden University, I explicitly choose to develop Data Analytics techniques to contribute to social values, such as health and sustainability. In several domains, the value for the individual and for society coincide. By collecting longitudinal data on a large population, much more personalised (and therefore better) health advice can be provided. The chance that a recommended treatment will lead to a positive result will then increase (Ruiter et al. 2006). More personalised advice, on the other hand, also allows for better predictions when treatment for a patient will be ineffective (Evans and McLeod 2003; Evans and Relling 2004). Unnecessary treatments can therefore be avoided with a positive effect on quality of life of patients and a reduction of the cost of providing care. By providing personal health data for research, an individual contributes to the population-level data collection needed to make treatment recommendations more personal. There is therefore a mutual dependence of the individual interest and the general interest.

This societal value in the various domains cannot, of course, be achieved with just data science. On the contrary, intensive cooperation with scientists from, for example, the health domain or behavioral sciences is a prerequisite. Real progress is made when data scientists understand the scientific challenges in the domain at hand and when experts from the domain discipline also understand the new opportunities generated by large-scale data processing. This interdisciplinary cooperation is crucial and fortunately more and more self-evident, as shown in the Leiden Data Science program.

In my research, I focus mainly on empowerment of the individual to positively influence their own health and quality of life. A second focus area is the application of Data Analytics for urban transitions for which a change of behavior is necessary, such as the energy transition.

My proposition is that collecting and analyzing detailed data of an individual and the environment in which he lives and comparing this with the data of a specific possible reference group brings value to both the individual and the group. A good example of this approach is self-monitoring, where citizens themselves collect and manage data about their health.

Self-monitoring can provide the empowerment of citizens under two conditions. First, specific domain knowledge is required to provide valid interpretations and advice. Secondly, special attention is needed for data governance to protect personal data.

2 Data analytics and health

2.1 Better healthcare and prevention by linking and analysing new data sources

Dear audience,

After these introductory words about the positioning of my research, I now want to further discuss the field of health as a data science application. First of all, I would like to say something about the systems approach to health. Then I will provide some examples of applications of data analytics in the context of the systems approach.

The domain of health is primarily associated with curative care, but health care is not just about treating illness. It is also a matter of prevention: preventing illness, for example by adopting a healthy lifestyle. It's also about understanding better why people get sick. This can be done, for example, by analyzing for large groups of people which factors can affect health and to distinguish subgroups. An important factor to which a lot of research has been devoted is the individual genetic profile, but lifestyle-related factors (such as nutrition, exercise and sleep) and environmental factors are also important (McGinnis, Williams-Russo and Knickman 2002).

In the health sector, an important movement is gaining ground; which I will use as a framework for discussing the examples. This movement promotes the so-called P4 concept of health. P4 stands for prediction, prevention, personalization and participation. An important motivation is the observation that healthcare is too focused on disease treatment and not enough on *prevention*. A second observation is that treatment and diagnosis are based on population averages. The top ten prescription drugs in the United States work in the best case for only one in four patients and in the worst case for one in twenty-five (Schork 2015). In some cases, the medication even has a negative effect. As more and more data is gathered, there is more room for precision treatment and medication, which is the P for *personalization*. This personalized treatment uses *predictive* models. These predict health outcomes based on longitudinal data, collected across different health dimensions.

Participation means involving the patient and making them the center of all actions concerning their health. The P4 concept originates in system biology (Kitano 2002), and thus is a system approach, capturing the interaction between different factors in a mathematical model. P4 has been proposed in the United States in recent years by Leroy Hood and in the Netherlands by Jan van der Greef of Leiden University and TNO (Van der Greef, Hankemeier and McBurney 2006; Hood and Friend 2011). The four principles together form a framework for improving health quality. In order to operationalize P4, it is necessary to collect and interpret data of value. In two P's, the importance of the individual is explicitly linked: personalization and participation. An important addition to this framework for a more personal and quality-oriented healthcare system is to measure patient reported outcome measures systematically over a longer period, a central element of the value-based health care of American economist Michael Porter (Porter 2010).

I will now give a few examples of research where individual citizens and patients themselves play an active role in improving their health. The collection, analysis and sharing of data plays an important role in all projects.

A first example of a project aimed at prevention, prediction and personalization is the project SWELL[4], part of the national ICT research program COMMIT /, which was conducted between 2011 and 2016.

2.2 Example: Self-management of mental and physical health in knowledge workers (personalization, prediction, prevention)

SWELL aims to develop data science techniques as a basis for self-management of mental and physical health of knowledge workers. Joint research by CBS and TNO shows that one in

seven employees in the Netherlands is affected by burn-out complaints, with a major impact on the person concerned and his environment but also the employer concerned. There are indications that the new possibilities for working anywhere and anytime can be a risk factor. Because of these new opportunities, employees have to make conscious choices, to establish structure in their work and take breaks at regular moments. This has positive sides, but not everyone can handle this freedom adequately (Slijkhuis 2012). The leading vision of the SWELL project is to develop a digital alter ego that is following you, and records your activities, fitness and fatigue. Based on those registrations, habits with negative consequences can be identified. The final step is to provide personalized feedback to adapt or change behavior. The strength of the approach is mainly in combining different types of sensor information, which are collected longitudinally in the form of a *lifelog* (Gemmell, Bell and Lueder 2006; Aizawa et al. 2004; Doherty and Smeaton 2008). When activities, social interaction, moments of concentration, emotion, physical and mental fitness are recorded in this digital diary, this can provide a lot of insight into someone's own *biopsychosocial* system. For example: *What situations do lead to positive emotions, which activities cost a lot of energy?* However, it is necessary to convert the heterogeneous raw sensor data into intelligent status information using machine learning techniques. Examples of the raw sensor data are: type of activity, duration, place and social context. In a broader health setting, sensor data can also include air quality and nutrition. Outcomes are for example physical and mental fitness.

In the SWELL project, we have created a controlled environment to investigate whether we can construct such a lifelog in a work setting. In a lab experiment, subjects have been asked to work on a variety of assignments for a whole afternoon, such as writing an essay or preparing a presentation. Subjects were equipped with adhesive sensors to measure ECG and skin conductance. In addition, they were filmed with a regular and 3D camera. Those signals were translated with algorithms to a standardized computer readable description of facial expression and posture (Koldijk et al. 2014).

Stress is a phenomenon which is difficult to measure. Unfortunately, it is not possible to simply measure stress with a wrist strap. My PhD student Saskia Koldijk has examined whether we can find new objective indicators for stress. It is well known that the hormone cortisol is an important indicator of stress. LUMC[5] colleague Meijer recently discussed this in his inaugural lecture (Meijer 2016). However, in our research we looked for an alternative to cortisol measurements because cortisol can only be determined in the lab and the method of measurement is quite intrusive. Saskia found in her experiments that the status of subjects does actually significantly differ between the control conditions and the stress conditions. The strongest differences can be found in posture, followed by facial expression. As you probably know from your own experience, there are significant differences between individuals. The study has shown that increased mental exertion is expressed in a limited number of ways. One group has little expression; A second group has tension around the eyes and a relaxed mouth; A third group has wide open eyes and a tense mouth (Koldijk, Neerincx and Kraaij 2016).

Of course, it is important to annotate the measured sensor data with context information, for example to recognize artifacts, but also to search for interactions between emotion and stress and specific activities and contexts.

That is why my PhD student Maya Sappelli has worked on recognizing different working contexts using computer interaction (such as keystrokes and mouse clicks). Contexts can refer to the various tasks in a day of a knowledge worker such as handling email, making a report or presentation. It appears that an approach based on a neural network can recognize the correct context based on the raw unlabeled computer interaction (Sappelli, Verberne and Kraaij 2016). The technique for recognizing and labeling contexts can be

used to automatically segment and index computer activity at the subject level. The context index makes it possible to search for interactions between activities and mental states such as emotions, fatigue and stress.

An even bigger challenge is the development of effective coaching strategies for self-management. After all, the human tendency is not to give up ingrained habits. Better information can help in raising awareness and this can be a first step toward behavioral change. In the development of apps to acquire new behavior in SWELL, inspiration has been sought from the theory of, among others, Canadian psychologist Albert Bandura by tailoring the e-coach messages to the individual level of self-efficacy, the phase in the change process and context (Bandura 1977). The essence of the matter is that the longitudinal set of health analytics, contexts and outcomes must be sufficiently specific to connect with any individual but, on the other hand, can be sufficiently generalized to have predictive value.

The registration of all these personal data, of course, requires special attention for safe storage. For the time being, the SWELL project has chosen an architecture where personal data is only accessible to the individual, although methods have been developed to share aggregated data (e.g., number of steps per day or sleep quality) in a user-friendly manner with peers. By comparing the interpreted measurement data with preset goals, there are possibilities for guidance and feedback. Thinking about the discrepancy between intended goals and achieved results may lead to behavioural adjustments. This might be quite effective if there is a moment of active reflection as part of a weekly routine (Schön 1987).

Showing the connection between a particular type of behaviour and the associated future health risks is just one of the ways in which to influence behavior. British health psychologists Charles Abraham and Susan Michie have described several dozens of behavioural change techniques (Abraham and Michie 2008). However, little is known about the effectiveness of techniques for different personality types.

In summary, in SWELL we have developed new techniques based on data analytics to measure stress and physical exertion, and to coach behavior. This is a good example therefore, of the application of the P4 elements prediction, prevention and personalization, which also reveals the crucial role of social sciences.

2.3 Comparison with population data: the importance of data governance and 'privacy by design'

The strict approach, where personal data is accessible to the person himself, provides maximum privacy. But the interpretation of personal data in the lifestyle domain (nutrition, exercise, sleep, computer use) can be much more meaningful if it can be compared to population data. This comparison allows for a quick assessment whether a particular condition is normal, for example growth of a baby (Van Buuren 2014). If growth is lagging behind (compared to children with a similar growth curve), it may be decided to give extra nutrition.

In an ideal situation, we aim to collect relevant health parameters from each individual from 10 months before birth to the end of life (Topol 2014), as this will provide insight into the reference population. You could then ask specific questions like: "*What is the impact of premature birth on school performance?*" Or *"How good is my fitness compared to men of the same age? How much do they sleep on average per night, how much exercise do they do?"* By systematically mapping health related parameters into a personal health record and by comparing individual data with similar individuals, it is made possible to assess someone's own health status, but also to deliver a personal prognosis. For this, it is necessary that historical data is available as well. The American Institute of Medicine advocates a continuous learning healthcare system that focuses on the collection of treatment outcomes and

patient-oriented care (Smith et al. 2012). With a well-designed infrastructure, it is possible to collect large-scale observational data, thereby contributing to evidence-based practice in healthcare. Controlled and observational studies fulfil a complementary role (Booth and Tannock 2014).

The more we want to make models personal, the more data we need to combine. In the Netherlands, plans have already been drawn up for a national, even perhaps eventually international, organized data infrastructure[6] that allows learning about distributed data sets for research, which cannot be stored centrally by legislation regarding personal data and security considerations. Maastricht UMC, LUMC and DTL, together with partners, are developing the personal health train infrastructure[7] to make data stored in different locations available for joint analysis. Semantic data interoperability (Wilkinson et al. 2016) is a prerequisite for bringing algorithms and models to the data for a distributed machine learning approach (Damiani et al. 2015). The Prana Data project[8] also perf*orms pilots with data encryption methods (Erkin et al. 2012) that support simple forms of data analysis, for example, computing population averages. We are now in a period of time where techniques have not yet been fully developed and different ideas are being tested. There is currently a pilot project in the province of Limburg with the Personal Health Train. In the city of Rotterdam, a trial is being prepared, on the initiative of Medical Delta to enable people to manage their own health data under the name of My Data Our Health. Performing this type of pilot helps to know in practice how techniques aimed at reducing privacy risks related to data analysis can best be applied. On the one hand, we try to minimize the risks for individuals by providing the citizen or patient with control over who has access to his data. On the other hand, the system architecture must also provide opportunities to conduct studies or generate personalized advice based on the data of people who have given consent.

It is my conviction that making long-term health and treatment information from a large population accessible, in a responsible manner (of course, with the necessary precautions) can help to accelerate the learning ability of our healthcare system. Ideally, physicians and patients can make a decision based on the best quality of life, with these longitudinal reference data as an important complementary source of evidence.

In summary: access to population data can play an important role in personalized advice, but good data governance is necessary to make citizens feel comfortable when sharing data.

2.4 Example: Active lifestyle for wheelchair users (Prediction, Personalization)

An example of intensive use of different reference populations is the project 'From data to action'[9], which will soon start in collaboration between VU Amsterdam, Amsterdam University, Campinas University in Brazil and Leiden University. This consortium is an interdisciplinary collaboration between rehabilitation research, movement science, nutrition science and data science. The purpose of this research is to develop personal exercising and nutritional advice for wheelchair users, using the recorded experiences of similar individuals in a secure database.

People who must use a wheelchair after a spinal cord injury or amputation soon suffer from the effects of insufficient physical exercise. Their activity level drops to 40% of the normal level (Van den Berg-Emons, Bussmann and Stam 2010). This leads to an increased risk of being overweight, having diabetes or cardiovascular disease, resulting in a reduced quality of life (Manns and Chad 1999). For the research, we want to use an existing digital platform consisting of a central database, learning algorithms, and personalized advice for exercise and nutrition applications. This platform is not yet suitable

for wheelchair users because of their specific limitations and capabilities. In the project, therefore, we will first investigate the determinants and factors that promote physical activity and health of wheelchair users. We will work with a systems approach and look at physiology, nutrition, balance between activity and rest, social and physical context and psychological factors. Everyone is different and there are indications that personalized advice is more effective than general advice (Krebs, Prochaska and Rossi 2010). This is certainly the case with wheelchair users with very different medical backgrounds.

Based on the new knowledge, we will adapt the digital platform for wheelchair users to support them in acquiring an active lifestyle. The use of sensors allows the recording of exercise in combination with the registration of relevant outcomes. The contribution of my group is primarily focused on the analysis of the large amounts of sensor data. We are also going to develop predictive models for the success of a specific exercise program, with a large number of input-related factors as input. Finally, we want to personalize the advice based on a self-learning algorithm.

For this recommendation algorithm, we want to use existing wheelchair user measurement data, but also allow the system to learn from new users of the platform. We expect to be able to refine predictive models and recommendations by making comparisons between different groups: wheelchair user versus non-wheelchair user, Amsterdam versus Sao Paulo, beginner athlete versus advanced athlete versus elite athlete. We want to learn from the data which factors contribute to the development of individuals in physical activity and what impeding factors there are.

In summary, this project works on empowerment of wheelchair users by providing the best possible personal advice based on a predictive model building on experiences of similar wheelchair users.

2.5 Example: Patient-forum-miner (Participatory, patient reported outcomes)

In the last example from the health domain, the Patient-Forum-Miner project, the P of participatory care is discussed.

Patients are becoming increasingly aware of their own knowledge and data position. They are more active in managing their own health. An example of this is that patients are communicating through internet discussion forums, with a demonstrable positive effect on their well-being (Van Uden-Kraan et al. 2009; Batenburg and Das 2014). Over the past two years we have conducted projects with the GIST Netherlands contact group, which unites patients with a rare form of gastrointestinal cancer. Because GIST prevalence is so small, there is relatively little knowledge about disease and therapy (Liegl-Atzwanger, Fletcher and Fletcher 2010). In this sense, GIST is a typical orphan disease. Patients from all over the world communicate with each other via a public Facebook group and a private email list. In a collaboration with the GIST contact group, we have archived the email and forum communication channels. Subsequently the messages were semantically indexed and filtered. In addition to the messages of social support, patient experiences are being exchanged, for example tips to reduce side effects of medication. By using text analytics and automatic summarization techniques, the information is made available to patients and cancer experts. New hypotheses can be generated based on statistical analysis of semantically indexed messages (Van Oortmerssen et al. 2017). For example, patients report that it helps to eat chocolate as pure as possible while taking Glivec, in order to reduce nausea. This is, of course, not a causal link, and it should be investigated, for example, if there is a hidden adverse effect, e.g. reduced Glivec's intake. Nevertheless, this experiential knowledge is something that has been unknown to doctors until recently. This project is a good example of citizen science, where patients themselves give guidance to research. I think that such an approach could also have a broader impact. There are many thousands of rare diseases. It is estimated that

7% of the EU population has an orphan disease. In addition, there are also positive results of the application of text mining techniques to messages in on-line communities for reporting new side effects (Sarker et al. 2015). By involving the patients themselves, a lot of additional knowledge can be collected that can supplement the knowledge from clinical trials. In future research, we want to investigate how we can determine the quality of the information in the filtered messages and possibly increase it, for example by linking up with certified sources of medical information.

In summary: By structuring and researching the experiences written by patients, the patient's knowledge position is strengthened. Patients are also better prepared to participate in decisions about care and research options.

3 Data Analytics for policy makers

3.1 Societal value in the context of urban regions

Finally, I want to outline how data analytics can play a role in a completely different domain, in support of urban policy processes aimed at a sustainable future. In recent decades, the liberalization of world trade and the development of the Internet has led to substantial changes in terms of employment, labor productivity, logistics, energy and climate.

It is a huge task for policy makers and governments to initiate a transformation to a sustainable society in cooperation with companies and in view of the interests of citizens. National governments and urban regions increasingly realize that major policy changes are needed to realize that transformation. In the metropolitan area of Rotterdam The Hague, a plan has been made by intensive cooperation between government, market and knowledge partners. Technological developments are an important opportunity to live more sustainable, for example in terms of energy, nutrition and mobility. Similarly as in the field of health, the benefits of investments are not immediately visible but will only pay off in the future. Public private partnerships can help prevent decisions being solely motivated by interests in the short term. Essentially, it is about realizing a transition agenda aimed at sustainable social values through collaboration between various stakeholders

I want to investigate whether a systems approach is possible for this transition and whether principles from the P4 framework can attain a new meaning in this context. Again, we want to quantify the interactions between the various factors and outcome measures based on longitudinal big data. When we use measurements per household for the data, there are opportunities for neighborhood-level aggregation, giving policy makers better insight and forming the basis for evidence-based policies. Such a bottom up approach offers opportunities to promote citizens' participation and make the measurement process more transparent. For example, comparing your monthly average energy usage with the median of the street can give you a lot of insight, but it can only work if fellow citizens contribute to the aggregated data. For such an approach, transparency at all levels of processing and weighing of data is crucial for acceptance.

3.2 Data Analytics for 'real time' policy indicators

This year we will launch a research project into the development of policy indicators based on large data sources in collaboration with the Faculty of Governance and Global Affairs and the Center for Big Data Statistics of CBS (including Leiden University and TNO). By bringing together model driven and data driven disciplines (Janssen and Kuk 2016), we expect a more robust methodology for policy recommendations.

Monitoring of indicators in various policy areas such as employment, mobility, health and safety is part of the standard practice of public administration. After all, policies can only be made based on reliable data. Traditionally, many such data are collected through questionnaires. This approach has disadvantages: the most recent figures often lag far behind on current issues so that , quantitative analyses of interventions

are only made available relatively late. It is therefore difficult to go through a feedback loop in policy processes. In addition, there is always the risk of bias in sampling.

In a big data approach, in general all available information is included in the analysis and can be updated more frequently. An example of such an approach is to analyze open source resources including social media through text mining techniques. These include techniques such as entity recognition, sentiment mining, event recognition. For example, corporate websites can be analyzed on features of job announcements. Other examples of available big data sources are traffic loop data or floating cardata of traffic flows. These resources have already provided some experience that can be used to develop applications for metropolitan practice. Again, data governance and privacy by design are important components in the research.

Our plan is to develop a methodology in the coming years in partnership with Center for Big Data Statistics (CBS with among others UL and TNO) to develop policy indicators for different policy areas in urban regions, beginning with The Hague and the Metropolitan Region of Rotterdam The Hague. Follow-up questions concern: (i) the extraction of new types of information by linking traditional CBS and The Hague data sets; ii) applying a systems approach to quantify the interactions between factors and to predict and visualize the effect of interventions in the best possible way; iii) generalization of the methodology for use in other urban regions in the EU. This research will also be closely linked to the curriculum of the new master ICT in Business focused on the public sector. This curriculum aims at providing future policy advisors and ICT experts working in the sector with an overview of the possibilities of using data analytics for policy. In addition, attention is paid to the legal and ethical framework of the methods.

In summary: There are great opportunities to analyze big data as a basis for evidence-based policies and to work on a learning (eco) system. Research is required into how the new resources can be linked to the current methodology for policy creation and evaluation.

4 Data Science challenges

After describing several societal challenges where, in my opinion, a data driven approach could lead to better choices and priorities, I would like to mention a number of challenges for data science research that will be the focus of my work in multidisciplinary context in Leiden, also in relation to my other employer TNO.

Firstly, I want to develop a system for the collection and long-term storage of data, for validated and robust indicators for health and value systems of sustainable urban areas. This primarily concerns collecting the right data and in some cases combining different modalities. It is important to include contextual information in that data collection. A second challenge is to develop a system architecture to perform analyses and aggregations that respect the rights of data subjects and give them control over who has access to personal data. Examples of these are *health data cooperatives or the personal health train*. A third challenge is to define and validate a similarity function for lifelogs, the longitudinal health data - how can we decide from the raw data which persons are similar to each other given the trends in their health parameters? The fourth challenge is to learn predictive models from longitudinally observed data and their associated uncertainty. One last major interdisciplinary challenge is the application of value based analytics to calculate the effects of different scenarios. I expect that the combination of data across domains and the optimization of the ensemble of value indicators (in policy terms: Integral Policy) has a great potential in comparison with the practice of optimization by policy area.

5 Closure

Ladies and gentlemen, I have explained in my lecture that data, data processing techniques, data science are intrinsically neutral. In this 21st century, internet companies have initiated a transformation by collecting large amounts of data and providing personalized services. In my research, I focus on strengthening the position of individual citizens by deploying data analytics for societal values that benefit society as a whole. I explained that systematically gathering data about, for example, behavior, environment and indicators for values such as health, sustainability, safety and quality of life can lead to new insights. These insights can be created by combining data from different domains and at different aggregation levels. I have illustrated the impact of this approach with several examples. A prerequisite for a successful data analytics approach focused on societal value is that privacy is effectively protected by incorporating privacy and data ownership into the design process.

6 Acknowledgements

I have now arrived at the end of my lecture. I would like to thank several people who have made a significant contribution to the creation of this chair.

First of all, I would like to thank the Board of Leiden University and the Board of the Faculty of Science for appointing me as a professor of applied data analytics.
Of course, I also would like to thank some LIACS colleagues:

Hooggeleerde Kok, dear Joost, it was because of your commitment in 2015, - you were still scientific director of LIACS at the time - that the process of realising this chair was so pleasant and smooth. At our very first meeting at LIACS, I felt that there was a lot of potential in a possible collaboration. We have already achieved several compelling results, and I think that more is yet to come. I also see great opportunities for expanding the Data Science program.

Hooggeleerde Plaat, dear Aske, my appointment came to be in February 2016. You had just started as the new scientific director. I would like to thank you for the fact that you have always spent time to help me to get to know the new organization. I would also like to thank you for your efforts to ensure that this institute has such a pleasant working environment, which is an important building block for excellent research and education.

At LIACS I am working with a large number of colleagues, unfortunately I cannot thank all of them individually. Two colleagues that I would like to mention are Jaap van den Herik and Cor Veenman. *Hooggeleerde* Van den Herik, dear Jaap, thanks to your activities for the Leiden Center of Data Science (LCDS), fruitful collaborations are being initiated. I hope to contribute to some of these activities in the future. Dear Cor, thank you for your commitment to creating a new master course together. Through your experience we can present a wide range of practical cases to the students.
I would also like to thank my other employer TNO. *Hooggeleerde* Keurentjes, *hooggeleerde* Werkoven, representing the TNO Board of Directors and the Board of TNO Technical Sciences. Dear Jos, dear Peter, I would like to thank the TNO organization for providing support to me to build my academic career. I hope that the combination of basic research at LIACS and applied research at TNO will lead to greater societal value.

In my lecture I spoke about the SWELL project of which I have been principal investigator in the context of the COMMIT / program. *Hooggeleerde* Smeulders, *hooggeleerde* Lagendijk, Dear Arnold, Dear Inald, thank you for involving me in building the COMMIT / community in the Netherlands. I owe a lot to it and enjoyed the experience.

There are also TNO colleagues present today as well as many familiar faces from the networks big data and health.

I appreciate the fact that you are present today and I look forward to continuing the cooperation.

I would also like to thank my master and PhD students for their presence. It has been an enriching experience to work with you and to guide you.

Dear friends and family, good to see you here! It gives me a special feeling to give this lecture in this city with a historical link with science but also with the Kraaij family.

Dear father and mother, there are so many things that I learned from you. I want to thank you for your continuous support. I hope you enjoy this day too.
Dear Lyne, Ruben and Gaël, what a beautiful life we have together! Thank you for your support and inspiration.

Ik heb gezegd.

7 References

Abraham, Charles and Susan Michie. 2008. "A Taxonomy of Behavior Change Techniques Used in Interventions." *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association* 27 (3): 379-87. doi:10.1037/0278-6133.27.3.379.

Aizawa, Kiyoharu, Datchakorn Tancharoen, Shinya Kawasaki and Toshihiko Yamasaki. 2004. "Efficient Retrieval of Life Log Based on Context and Content." In *Proceedings of the the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, 22-31. CARPE'04. New York, NY, USA: ACM. doi:10.1145/1026653.1026656.

Bandura, Albert. 1977. "Self-Efficacy: Toward a Unifying Theory of Behavioral Change." *Psychological Review* 84 (2): 191-215. doi:10.1037/0033-295X.84.2.191.

Batenburg, Anika and Enny Das. 2014. "Emotional Approach Coping and the Effects of Online Peer-Led Support Group Participation Among Patients With Breast Cancer: A Longitudinal Study." *Journal of Medical Internet Research* 16 (11): e256. doi:10.2196/jmir.3517.

Berg-Emons, Rita J. van den, Johannes B. Bussmann and Henk J. Stam. 2010. "Accelerometry-Based Activity Spectrum in Persons with Chronic Physical Conditions." *Archives of Physical Medicine and Rehabilitation* 91 (12): 1856-61. doi:10.1016/j.apmr.2010.08.018.

Booth, C.M. and I.F. Tannock. 2014. "Randomised Controlled Trials and Population-Based Observational Research: Partners in the Evolution of Medical Evidence." *British Journal of Cancer* 110 (3): 551-55. doi:10.1038/bjc.2013.725.

Buuren, Stef van. 2014. "Curve Matching: A Data-Driven Technique to Improve Individual Prediction of Childhood Growth." *Annals of Nutrition & Metabolism* 65 (2–3): 227–33. doi:10.1159/000365398.

Damiani, Andrea, Mauro Vallati, Roberto Gatta, Nicola Dinapoli, Arthur Jochems, Timo Deist, Johan van Soest, Andre Dekker and Vincenzo Valentini. 2015. "Distributed Learning to Protect Privacy in Multi-Centric Clinical Studies." In *The 15th Conference on Artificial Intelligence in Medicine*, edited by J. H. Holmes, R. Bellazzi, L. Sacchi, and N. Peek, 65–75. Pavia, Italy: Springer. http://eprints.hud.ac.uk/23905/.

Doherty, A.R. and A.F. Smeaton. 2008. "Automatically Segmenting LifeLog Data into Events." In *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, 20-23. doi:10.1109/WIAMIS.2008.32.

Erkin, Zekeriya, Thijs Veugen, Tomas Toft and Reginald L. Lagendijk. 2012. "Generating Private Recommendations Efficiently Using Homomorphic Encryption and Data Packing." *IEEE Transactions on Information Forensics and Security* 7 (3): 1053-1066.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542 (7639): 115–18. doi:10.1038/nature21056.

Evans, William E. and Howard L. McLeod. 2003. "Pharmacogenomics - Drug Disposition, Drug Targets, and Side Effects." *New England Journal of Medicine* 348 (6): 538-49. doi:10.1056/NEJMra020526.

Evans, William E. and Mary V. Relling. 2004. "Moving towards Individualized Medicine with Pharmacogenomics." *Nature* 429 (6990): 464-68. doi:10.1038/nature02626.

Ford, Daniel A., Reiner Kraft and Gaurav Tewari. 2003. System and technique for dynamic information gathering and targeted advertising in a web based model using a live information selection and analysis tool. US6606644 B1, filed February 24, 2000, and issued August 12, 2003. http://www.google.com/patents/US6606644.

Gemmell, Jim, Gordon Bell and Roger Lueder. 2006. "MyLifeBits: A Personal Database for Everything." *Commun. ACM* 49 (1): 88-95. doi:10.1145/1107458.1107460.

Greef, Jan van der, Thomas Hankemeier and Robert N McBurney. 2006. "Metabolomics-Based Systems Biology and Personalized Medicine: Moving towards N = 1

Clinical Trials?" *Pharmacogenomics* 7 (7): 1087-94. doi:10.2217/14622416.7.7.1087.

Hood, Leroy and Stephen H. Friend. 2011. "Predictive, Personalized, Preventive, Participatory (P4) Cancer Medicine." *Nature Reviews Clinical Oncology* 8 (3): 184-87. doi:10.1038/nrclinonc.2010.227.

Janssen, Marijn and George Kuk. 2016. "Big and Open Linked Data (BOLD) in Research, Policy, and Practice." *Journal of Organizational Computing and Electronic Commerce* 26 (1-2): 3-13. doi:10.1080/10919392.2015.1124005.

Kitano, Hiroaki. 2002. "Computational Systems Biology." *Nature* 420 (6912): 206-10. doi:10.1038/nature01254.

Koldijk, S., M.A. Neerincx and W. Kraaij. 2016. "Detecting Work Stress in Offices by Combining Unobtrusive Sensors." *IEEE Transactions on Affective Computing* PP (99): 1-1. doi:10.1109/TAFFC.2016.2610975.

Koldijk, Saskia, Maya Sappelli, Suzan Verberne, Mark A. Neerincx and Wessel Kraaij. 2014. "The SWELL Knowledge Work Dataset for Stress and User Modeling Research." In *Proceedings of the 16th International Conference on Multimodal Interaction*, 291-298. ICMI '14. New York, NY, USA: ACM. doi:10.1145/2663204.2663257.

Krebs, Paul, James O. Prochaska and Joseph S. Rossi. 2010. "A Meta-Analysis of Computer-Tailored Interventions for Health Behavior Change." *Preventive Medicine* 51 (3-4): 214-21. doi:10.1016/j.ypmed.2010.06.004.

Liegl-Atzwanger, Bernadette, Jonathan A. Fletcher and Christopher D. M. Fletcher. 2010. "Gastrointestinal Stromal Tumors." *Virchows Archiv* 456 (2): 111-27. doi:10.1007/s00428-010-0891-y.

Manns, P.J. and K.E. Chad. 1999. "Determining the Relation between Quality of Life, Handicap, Fitness, and Physical Activity for Persons with Spinal Cord Injury." *Archives of Physical Medicine and Rehabilitation* 80 (12): 1566-71.

McGinnis, J. Michael, Pamela Williams-Russo and James R. Knickman. 2002. "The Case For More Active Policy Attention To Health Promotion." *Health Affairs* 21 (2): 78-93. doi:10.1377/hlthaff.21.2.78.

Meijer, O.C. 2016. *Cortisol van Kop Tot Teen: Over "Goed En Kwaad" van Een Stresshormoon*. Leiden: Universiteit Leiden.

Oortmerssen, Gerard van, Stephan Raaijmakers, Maya Sappelli, Erik Boertjes, Suzan Verberne, Nicole Walasek and Wessel Kraaij. 2017. "Analyzing Cancer Forum Discussions with Text Mining." In *Proceedings of Second International Workshop on Extraction and Processing of Rich Semantics from Medical Texts*. Vienna.

Porter, Michael E. 2010. "What Is Value in Health Care?" *New England Journal of Medicine* 363 (26): 2477-81. doi:10.1056/NEJMp1011024.

Roosendaal, Arnold. 2013. "Digital Personae and Profiles in Law: Protecting Individuals' Rights in Online Contexts." SSRN Scholarly Paper ID 2313576. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2313576.

Ruiter, Robert A. C., Loes T. E. Kessels, Bernadette M. Jansma and Johannes Brug. 2006. "Increased Attention for Computer-Tailored Health Communications: An Event-Related Potential Study." *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association* 25 (3): 300-306. doi:10.1037/0278-6133.25.3.300.

Sappelli, Maya, Suzan Verberne and Wessel Kraaij. 2016. "Adapting the Interactive Activation Model for Context Recognition and Identification." *ACM Trans. Interact. Intell. Syst.* 6 (3): 22:1-22:30. doi:10.1145/2873067.

Sarker, Abeed, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya and Graciela Gonzalez. 2015. "Utilizing Social Media Data for Pharmacovigilance: A Review." *Journal of Biomedical Informatics* 54 (April): 202-12. doi:10.1016/j.jbi.2015.02.004.

Schön, Donald A. 1987. *Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions*. Vol. xvii. Jossey-Bass Higher Education Series. San Francisco, CA, US: Jossey-Bass.

Schork, Nicholas J. 2015. "Personalized Medicine: Time for One-Person Trials." *Nature News* 520 (7549): 609. doi:10.1038/520609a.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484-9. doi:10.1038/nature16961.

Slijkhuis, Jannette Marieke. 2012. "A Structured Approach to Need for Structure at Work." http://www.rug.nl/research/portal/publications/a-structured-approach-to-need-for-structure-at-work(08debcfa-ba7a-42c1-9139-8f7d9702976f).html.

Smith, Mark, Robert Saunders, Leigh Stuckhardt, and J. Michael McGinnis, eds. 2012. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Institute of Medicine. https://www.nap.edu/catalog/13444/best-care-at-lower-cost-the-path-to-continuously-learning.

Topol, Eric J. 2014. "Individualized Medicine from Prewomb to Tomb." *Cell* 157 (1): 241-53. doi:10.1016/j.cell.2014.02.012.

Uden-Kraan, C.F. van, C.H.C. Drossaert, E. Taal, E.R. Seydel and M.F.J. van de Laar. 2009. "Participation in Online Patient Support Groups Endorses Patients' Empowerment." *Patient Education and Counseling* 74 (1): 61-69. doi:10.1016/j.pec.2008.07.044.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (March): 160018. doi:10.1038/sdata.2016.18.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun et al. 2016. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *arXiv:1609.08144 [Cs]*, September. http://arxiv.org/abs/1609.08144.

8 Noten

1 www.wetenschapsagenda.nl/
2 Leiden Institute of Advanced Computer Science liacs.leidenuniv.nl
3 autoriteitpersoonsgegevens.nl/nl/nieuws/1-jaar-meldplicht-datalekken
4 www.swell-project.net
5 Leiden University Medical Center
6 www.dtls.nl/health-ri
7 www.dtls.nl/fair-data/personal-health-train/
8 pranadata.nl
9 NWO: *From Data to Action: Promoting Active Lifestyle in Wheelchair Users with Spinal Cord Injury or Amputation*

Prof.dr.ir. Wessel Kraaij

2016-present	Professor Applied Data Analytics, Leiden University
2008-2016	Professor Information Filtering and Aggregation, Radboud University
2015-present	Principal Scientist, TNO
1995-2014	Senior Scientist, TNO
2004	PhD Computer Science, University of Twente
1999-2000	Visiting researcher, Université de Montréal, Canada
1994-1995	Research Institute for Language and Speech, Utrecht University
1988-1993	Institute for Language Technology and Artificial Intelligence, Tilburg University
1987-1988	Master thesis, Institute for Perception Research - IPO, Eindhoven
1981-1988	Master Electrical Engineering, Technische Universiteit Eindhoven

In 'Data of Value', Wessel Kraaij shows how data analytics techniques and data science can be used in a broader sense to work on societal challenges, such as personalized health care or the transition towards a sustainable society. The systematic collection of everyday personal data, such as lifestyle data, can already help to provide citizens with insight into health risks. Combination of this data about a population can help to approach health from the perspective of prevention, prediction, participation and personalization. It is therefore necessary to gather longitudinal data on various factors that affect health, such as physiology, lifestyle, social environment, mental status and relevant outcomes. The wider applicability of this system approach is illustrated by a case study big data for policy interventions in urban contexts.

Wessel Kraaij focuses his research on developing search technology and finding new insights in large amounts of unstructured information. Initially, that was text and digital video, later it was expanded with sensor data. He uses probabilistic models and machine learning techniques. Currently, he is focusing on the E-health domain: how citizens and patients can contribute to improved health care through the collection and sharing of health data. Citizen data ownership and analysis tools safeguarding privacy are important conditions for this.

Universiteit Leiden
The Netherlands