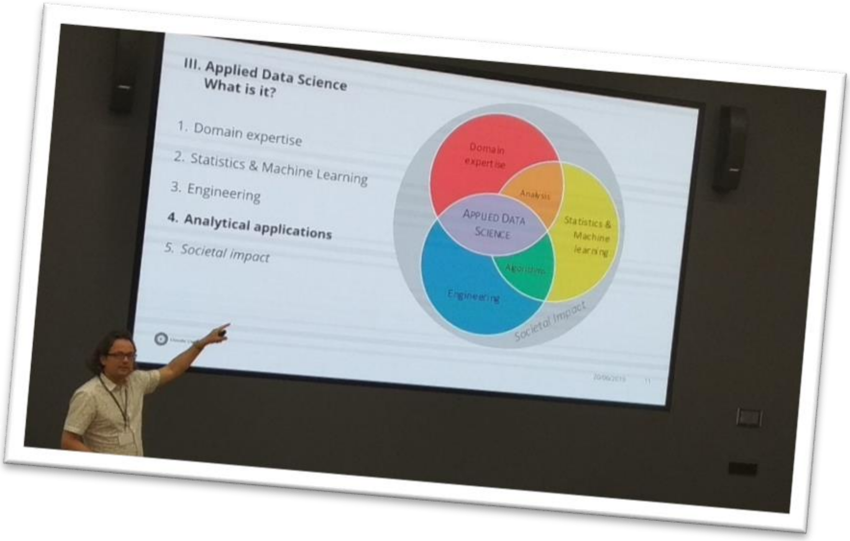


III. Applied Data Science
What is it?

1. Domain expertise
2. Statistics & Machine Learning
3. Engineering
4. **Analytical applications**
5. *Societal impact*

Senior Research Qualification portfolio

"SKOz"

Dr. Marco Spruit :: m.r.spruit@uu.nl :: July, 2019

Abstract

This portfolio showcases my research leadership. First, I introduce Applied Data Science as a separate research discipline and discuss current research challenges and future research directions in Applied Data Science based on my two position papers. Then, I reconstruct my personal research journey and summarise my main scientific contributions thus far. Finally, I formulate the following research objective for the coming years: *I want to establish and lead an authoritative national infrastructure for Dutch natural language processing (NLP) to facilitate and popularise self-service data science.*

Natural Language Processing

Applied Data Science

Preface

I present my Senior Qualification Research (SKOz) portfolio. It was an enlightening experience to fill out the key sections, according to the following documentation that was provided to me:

- Regeling Seniorkwalificatie Onderzoek (SKOz), Faculteit Wiskunde en Informatica (1998).
- Bijlage Lijst van bekwaamheden m.b.t. de SKOz (1998).
- Assessment Form SKOz, Department of Information and Computing Sciences, UU: Internal document for the Assessment Committee (2018).

I aim to facilitate the assessment by adhering to the assessment criteria structure. I have highlighted the key sections in **bold** face in the table of contents below.

Table of Contents

I. Research field qualities

R2: well-founded vision on research field's place and relevance in context

Scientific Embedding

My research field of expertise is Applied Data Science (ADS), which is a subfield within the broader domain of Data Science, which, in turn, is a subfield within Computer Science. Applied Data Science has only recently emerged as a separate discipline, as can also be seen from the relatively few publications that include this term (merely 615 results in Google Scholar as of July 2019).

The first academic mention of the term Applied Data Science can be found in Stadelmann *et al.* (2013). However, its first definition was published by Walker (2015) who defines Applied Data Science as *"[...] a branch of Data Science oriented towards the development of practical applications, technologies and other interventions including engineering practices. Applied Data Science bridges the gap between Basic Data Science and the engineering domains to provide predictable, usable tools to industries including standard methods and practices"*. This is in close alignment with the following definition of Spruit & Jagesar (2016): *"Applied Data Science (ADS) is the knowledge discovery process in which analytical applications are designed and evaluated to improve the daily practices of domain experts"*. The latter publication also visualises the ADS research field, shown in in Figure 2, by building on the widely acknowledged visual definition of Data Science skills by the US National Institute of Standards and Technology (NIST; Prizker & May, 2015), reprinted in Figure 1.

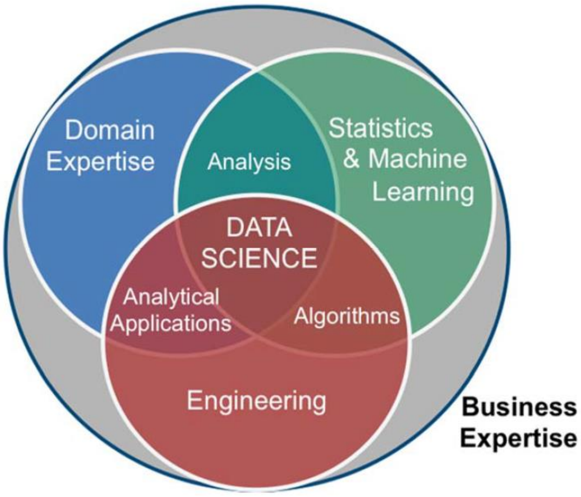


Figure 1: Skills needed in Data Science (Prizker & May, 2015).

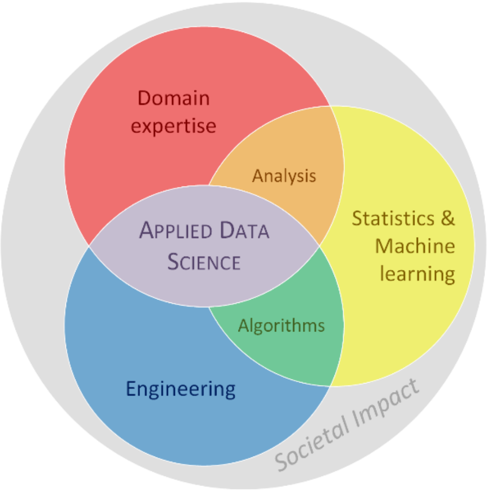


Figure 2: Applied Data Science (Spruit & Jagesar, 2016).

Knowledge Discovery

The added value of Spruit & Jagesar's (2016) definition of Applied Data Science is the explicit scientific relation with the relatively mature research field of Knowledge Discovery, which may have seemingly lost its appeal but which is still foundational to much recent research where the focus is on applying Data Science techniques to solve real-world problems. As noted in Spruit & Jagesar (2016), although more elaborately in Spruit & Lytras (2018), the Knowledge Discovery Process (KDP) *"creates the context for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of business, manufacturing, scientific, and personal information"* (Fayyad *et al.*, 1996b). Even though many process models have been developed, the *CRoss-Industry Standard Process for Data Mining (CRISP-DM)* is widely considered to be the #1 Knowledge Discovery Process guideline in the Data Science industry to perform Applied Research based on 20 years of best practices (Chapman *et al.*, 2000;

KDNuggets.com, 2019). Another influential KDP model is Knowledge Discovery in Databases (KDD; Fayyad *et al.*, 1996), because it is the first model to emphasise the importance of the process itself while performing a data analysis.

Data Science & Citizen Data Science

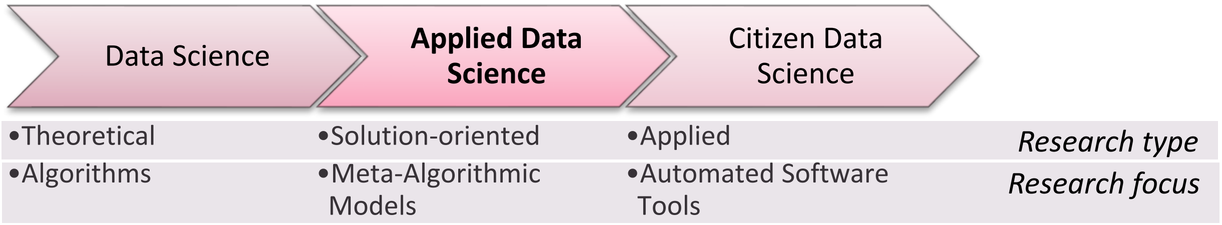


Figure 3: Applied Data Science within a spectrum of related research fields regarding research types and research focus.

Figure 3 places Applied Data Science as a separate research field into a broader context of related research areas with respect to their research types and research focus. On the one hand, Data Science (*i.e.* "Foundational" Data Science) research is often of a primarily theoretical or (applied) mathematical nature, where new algorithms are presented with either mathematical proofs of their correctness or based on a series of computational experiments, thus, in a controlled laboratory setting. On the other hand, Citizen Data Science is most often of a purely applied research nature, using standard data analysis tools including Excel, SPSS or Tableau whenever possible. The primary objective of the data analysis process is to obtain domain-relevant answers to the posed questions. In other words, and in constrast to Data Science, no novel data science techniques will be developed.

However, in Applied Data Science we aim for the best of both worlds. Similar to Citizen Data Science, ADS projects are usually initiated to find a usable solution for a problem that is deemed relevant, timely and urgent within a specific application domain by the domain experts. The primary objective, therefore, is to improve the daily practices of the domain experts. That means that a computational experiment to evaluate the data analysis model(s) is only a small part of the solution. The most time-consuming process steps are getting good-quality, real-world and well-prepared data, which account normally for at least 80% of the available project time, as well as getting the proposed solution accepted for deployment in daily practices, which requires end-user acceptance and regulatory compliance, among others. In other words, ADS addresses the entire cycle of the Knowledge Discovery Process. To codify best practices and quickstart new projects, Spruit & Jagesar (2016) introduce the concept of Meta-Algorithmic Modelling (MAM; See section C1) for improved knowledge reusability and transparancy.

Information Science & Information Systems

In part due to its explicit position in our Department of Information and Computing Sciences, and my current appointment at UU as Associate Professor Information Science, a reflection upon the relation of Applied Data Science to the related field of Information Science is appropriate. Note that this field is quite often also refered to as Information Systems, and I simply consider them to be synonymous throughout this text, although I observe that the term Information Science is becoming increasingly associated with "Librarian Science". Therefore, I prefer to use Information Systems instead.

As Marchionini (2017) notes, *"[...] information science strongly informs data science by considering the entire data life cycle rather than the storage and analytics alone. This end-to-end focus is especially important for the veracity and value components of data science. [...] It could be argued that data science is a subset of information science and some data science training programs may be housed in information schools, however, it is more strategic to view information science as an essential component of data science so that the emerging field can benefit from the diversity of perspectives that interdisciplinary collaborations bring."* Adding to this viewpoint, whereas the field of Knowledge Discovery provides a data-centric process perspective, Information Systems thus provides an ICT-centric process perspective for

Applied Data Science. In my research theme this is captured in the Information Infrastructure component, which is further described in Section C1. Interestingly, Marchionini (2017)'s visualisation of the related disciplines to the Data Science field extends Prizker & May (2015)'s Venn diagramme as shown in Figure 1 by adding the Information Science discipline, as can be seen in Figure 4.

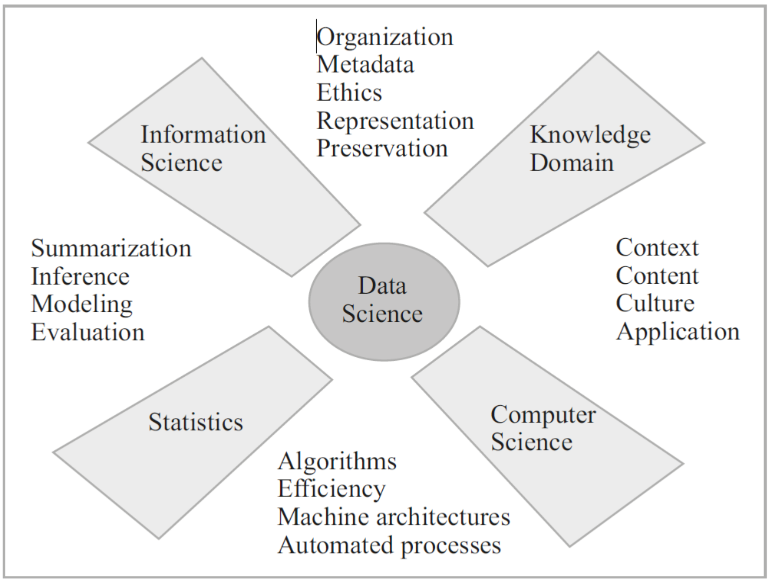


Figure 4: Information Science as a related discipline to Data Science, according to Marchionini (2017).

Utrecht University

From a Utrecht University perspective, the Data Science spectrum in Figure 3 highly resonates with the Department of Information and Computing Sciences (ICS)'s Profile Plan 2017-2020, which also identifies three research types (in alignment to Figure 3): Foundational, Application-oriented, and Applied research. Whereas the latter research type is not considered a useful research approach for ICS, since in that case the research's relevance and outcomes do generally not contribute to the field of ICS itself but to the application domain instead. However, the potential synergy between application-oriented and foundational research is noted, thereby highly conforming to Applied Data Science research objectives.

Moreover, Utrecht Applied Data Science (UADS) was founded in 2017 to bring together researchers from all fields who apply Data Science. The "Starting document for the Focus area of the Utrecht Platform for Applied Data Science" states that *"the definition of applied data science encompasses all applications of data-science methodology and engineering to a scientific domain. Applied data science thereby includes data-science research from scientific domains as well as fundamental research in which (new) methodology and tools are developed and studied from an applications-oriented perspective and in conjunction with one or more domains. In using an inclusive definition of data science, data science also includes legal aspects concerning data, such as privacy protection and ownership"*.

International Community

From a brief historical perspective, one may consider 2017 as the year that the research field of Applied Data Science more broadly established itself, through its prominent appearance at KDD 2017, the 23rd SIGKDD Conference on Knowledge Discovery and Data Mining, which is *"[...] a premier interdisciplinary conference bringing together researchers and practitioners from data science, data mining, knowledge discovery, large-scale data analytics, and big data"* (https://www.kdd.org/kdd2017). At this conference, Applied Data Science was introduced as a key topic for invited talks, invited panels, and separate paper track with best paper award.

Finally, in 2018 I co-edited a special issue for the medium high-impact factor Elsevier journal "Telematics and Informatics" on *Applied Data Science in Patient-centric Healthcare: Adaptive Analytic Systems for*

Empowering Physicians and Patients, which introduces several novel concepts for furthering Applied Data Science research, as described below in sections R1 and C1.

R1: profound knowledge of research field(s) within CS

In this section I briefly review my contributions to the field of Applied Data Science, as a way to demonstrate my expertise regarding this topic, in addition to the contributions which I have mentioned already in Section R2 above as part of the overall Applied Data Science storyline. This overview is based on two position papers: (Spruit & Jagesar, 2016) and (Spruit & Lytras, 2018).

Spruit & Jagesar (2016)

In Spruit & Jagesar (2016) we first define the research field of Applied Data Science in relation to the Knowledge Discovery Process, as shown in Figure 2. Then, we identify the *three main challenges* in correctly applying Machine Learning techniques in Knowledge Discovery projects:

1. *Depth versus breadth:* The ML field knows many different use cases, each of which has a sizeable body of literature surrounding the specific cases. The literature is usually found to be heavy on mathematical terminology and aimed at the computer science community. This prevents researchers from other fields in learning and correctly applying machine learning techniques in their own research (Domingos, 2012).
2. *Selection versus configuration:* In line with the aforementioned, applying machine learning techniques confronts users with many degrees of freedom in how to assemble and configure a learning system. One example of this is the fact that algorithm performance is largely determined by parameter settings, these settings are specific for each class of algorithm. However, in practice end users usually do not have enough knowledge on how to find optimal parameter settings (Yoo *et al.*, 2012). Many users leave the parameters to their default settings and base algorithm selection on reputation and/or intuitive appeal (Thornton *et al.*, 2013). This may lead to researchers using underperforming algorithms and gaining suboptimal results.
3. *Accuracy versus transparency:* Concerning the creation of models, ML findings show that currently there is a trade-off to be had between accuracy and transparency (Kamwa *et al.*, 2012). In practice this means that algorithms which yield a high amount of insight into the data do not perform as well as their non-transparent (black box) counterparts and the other way around.

The third and final contribution of this position paper is its formalisation of the concept of Meta-Algorithmic Modelling (MAM) as a way to improve the transparancy, usability, reproducibility and success rate of the knowledge discovery process. Section C1 describes MAM in more detail.

Spruit & Lytras (2018)

In Spruit & Lytras (2018) we propose a solution-oriented research framework to address the three key dilemmas in the emerging *"post-algorithmic era"* of data science as raised in Spruit & Jagesar (2016), as visualised in Figure 5 below. I elaborate upon nine aspects of the research framework below.

First, the right hand side which visualises the framework's scientific embedding, is the standard *Data Science* skills Venn diagramme by Pritzker & May (2015) as introduced in Section R2. Second, the left hand side is embedded within the six phases of the CRISP-DM cycle as the *de facto Knowledge Discovery Process (KDP)* model (Chapman *et al.*, 2000). Third, the second embedding on the left hand side is the *Design Science Research (DSR)* approach, which is a solution-oriented paradigm that *"seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts"* (Hevner *et al.*, 2004; Wieringa, 2014). Fourth, we align the Design Science cycle and CRISP-DM's Knowledge Discovery process by linking DSR's problem investigation phase to CRISP-DM's domain understanding and data understanding, DSR's treatment design to data preparation and data modelling, and DSR's treatment validation to model evaluation and knowledge deployment.

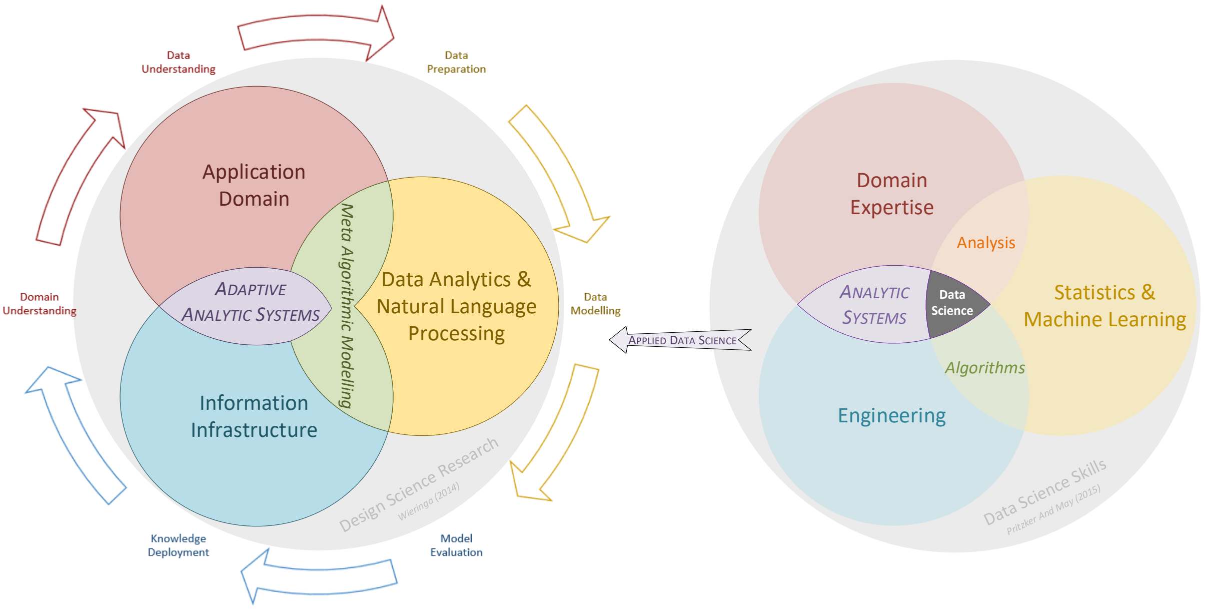


Figure 5: Research framework for Applied Data Science within an application domain (Spruit & Lytras, 2018).

Fifth, from such a methodological viewpoint, then, an Analytic System prototype (in purple) functions as a research intervention instrument, defined as follows: *"An analytic system is a specialised information system for performing analytical tasks based on problem-specific data input characteristics and process preferences"*. Such a prototype is used to evaluate the Design Science artifact under development (*e.g.* a method, model, process, framework, or architecture), employing metrics such as effectiveness, efficiency and usability to determine the analytic system's societal impact. Sixth, we thus conceive *Adaptive Analytic Systems (AAS)* as an engineering approach to investigate intertwining aspects of the three dual phases of the knowledge discovery process with a solution-oriented design science research approach within an applied data science context. The *adaptive* component relates to their focus on data-driven personalisation aspects within such systems. Seventh, the concept of *Meta-Algorithmic Modelling (MAM)* which was introduced in (Spruit & Jagesar, 2016) is now more more embedded to better facilitate and improve the labour-intensive phase of Data Preparation and the crucial step of Data Modelling, as further explained in Section C1.

Eighth, this also applies for the inclusion of Natural Language Processing (NLP) techniques as part of the methodological toolbox of applied data scientists, because in daily practices, the most valuable information is often not available in structured data format, but often as freetext snippets instead. Simon (1960) first formulated this widely acknowledged observation, was further popularised by Grimes (2008) as the "80 percent rule" and still holds true (Das & Kumar, 2013). A well-known example are patients' Electronic Health Records (EHRs) where physicians often still enter insightful information such as diagnoses and daily observations as freetext only. But it is not just because of my background as a computational linguist, that I observe the importance of *NLP for Applied Data Science* in particular.

This NLP4ADS trend has also been confirmed from a non-academic societal impact perspective, by the software industry, which estimates a tripling of the global NLP market within seven years (*e.g.* FutureResearchMarkets.com, 2017). Moreover, it is noteworthily observed that *"NLP can enhance the completeness and accuracy of electronic health records by translating free text into standardized data"*. (ResearchAndMarkets.com, 2019) concurs, as visualised in Figure 6: *"[…] In 2017, the healthcare sector was the most lucrative industrial vertical in overall NLP market. NLP enables a physician to excerpt and summarize information of any drug dosage, symptoms and response data with the purpose of categorizing* *possible side effects of any medicine"*. The importance of these quotes remains somewhat hidden between the lines but is the following:

The added value of Natural Language Processing in Applied Data Science is to integrate structured, semi-structured and unstructured data sources within an information infrastructure that enables more insightful data analyses.

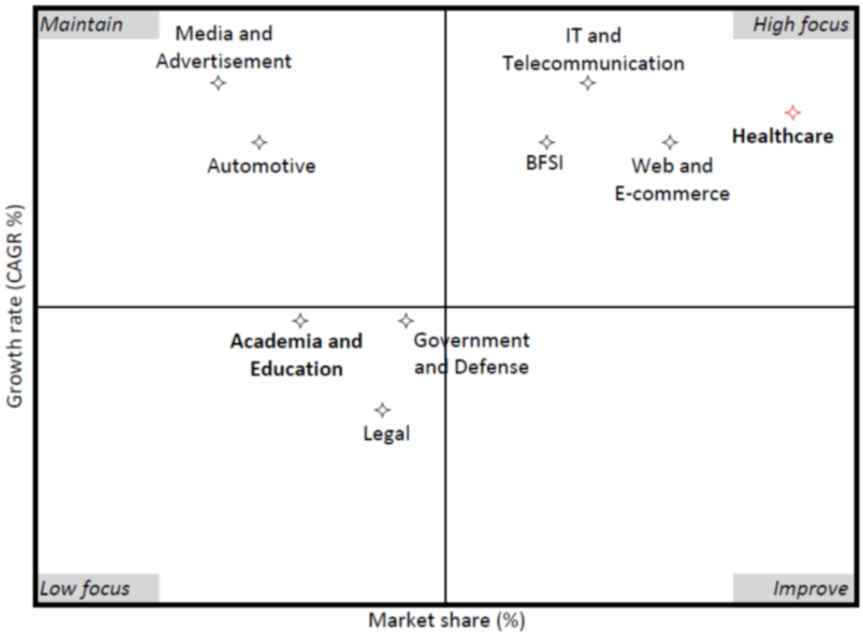


Figure 6: Healthcare is an important NLP application domain (ResearchAndMarkets.com, 2017).

Ninth and finally, as Spruit & Lytras (2018) proclaim, *"[…] we interpret the rise of applied data science and citizen data science as emerging research disciplines as the onset of the post-algorithmic era, where non-data scientists are empowered with automated software tools and meta-algorithmic models to self-service their own data analyses on their own data sources in a reliable, usable and transparent manner"*. Obviously, it is far from trivial to have domain professionals such as physicians and consultants perform their own data analyses in a reliable manner. I consider this fascinating aspect of enabling Self-Service Data Science as the strategic ADS **research problem** which I want to address. The concept of Self-Service Capability is *"to empower non-data scientists with automated software tools and meta-algorithmic models to self-service their own data analyses on their own data sources in a reliable, usable and transparent manner"*. To conclude, as shaped by my two position papers and with respect to my strategic positioning within the research area of Applied Data Science, I have formulated the following overarching ADS **research challenge**:

"How, and to what extent, can we best address the three key dilemmas related to the emerging importance of self-service data science:

1. *Depth versus breadth,*
2. *Selection versus configuration, and*
3. *Accuracy versus transparency?"*

R3: insight into possibilities for inter- and multidisciplinary collaboration

Increasingly, experts from many domains request my research collaboration, in part simply due to the nature of the Appied Data Science field itself, where inter- and multidisciplinary collaborations in other domains are inherently implied.

I have been focusing in particular on collaborations in the Healthcare domain with various research groups within UMC Utrecht, to elicit and utilise their domain expertise. For example, we work with UMCU/Psychiatry (prof. F Scheepers) in the Big Data Psychiatry research programme (PRAISE). With UMCU/Geriatrics (dr. W Knol) we collaborate in various STRIP Assistant RCT-based research projects (OPERAM, STRIMP, OPTICA), in the case of STRIMP also including UMCU/Julius Centre (prof. N de Wit). With UMCU/Cell Biology (prof. J Klumperman) we have worked together since 2012 through a jointly financed PhD student which has resulted in the successful start-up CoreLifeAnalytics BV. With UU/Bioinformatics (prof. B Snel) we jointly supervise a PhD student on research data management in Life Sciences labs. With UMCU/WKZ (prof. M Benders), UU/Social Sciences (dr. R Corten), UU/Education Sciences (prof. S Akkerman), and UU/DGK (dr. M Hostens) we have been jointly supervising master thesis projects, often following up on inspiring guest lectures in my Data Science & Society master course.

Furthermore, UU-external collaborations include RUG/Behavioral Neuroscience (prof. M Kas, BeHAPP), OU/Information Science (prof. R Helms), TUe/Information Systems (prof. U Kaymak, COVIDA), Ministry of Security and Justice/WODC (dr. S Choenni, DATASPACE), Switzerland's UBERN/Internal Medicine (prof. N Rodondi, OPERAM), Switzerland's Fachhochschule Nordwestschweiz/Software engineering (prof. S Fricker, SMESEC), and Norway's Arctic University/Fisheries management (dr. M Borit, SAF21), among many others. Finally, on just as many occasions, collaborations don't take off, mostly due to lack of project funding. This has been the case for UMCU/Julius Centre, UMCU/Cardiology and UU/Pharmaceutical sciences in particular.

Due to the rather long list of active research projects, I sometimes have to decline collaboration requests. Nevertheless, I list below my current preliminary participations in new research collaborations in the Healthcare domain through ZonMW/NWA research proposals which are currently under review (confidential as of July 2019) below (from my website):

1. Building a patient sensitive framework in psychiatry (PASENPSY). (2019/06/05). *Financer(s):* NWA route, Personalized medicine, the individual at the centre. *Applicant(s):* Scheepers,F. et al. *Remark:* First round, Grant total: 4M EUR.
2. Redesigning Mental Healthcare (REMEHE). (2019/06/05). *Financer(s):* NWA route, Health care research, sickness prevention and treatment. *Applicant(s):* Tromp,N. et al. *Remark:* First round, Grant total: 4.5M EUR.
3. Prevention of readmissions in elderly patients with hyperpolypharmacy (HYPPO). (2019/03/04). *Financer(s):* ZonMW Goed Gebruik Geneesmiddelen, UA 8. *Applicant(s):* Kramers,C. et al. *Remark:* First round, Grant total: 348K EUR.
4. Copilot for patients with comorbid Heart Failure and COPD (COPILOT). (2019/04/01). *Financer(s):* IMDI, DCVA-IMDI – PI. *Applicant(s):* Trappenburg,J. et al. *Remark:* First round.

II. Scientific creativity, productivity and recognition by peers

C1: initiated and conducted groundbreaking or at least impactful research

Below I outline my scientific contributions over the years as a mostly chronological journey through the related fields within the overarching research field of Applied Data Science, as shown in Figure 7. Please refer to section R2 for a visual overview which relates all the subfields discussed within this section.

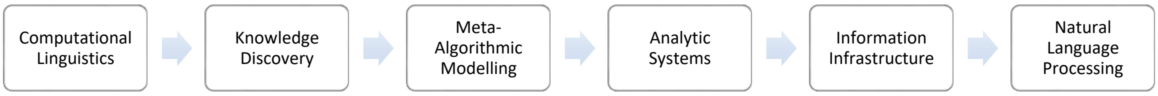

Computational Linguistics
Knowledge Discovery
Meta-Algorithmic Modelling
Analytic Systems
Information Infrastructure
Natural Language Processing

Figure 7: My personal chronological research journey within the field of Applied Data Science.

Computational Linguistics

In 1995 I obtained my master's degree in *Computational Linguistics* on the topic of Unsupervised Learning using *Artificial Neural Networks* (Spruit, 1995) in which I developed the FILTER prototype for Big Data processing in computational experiments which compared fuzzy symbolic versus neural net techniques for information retrieval in a libraries context. Then, during my PhD research I have introduced association rule mining as a *Data Mining* technique to uncover *asymmetrical* relationships among 500+ variables on syntactic variation in the Dutch language area (Spruit, 2007). Additionally, I furthered the state-of-the-art in Dutch language understanding by introducing a multi-level residual-based regression analysis to uncover the underlying associations among aggregate pronunciational, lexical and syntactic variables (Spruit, Heeringa & Nerbonne, 2009). Together with my dissertation, which includes these papers, this research has been cited 150 times as of July 2019, according to Google Scholar.

Knowledge Discovery

Furthering this line of Data Mining research within the realm of Computational Linguistics, I have subsequently introduced the Cross-Industry Standard Process for Data Mining (CRISP-DM) as a best-practice *Knowledge Discovery* process. I have positioned this methodological approach in Spruit (2009) *"[…] as a relevant method in the context of the prestigious Common Language Resources and Technology Infrastructure (CLARIN)"* programme. Next, the importance of holistically focusing on the entire knowledge discovery process in order to obtain optimal solutions (by effectively mitigating "the weakest link" impact of any step within the whole discovery process), led to various contributions to the body-of-knowledge in the field of knowledge discovery. Recent contributions include tailored Knowledge Discovery models for Distributed Computing Workflows (Spruit & Meijers, 2019) and Outsourcing contexts (Ooms, Spruit & Overbeek, 2019). Furthermore, in various contributions we performed exploratory data analyses in unchartered application domains such as Long-term *Healthcare* (Spruit, Vroon & Batenburg, 2014), Surgical Healthcare (Spruit & Rijnst, 2019) and Psychological Healthcare (Eskes, Spruit, Brinkkemper, Vorstman & Kas, 2016).

Meta-Algorithmic Modelling

Vleugel, Spruit & Daal (2010) introduce a novel process modelling viewpoint to facilitate the application of the knowledge discovery process for less experienced data analysts based on Process Deliverable Diagrams (PDDs; Weerd & Brinkkemper, 2008), which in (Pachidi, Spruit & Weerd, 2014) was first mentioned as the concept of *Meta-Algorithmic Modelling* (MAM) and which was further defined in (Spruit & Jagesar, 2016) as *"[…] an engineering discipline where sequences of algorithm selection and configuration activities are specified deterministically for performing analytical tasks based on problem-specific data input characteristics and process preferences"*. A recent example of a MAM receipe can be found in Figure 8 (Menger, Spruit, Klift & Scheepers, 2019), where a meta-algorithmic model is presented to facilitate domain experts and data analysts to effectively apply cluster ensembles to more robustly identify psychiatric patient subgroups. Meta-algorithmic Modelling can be viewed as a key contribution for a trustworthy realisation of *Self-Service Data Science*, in which the aim is to empower non-data scientists (*e.g.* medical professionals, business owners, citizens) with *"automated software tools and meta-algorithmic models to self-service their own data analyses on their own data sources in a reliable, usable and transparent manner"* (Spruit & Lytras, 2018).

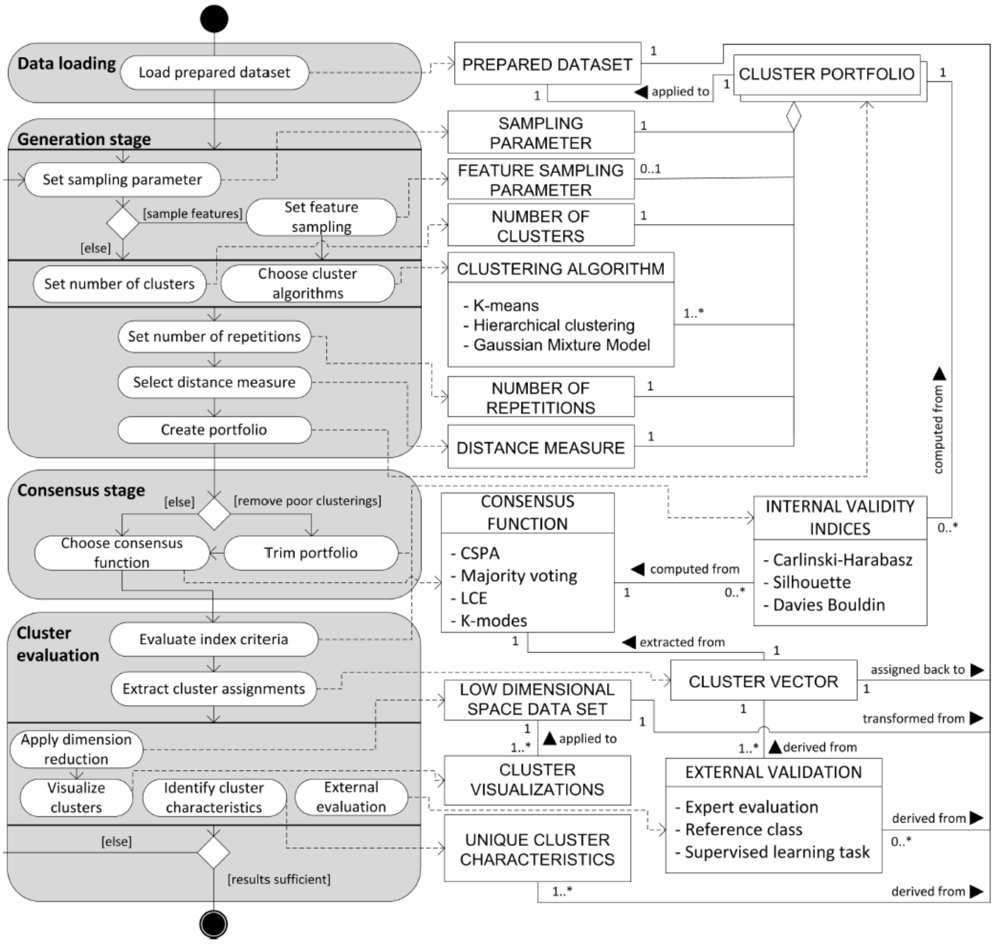


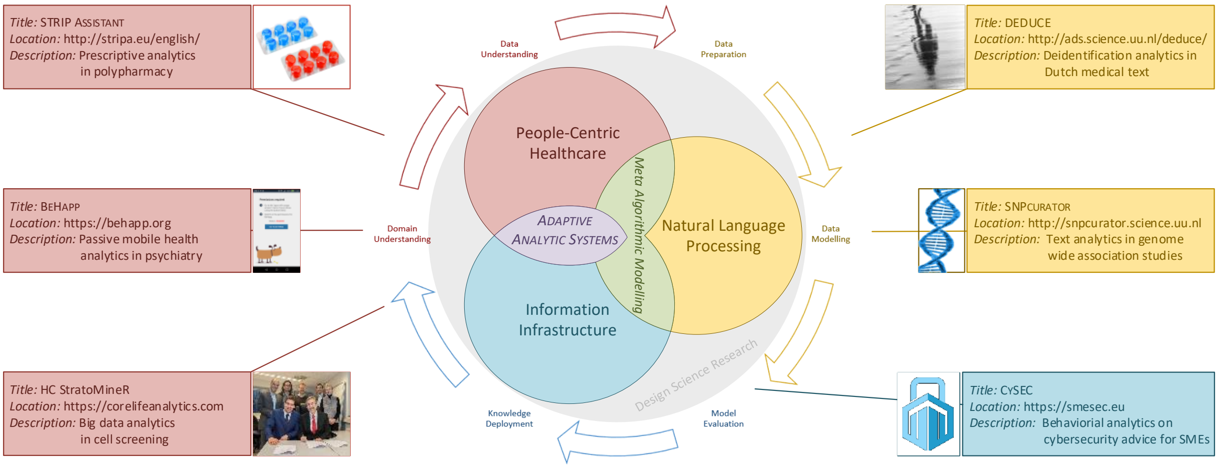
Figure 8: Part of the Meta-Algorithmic Model for cluster ensemble modelling and evaluation (Menger et al., 2019).

Maturity Modelling

A related modelling perspective for organisational process improvement is *Maturity Modelling*. Compared to a meta-algorithmic model, which provides a guideline of ordered advices for reliably and transparantly performing analytic tasks within an organisation, a maturity model provides incremental process improvement advices to improve an organisation in any functional domain. Another key distinction is the implicitly data-driven, quantitative nature of meta-algorithmic modelling which contrasts with the knowledge-based, qualitative approach in maturity modelling. Therefore, Maturity Models can also be used to kickstart organisational improvement and data gathering processes. Utrecht University is well-known for its innovative research on Focus Area Maturity (FAM) models (*e.g.* Steenbergen, Bos, Brinkkemper, Weerd & Bekker, 2010), which improve upon the Capability Maturity Model (CMM) in application domains such as Software Product Management (*e.g.* the Situational Assessment Method (SAM); Bekkers & Spruit, 2010) and Information Security (*e.g.* Information Security Focus Area Maturity model (ISFAM); Spruit & Roeling, 2014). My scientific contributions in this field address the key limitation of maturity models in general: the rigid oversimplication of the real world which evidently is *not* one-size-fits-all. Therefore, in the ISFAM research strand we have refined FAMs by determining the organisational characteristics profiles of the ISFAM users (Mijnhardt, Baars & Spruit, 2016) and analysing the influence that these organizational characteristics have on individual focus areas within the ISFAM, resulting in the novel concept of *Adaptive Maturity Modelling* (Baars, Mijnhardt, Vlaanderen & Spruit, 2016). Ozkan & Spruit (2019) contribute further enrichments in the questionnaire structure and scoring metrics within the context of the Horizon2020 SMESEC project.

Analytic Systems

In many evaluations of our modelling artifacts we have observed that domain experts require more structured assistance than a graphical MAM receipe on paper. To empower domain experts and other non-data scientists to explore their own data analyses, a decision support system is in fact required to satisfactorily walk the user through the meta-algorithmic model. Put somewhat bluntly, no analysis script is likely to ever empower a domain expert to perform an analysis on their own dataset, no matter how elaborate the code comments and documentation. This is why Spruit & Lytras (2018) define an *Analytic System* as *"a specialised information system for performing analytical tasks based on problem-specific data input characteristics and process preferences"*, and subsequently introduce the concept of *Adaptive Analytic Systems* (AAS) which can be thought of as extended and user-friendly analysis scripts or webapps, implementing a significant portion of the knowledge discovery process including a meta-algorithmic model. Tawfik & Spruit (2018) provide a relatively barebones example with SNPcurator, which facilitates medical researchers to quicky explore medical literature in the PubMed repository by interactively providing automatically extracted single-nucleotide polymorphism (SNP) associations of any given disease and its reported statistical significance, odd ratio as well as cohort information such as size and ethnicity. Other examples are shown in Figure 9. Finally, as an example to further clarify the user-friendliness assessment, (Meulendijk, Spruit, …, 2015) evaluate STRIPA's GUI user-friendliness by performing a mixed-methods usability evaluation, consisting of interviewing potential users, having iterative and interactive prototype testing sessions, and measuring its performance with the System Usability Scale (SUS). In subsequent work related to (Shen, Meulendijk & Spruit, 2016), a Post-Study System Usability Questionnaire (PSSUQ), and a 'think aloud' method were also performed to further improve our understanding of the system's usability.


Title: STRIP ASSISTANT
Location: http://stripa.eu/english/
Description: Prescriptive analytics in polypharmacy

Title: BEHAPP
Location: https://behapp.org
Description: Passive mobile health analytics in psychiatry

Title: HC StratoMineR
Location: https://corelifeanalytics.com
Description: Big data analytics in cell screening

Title: DEDUCE
Location: http://ads.science.uu.nl/deduce/
Description: Deidentification analytics in Dutch medical text

Title: SNPCURATOR
Location: http://snpcurator.science.uu.nl
Description: Text analytics in genome wide association studies

Title: CYSEC
Location: https://smesec.eu
Description: Behaviorial analytics on cybersecurity advice for SMEs

Data Understanding
Data Preparation
Domain Understanding
Data Modelling
Knowledge Deployment
Model Evaluation
People-Centric Healthcare
Natural Language Processing
Information Infrastructure
ADAPTIVE ANALYTIC SYSTEMS
Meta Algorithmic Modelling
Design Science Research

Figure 9: Examples of Analytic Systems that are being developed in my Applied Data Science Lab.

Information Infrastructure

As an example of a more mature Adaptive Analytic System, with current Technology Readiness Level (TRL) 6, the STRIP Assistant (STRIPA) implements the Systematic Tool to Reduce Inappropriate Prescribing (STRIP) guideline, similar to a meta-algorithmic model. STRIPA is a web-based clinical decision support system that advices physicians during the pharmacotherapeutic analysis of patients' health records (Meulendijk, Spruit, Jansen, Numans & Brinkkemper, 2015). STRIPA has been proven to be effective (Meulendijk, Spruit, …, Knol, 2015) and efficient (Meulendijk, Spruit…, 2016), and has been the intervention instrument in the large-scale multinational, multilingual Horizon2020-funded OPERAM randomised controlled trial (RCT; Shen, Meulendijk & Spruit, 2016; Adam *et al.*, 2019). The adaptive component in STRIPA is operationalised through developing a risk-aware model for association rule mining (Meulendijk, Spruit & Brinkkemper, 2017). In two nationally funded spinoff projects (STRIMP and

OPTICA) STRIPA is our Analytic System of choice to contribute to the state-of-the-art in the under-researched *Knowledge Deployment* phase within the knowledge discovery process, *i.e.* Implementation Science for Knowledge Discovery. The continual focus on the actual deployment of analytic systems in daily practices, naturally places this research within the well-defined realm of *Information Infrastructure*, which Borgman (2010:19) defines as “[…] the technical, social, and political framework that encompasses the people, technology, tools, and services used to facilitate the distributed, collaborative use of content over time and distance”.

To briefly recapitulate the chronological research journey above, I made my first scientific contributions in the field of Computational Linguistics by introducing novel data analysis techniques, which I subsequently embedded within a more holistic and application-oriented Knowledge Discovery context, where I focus on the empowerment of domain experts by contributing meta-algorithmic and maturity models which are implemented in analytic systems to enable the expert’s self-service capability. The latter requires particular attention to the knowledge deployment phase and aligns closely to the already-established discipline of information infrastructure .

Natural Language Processing

The research strands above have been coming together after observing that in many daily practices, the required data for performing analytical tasks within a knowledge discovery process context is available in both structured and unstructured formats. For example within the domain of Healthcare, even though all information is entered and maintained in Electronic Patient Records (EPRs) within various Information Systems as required by law, the most valuable information—such as a patient’s medical diagnosis—is often only available in (semi- or) unstructured freetext format, making it unavailable for automated data analysis. Therefore, within the context of Phase 3 of the knowledge discovery process (*i.e.* Data Preparation), in recent years I have been completing my personal research journey by coming full circle, returning home to the field of Computational Linguistics, which, when focused in particular on real-world applications instead of the computational analysis of latent linguistic structures, is more commonly refered to as *Natural Language Processing* (NLP).

NLP can be pursued from both a symbolic (rule-based) or probabilistic approach to automatically derive or learn lexical and structural preferences from corpora (Manning and Schütze, 2001). I have contributed to both linguistic perspectives, and aim to further contribute in particular by integrating these NLP approaches in search for “the best of both worlds”. Recent rule-based approaches apply NLP for pattern matching to automatically de-identify Dutch medical texts (DEDUCE; Menger, Scheepers, Wijk & Spruit, 2018), and for processing internal business policies (Spruit & Ferati, 2019), among others. Recent Deep Learning-based NLP research includes work on contextualized word embeddings in Open Information Retrieval (Sarhan & Spruit, 2019) and on recognition of textual entailment in the biomedical domain (Tawfik & Spruit, 2019a). The latter work develops a Deep Learning architecture which embeds Machine Learning features representing linguistic (*i.e.* symbolic) features, effectively integrating representations of lexical, contextual and compositional semantics into the neural network. We further refined this hybrid approach in (Tawfik & Spruit, 2019b) during our participation in the Association for Computational Linguistics (ACL) 2019 MEDIQA Challenge, in which we ended up within the final Top 10.

In Section P1 I provide a showcase example of the strategic direction of my research which synthesises the work described above.

C2: substantial set of publications in international peer-reviewed venues

Figure 10 shows an overview of my publications based on Google Scholar data. Furthermore, I list a selection of recent high impact-journal and top-tier conference publications. Please refer to my CV for a complete record of my publications.

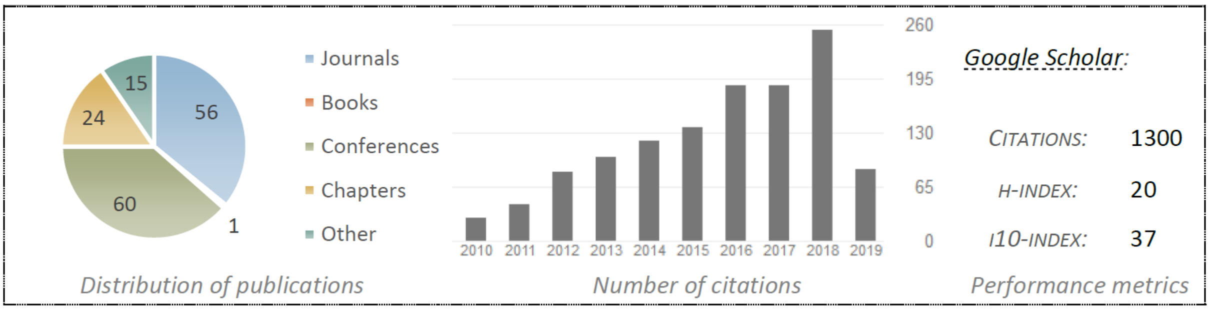


Figure 10: Performance metrics of my publications.

Impact journal publications

I have published 30 ISI impact journal articles so far, for which I list my favorite Top 5 of the last 5 years. Please refer to my CV for a complete listing.

[1] **Spruit,M.**, & Lytras,M. (2018). Applied Data Science in Patient-centric Healthcare: Adaptive Analytic Systems for Empowering Physicians and Patients. *Telematics and Informatics, 35*(4), Special Issue: Patient Centric Healthcare, 643–653. **[IF: 3.398]** [pdf] [online] [17 cites]

[2] Menger,V., Scheepers,F., & **Spruit,M.** (2018). Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text. *Applied Sciences, 8*(6), Data Analytics in Smart Healthcare, 981. **[IF: 2.217]** [pdf] [online] [4 cites]

[3] Syed,S., Borit,M., & **Spruit,M.** (2018). Narrow lenses for capturing the complexity of fisheries: A topic analysis of fisheries science from 1990 to 2016. *Fish and Fisheries, 19*(4), 643–661. **[IF: 9.013]** [pdf] [online] [9 cites]

[4] **Spruit,M.**, Heeringa,W., & Nerbonne,J. (2009). Associations among linguistic levels. *Lingua, 119*(11), The forests behind the trees, 1624–1642. **[IF: 0.578]** [pdf] [url] [60 cites]

[5] Meulendijk,M., **Spruit,M.**, Drenth-van Maanen,C., Numans,M., Brinkkemper,S., Jansen,P., & Knol,W (2015). Computerized decision support improves medication review effectiveness: an experiment evaluating the STRIP Assistant's usability. *Drugs & Aging, 32*(6), 495–503. **[IF: 2.769]** [pdf] [online] [43 cites]

To clarify, [1] is my strategic position paper which introduces Applied Data Science in Healthcare for the empowerment of physicians, advocating the concept of Self-Service Data Science as a consequence of the onset of the "Post-Algorithmic Era". [2] illustrates my interest in comparing and combining different approaches to NLP for smarter Healthcare. [3] is the Top journal in the application domain of Fisheries and represents my highest impact factor publication on NLP. [4] has become a rather influential publication in the domain of Computational Linguistics for Dutch language understanding. [5] is a landmark publication on the effectiveness of the STRIPA Analytic System, my longest running Healthcare project line (since 2009).

Recent top conference publications

I have 60 conference proceedings so far, for which I compiled my favorite Top 5 of the last 5 years. Please refer to my CV for a complete listing.

[1] Tawfik,N., & **Spruit,M.** (In press). UU_TAILS at 2019 MEDIQA Challenge: Learning Textual Entailment in the Medical Domain. *Proceedings of the BioNLP 2019 workshop*. Florence, Italy. [online]

[2] Menger,V., **Spruit,M.**, Klift,W. van der, & Scheepers,F. (2019). Using Cluster Ensembles to Identify Psychiatric Patient Subgroups. *17th Conference on Artificial Intelligence in Medicine*. (pp. 252–262). AIME 2019, Poznan, Poland, June 26-29, 2019: Springer. [pdf] [online]

[3] Syed,S., & **Spruit,M.** (2017). *Full Text or Abstract - Examining Topic Coherence Scores Using Latent Dirichlet Allocation*. 4th IEEE International Conference on Data Science and Advanced Analytics (pp. 165–174). DSAA 2017, Oct 19-21, 2017, Tokyo, Japan: IEEE. [pdf] [16 cites]

[4] Meulendijk,M., **Spruit,M.**, & Brinkkemper,S. (2017). *Risk mediation in association rules: the case of decision support in medication review*. In Teije,A. ten, *et al.* (Eds.), LNAI 10259, 16th Conference on Artificial Intelligence in Medicine (pp. 327 ff). AIME 2017, June 21-24, Vienna, Austria: Springer. [pdf] [5 cites]

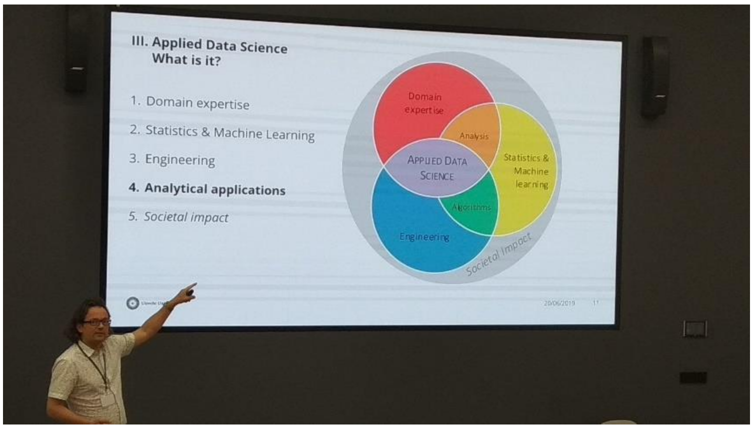
[5] **Spruit,M.**, & Jagesar,R. (2016). *Power to the People! Meta-algorithmic modelling in applied data science*. In Fred,A. *et al.* (Ed.), Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (pp. 400–406). KDIR 2016, November 11-13, 2016, Porto, Portugal: ScitePress. [pdf] [online] [14 cites]

To clarify, [1] describes our participation in the Association for Computational Linguistics (ACL) challenge in the biomedical NLP domain on natural language inference and question answering benchmarking 30 different Deep Learning architectures. [2] introduces Meta-Algorithmic Modelling at the AIME top conference in my field. [3] is quickly becoming a very popular paper which investigates configurations and implications of a critical hyperparameter in NLP: the (required amount of) freetext. [4] introduces an Adaptive Analytic System at the AIME top conference in my field. [5] first defined Applied Data Science and Meta-Algorithmic Modelling as emerging and related research fields.

C3: invited or keynote presentations at symposia, workshops, conferences

I have given invited talks at the following occasions:

- *Applied Data Science for Student Empowerment* (20/06/2019). International Distance Education Conference (DisCo), Prague, Czech Republic, at Microsoft [60 min]. See also Figure 11.
- *Applied Data Science masterclass* (16/04/2019). Rotterdam, ESRI Nederland [5*45 min].
- *Applied Data Science masterclass* (19/12/2019). Rotterdam, ESRI Nederland [5*45 min].
- *Health Analytic Systems in Applied Data Science: Design science for societal impact* (4/12/2017). Monthly Monday Morning Talk, AMC/Clinical Informatics department, Amsterdam [60 min].
- *Establishing Infrastructures for Big Data Research: Design for Societal Impact* (8/1/2016). Workshop: Exposome and Big Data on Geospatial Exposure and Health, Figi conference centre, Zeist [30 min].
- Towards healthcare business intelligence in long-term care: An explorative case study in the Netherlands (05/03/2015). Dutch Healthcare Authority (NZa), Utrecht, Netherlands [50 min].
- *PSGF: The Pricing Strategy Guideline Framework for SaaS Vendors* (16/09/2014). How to Price My Saas? Bootcamp, Brussels, Belgium [120 min].



III. Applied Data Science
What is it?

1. Domain expertise
2. Statistics & Machine Learning
3. Engineering
4. **Analytical applications**
5. *Societal impact*

Figure 11: Photo taken during my DisCo keynote on Applied Data Science (Prague, 20 June 2019).

C4: other recognition among peers, including committees, boards, citations

Below is a selection of recent academic services which I performed. See my CV for a complete listing.

Journals: Special issue editor

- 2018: Special Issue on Patient-centric business intelligence and data analytics systems in healthcare. **Spruit,M.** & Lytras,M. (Eds.), *Telematics and Informatics*, Elsevier. **[IF: 3.398]**
- 2017: Special Issue on Advanced Software Engineering for Data Mining in Business, Health, Education and Social Networks (B-H-E-SN). Lytras,M., **Spruit,M.**, Mathkour,H., & Yáñez-Márquez,C. (Eds.), *IET Software, 11*(5), IET. **[IF: 0.473]**

Journals: Associate editor

- 2017-2012: Decision Analytics, Springer Verlag.
- 2017-present: Int. J. of Semantic Web and Information Systems (IJSWIS), IGI Global. **[IF: 1.5]**

Journals/Book series: Editorial board member

- 2017: 73-volume book series on Advances in Business Information Systems and Analytics (ABISA), IGI Global, edited by prof. Madjid Tavana.
- 2013-present: J. of Computer Information Systems (JCIS), Taylor & Francis. **[IF: 0.822]**
- 2012-present: Int. J. of Business Intelligence Research (IJBIR), IGI Global.

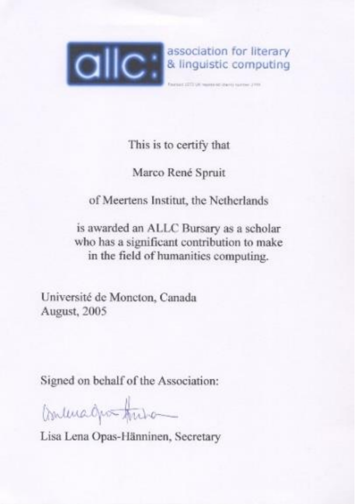
Conferences: Associate editor

- ICIS 2019-2016: International Conference on Information Systems: IS in healthcare track. **[IF: A+]**
- ECIS 2018-2009: European Conference on Information Systems: Digital Health Initiatives track. **[IF: A]**

Conferences: Programme committee

- WWW 2018-2017: 27th WWW conference, Lyon, France: Cognitive Computing track. **[IF: A+]**
- WWW 2017: 26th WWW conference, Perth, Australia: Cognitive Computing track. **[IF: A+]**
- HEALTHINF 2018: 11th International Conference on Health Informatics, Funchal, Portugal.

To conclude, I was awarded an Association for Literary and Linguistic Computing (ALLC) *PhD Bursary* Award of **EUR 750** in as "a scholar who has a significant contribution to make in the field of Humanities Computing", as shown in Figure 12.



allc: association for literary & linguistic computing

This is to certify that

Marco René Spruit

of Meertens Institut, the Netherlands

is awarded an ALLC Bursary as a scholar
who has a significant contribution to make
in the field of humanities computing.

Université de Moncton, Canada
August, 2005

Signed on behalf of the Association:

Lisa Lena Opas-Hänninen, Secretary

Figure 12: ALLC award certificate.

P1: well-founded vision on the strategic direction of his research program

To exemplify my strategic research direction for the coming years, I introduce the recently awarded project which I conceived and lead and that ties all my research strands, as described above in Section C1, together into one highly relevant research niche: COVIDA. This is an acronym for "**Co**mputing **Vi**sits **Da**ta" for Dutch Natural Language Processing in Mental Healthcare, and is now sponsored through the UU-UMCU-TUE Strategic Alliance Fund with 500K. In this project, which is conceptually visualised in Figure 13, I collaborate with TUE's Information Systems department (prof. Uzay Kaymak and dr. Pieter van Gorp) and UMCU's Psychiatry department (prof. Floortje Scheepers and Karin Hagoort).

COVIDA aims to develop a hybrid Dutch language model to better understand human language in general, and Dutch Mental Healthcare language use in particular. We operate within the Design Science Research paradigm to model our computational experiments from both Computational Linguistics (*i.e.* knowledge-based, reasoning-oriented) and Machine Learning (*i.e.* data-driven, learning-oriented) inspired representations through Meta-Algorithmic Models. *COVIDA also delivers an Information Infrastructure* *deploying a Self-Service Data Science facility for Natural Language Processing (NLP) of already routinely collected Dutch medical texts*, thereby enabling healthcare professionals throughout the Dutch language area to reuse their daily clinical notes by nurses and doctors from patients' EHRs to predict inpatient violence risk assessment, depression, and more.

Therefore, COVIDA effectively covers all subfields discussed in Section C1, thus making it an accurate and strategic illustration for my forthcoming research plans in my research theme *Model-Driven Analytic Systems for Self-Service Data Science*.

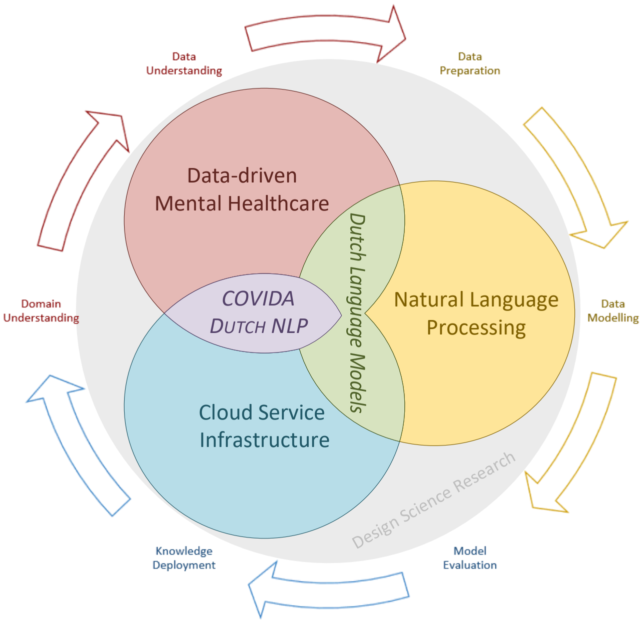


Figure 13: "Computing Visits Data" for Dutch Natural Language Processing in Mental Healthcare (COVIDA) overview.

Introspective

To conclude, I briefly recapitulate several considerations which have supported my decision to pursue this strategic research direction. First, I focus on *NLP* since I have a solid and certifiable background (MA, PhD) in this field, which is an important criterion in research grant proposals. Second, by focusing on *Dutch* language processing in particular, I narrow down my research niche, and thus, the number of direct heavy-weight competitors such as Stanford, MIT and the "Big Five Tech Companies". Furthermore, I implicitly capitalise on my Dutch native skills. Third, my strategic application domain is *Healthcare*, where I am particularly interested in understanding clinical notes within EHRs. I already have a strong and successful Applied Data Science portfolio in the Healthcare domain, including published research in Clinical NLP onto which I can further improve upon (*e.g.* Menger *et al.*, 2018-2019; Tawfik & Spruit, 2019). With my differentiating track record on deploying Knowledge Discovery innovations in daily practices I believe that through COVIDA I can also achieve my 3-5 years objective:

> I want to establish and lead an authoritative national infrastructure for Dutch natural language processing (NLP) to facilitate and popularise self-service data science.

This research objective aims to empower all kinds of domain professionals and even citizens to maximise the societal impact of natural language processing technologies throughout daily practices: *The Dutch language area is our Lab!* More details are available in Spruit and Lytras (2018).

On a final note, through COVIDA I can now finally pursue my original Knowledge Discovery plans in alignment with the *Common Lab Infrastructure for the Arts and the Humanities (CLARIAH, formerly*

CLARIN) as already proposed in Spruit (2009). My strategic aim will be to extend the CLARIAH Social Science and Humanities information infrastructure with Healthcare capabilities, which would potentially allow me to piggyback towards a Europe-wide deployment of my envisioned long-term authoritative infrastructure for European natural language processing (NLP) to facilitate and popularise *Self-Service Data Science*.

P2: from initiation to successful conclusion of at least one research program

I have already completed several research programmes, which I define here as a PhD student or postdoc research project. Figure 14 shows most of the **24** projects — representing a grand total of **~EUR 2,500,000** in allocated research resources — that have driven my research efforts over the years. More specifically, these efforts have contributed **>1.2M EUR** to my department's budget. See also the illustration below, where shaded boxes indicate already completed several research programmes.

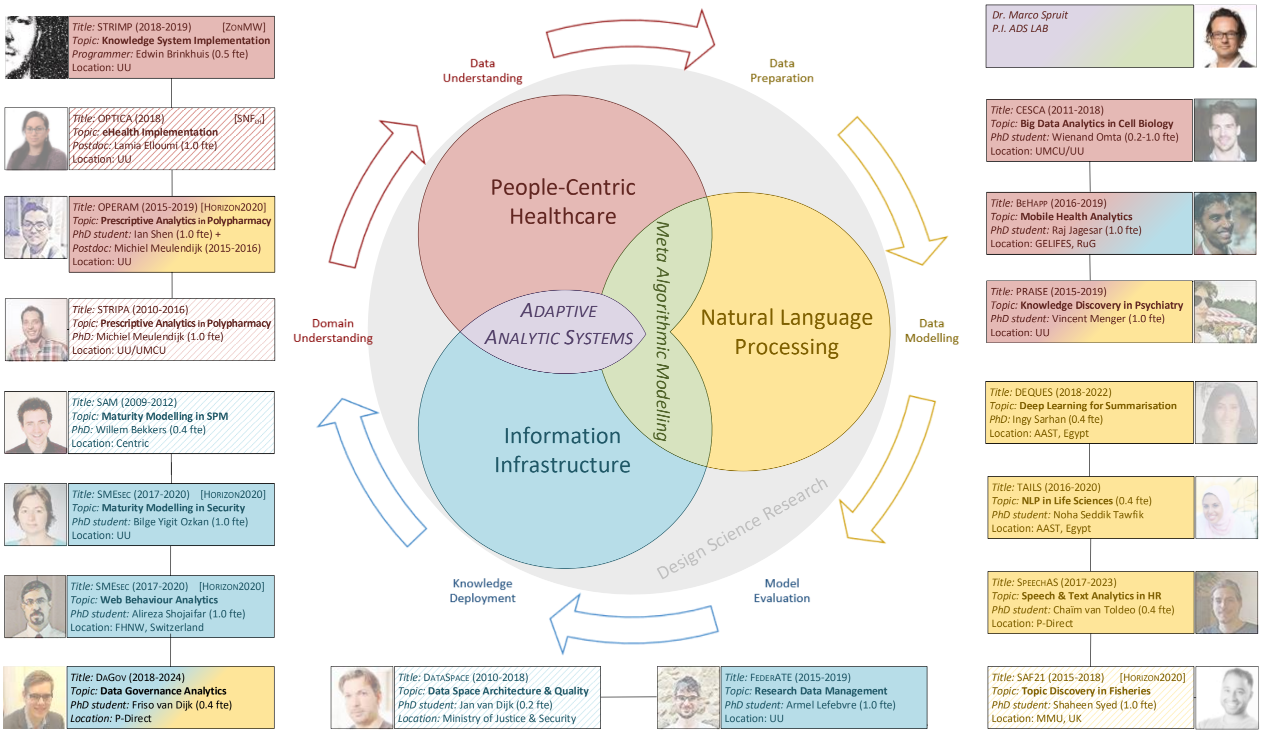

Title: STRIMP (2018-2019) [ZonMW]
Topic: Knowledge System Implementation
Programmer: Edwin Brinkhuis (0.5 fte)
Location: UU
Title: OPTICA (2018) [SNFCH]
Topic: eHealth Implementation
Postdoc: Lamia Elloumi (1.0 fte)
Location: UU
Title: OPERAM (2015-2019) [Horizon2020]
Topic: Prescriptive Analytics in Polypharmacy
PhD student: Ian Shen (1.0 fte) +
Postdoc: Michiel Meulendijk (2015-2016)
Location: UU
Title: STRIPA (2010-2016)
Topic: Prescriptive Analytics in Polypharmacy
PhD: Michiel Meulendijk (1.0 fte)
Location: UU/UMCU
Title: SAM (2009-2012)
Topic: Maturity Modelling in SPM
PhD: Willem Bekkers (0.4 fte)
Location: Centric
Title: SMEsec (2017-2020) [Horizon2020]
Topic: Maturity Modelling in Security
PhD student: Bilge Yigit Ozkan (1.0 fte)
Location: UU
Title: SMEsec (2017-2020) [Horizon2020]
Topic: Web Behaviour Analytics
PhD student: Alireza Shojaifar (1.0 fte)
Location: FHNW, Switzerland
Title: DaGov (2018-2024)
Topic: Data Governance Analytics
PhD student: Friso van Dijk (0.4 fte)
Location: P-Direct
Data Understanding
Data Preparation
Domain Understanding
Data Modelling
Knowledge Deployment
Model Evaluation
People-Centric Healthcare
Natural Language Processing
Information Infrastructure
Adaptive Analytic Systems
Meta Algorithmic Modelling
Design Science Research
Title: DataSpace (2010-2018)
Topic: Data Space Architecture & Quality
PhD student: Jan van Dijk (0.2 fte)
Location: Ministry of Justice & Security
Title: FederATE (2015-2019)
Topic: Research Data Management
PhD student: Armel Lefebvre (1.0 fte)
Location: UU
Dr. Marco Spruit
P.I. ADS LAB
Title: CESCA (2011-2018)
Topic: Big Data Analytics in Cell Biology
PhD student: Wienand Omta (0.2-1.0 fte)
Location: UMCU/UU
Title: BeHapp (2016-2019)
Topic: Mobile Health Analytics
PhD student: Raj Jagesar (1.0 fte)
Location: GELIFES, RuG
Title: PRAISE (2015-2019)
Topic: Knowledge Discovery in Psychiatry
PhD student: Vincent Menger (1.0 fte)
Location: UU
Title: DEQUES (2018-2022)
Topic: Deep Learning for Summarisation
PhD: Ingy Sarhan (0.4 fte)
Location: AAST, Egypt
Title: TAILS (2016-2020)
Topic: NLP in Life Sciences (0.4 fte)
PhD student: Noha Seddik Tawfik
Location: AAST, Egypt
Title: SpeechAS (2017-2023)
Topic: Speech & Text Analytics in HR
PhD student: Chaïm van Toldeo (0.4 fte)
Location: P-Direct
Title: SAF21 (2015-2018) [Horizon2020]
Topic: Topic Discovery in Fisheries
PhD student: Shaheen Syed (1.0 fte)
Location: MMU, UK

Figure 14: My research projects portfolio. Shaded boxes denote former ADS Lab members.

A complete list of completed research projects for which I received competitive funding as a project partner is shown below:

1. SAF21*: Text Analytics for 21st Century Fisheries.* **EUR 200,000** (2014–2018). PhD Project for Early Stage Researcher (ESR) 7 at Manchester University. *Financer(s):* MSCA-ITN-2014-ETN: Marie Skłodowska-Curie Innovative Training Networks (ITN-ETN). *Applicant(s):* Borit, M. et al. *Remark:* EU project 642080; grant total: 2.7M EUR. [IN]
2. BeHapp: *Using the smartphone to longitudinally monitor adolescent social behavior in real life*. **EUR 100,000**. (2015 – 2016). Financer(s): Utrecht University Strategic Theme Dynamics of Youth (DoY) 2015. Applicant(s): Kas,M. [HI]
3. SNF: *Social Network Forensics.* **EUR 50,000** (2013 – 2014). Inter-ethnic relations and ethnic identity of Dutch adolescents in offline and online networks based on the Linguistic Engineering for Business Intelligence (LEBI) framework. *Financer(s):* Utrecht University Strategic Theme Dynamics of Youth (DoY) 2014. *Applicant(s):* Corten,R. et al. Seed fund for UU strategic theme Dynamics of Youth. [HIN]

4. POMP: *Polypharmacy Optimisation Method Platform.* **EUR 50,000** (2010 – 2011). Feasibility study on a knowledge platform to assist physicians, especially general practitioners, in optimising polypharmacy in elderly patients. *Financer(s):* Agentschap NL/SBIR. *Applicant(s):* Spruit,M. Small Business Innovation Research programme. [HI]
5. CURP: *Curriculum Planning app* **EUR 15,000**. (2014 – 2014). Development of an interactive serious game to support collaborative curriculum course design for education management, staff and students. *Financer(s):* Utrechts Stimuleringsfonds Onderwijs. *Applicant(s):* Spruit,M. [I]
6. CURP 1.1: *CURsus Planning app - Rev1* **EUR 5,000**. (2015/09/03). *Financer(s):* Utrechts Stimuleringsfonds Onderwijs 2015. *Applicant(s):* Spruit,M. [I]

A complete list of completed research projects which I supervised without obtaining competitive funding is shown below:

1. PRAISE: *Psychiatry Research Analytics InfraStructurE.* **EUR 200,000** (2015-2019). PhD Project (defense: 2019/10/02) in the Big Data Psychiatry Doorbraak programme. Towards predicting psychiatric conditions such as schizofrenia and autism through patient fingerprinting techniques enabled by an interorganisational information integration architecture which combines best practices knowledge with big data analytics explorations. *Financer(s):* UMC Utrecht/Psychiatry Department. [HIN]
2. STRIPA: *STRIP Assistant*. **EUR 200,000** (2010 – 2015). PhD Project (defense: 2016/01/13) on the Systematic Tool to Reduce Inappropriate Prescribing (STRIP) Assistant, an online decision support platform integrated in GPISs to support general practitioners with optimising polypharmacy in elderly patients through periodical prescription reviews. *Financer(s):* Expertise centre Pharmacotherapy in Elderly (EPHOR), UU/ICS. [HI]
3. SAM: *PhD Project SAM-SPM*. **EUR 80,000** (2009 – 2012). External PhD research (defense: 2012/11/13) on how software product management (SPM) practices can be improved in a situational manner. *Financer(s):* Software quality group, Centric IT Solutions BV. Financed as a 60% business – 40% research contract. [I]
4. DataSpace: *PhD Project Data Space*. **EUR 80,000** (2010 – 2018). External PhD research on a Data Space architecture at the crossroads of Data Warehousing, Privacy Preservation, Semantic Web, and Data Quality. *Financer(s):* Research and Documentation Centre (WODC), Ministry of Security and Justice. Financed as a 80% business – 20% research contract. [I]
5. *UBIL: PhD Project UBIL*. **EUR 30,000** (2013 – 2015). External PhD research on a unified business intelligence language (UBIL) for vendor and technology independent BI-chain modeling through a data lineage framework. *Financer(s):* CSB-System Benelux BV, Kadenza BV. Financed as a 80% business – 20% research contract. [I]

NB: For external PhD students, their research time capacity is roughly estimated on average as follows: 10K/y/day, and is not transferred to UU but directly (as salary) to the external PhD student.

P3: acquired external (outside UU) funding for research programs

I have acquired ample funding for my research projects. My top 5 in competitively funded active projects:

1. *COVIDA: Computing Visits Data for Dutch Natural Language Processing in Mental Healthcare*. **EUR 247,000** (2019-2021). Two-year postdoc, parttime scientific programmer, and deployment infrastructure. Financer(s): Alliance Fund UU-UMCU-TUE. Applicant(s): Spruit,M., Scheepers,F., & Kaymak,U. et al. Remark: Grant total: **492K EUR**. [HIN]
2. OPERAM: *OPtimising thERapy to prevent Avoidable hospital admissions in the Multimorbid elderly*. **EUR 250,000** (2015-2020). PhD Project and postdoc position in WP2: Software Tool for Optimising Medication. OPERAM's objective is to run a RCT to evaluate STRIPA 2.0. *Financer(s):* PHC 17-2014: Comparing the effectiveness of existing healthcare interventions in the elderly. *Applicant(s):* Rodondi,N. et al. *Remark:* EU project 634238; grant total: **6.6M EUR**. [HIN]
3. SMESEC: *Protecting Small and Medium-sized Enterprises digital technology through an innovative cyber-SECurity framework.* **EUR 278,400** (2017-2020). Two PhD students and extra research time to further develop my overarching, personalisable maturity model for incremental organizational improvement throughout all security domains. *Financer(s):* H2020-DS-2016-2017: Secure societies –

Protecting freedom and security of Europe and its citizens. *Applicant(s):* Diaz,R. et al. *Remark:* EU project 740787; grant total: **5.6M EUR**. [I]

4. STRIMP: *Implementatie van de STRIP Assistent ter verbetering van de STRIP medicatiebeoordeling.* **EUR 90,000** (2018-2019). Programmer for two years to integrate the STRIP Assistant within Dutch daily primary care. *Financer(s):* ZonMW/Goed Gebruik Geneesmiddelen – Stimulering Toepassing In de Praktijk (GGG – STIP Ronde 3). *Applicant(s):* Wit,N. de, Spruit,M. *et al.*. *Remark:* Complete grant total: **350K EUR**. [HI]
5. OPTICA: *Optimising PharmacoTherapy In the multimorbid elderly in Primary CAre: a cluster randomised controlled trial*. **EUR 22,110** (2017-2019). Parttime postdoc position to prepare an RCT to implement STRIPA 2.0 in daily GP practices. *Financer(s):* Research Plan NRP 74 "Smarter Health Care" Division IV, National Research Programmes (NRP), Switzerland. *Applicant(s):* Rodondi,N., Streit,S., Schwenkglenk,M., Trelle,S., Spruit,M., & Schilling,G.. *Remark:* Swiss National Science Foundation (SNF) project; grant total: **475K EUR**. [HI]

My top 5 in other funded active projects:

1. *CESCA: CEll SCreening Architecture & Analytics*. **EUR 200,000** (2011 – 2019). PhD Project on HC StratoMineR, an online platform for high content screening and cloud-based data analysis services for drug target discovery and validation, leads discovery and optimization, and the assessment of cellular toxicity. *Financer(s):* UMC Utrecht / UU-ICS. [HI]
2. *PRAISE*: *Psychiatry Research Analytics InfraStructurE.* **EUR 200,000** (2015-2019). PhD Project in the Big Data Psychiatry Doorbraak programme. Towards predicting psychiatric conditions such as schizofrenia and autism through patient fingerprinting techniques enabled by an interorganisational information integration architecture which combines best practices knowledge with big data analytics explorations. *Financer(s):* UMC Utrecht/Psychiatry Department. [HIN]
3. *FeDerATE*: *Fair Data and context ArchiTEcture*. **EUR 200,000** (2015-2019). PhD Project on data stewardship focusing on research reproducibility and interactive text and data analytics in the *omics domains. *Financer(s):* Utrecht University IT Services (UU-ITS) / Utrecht Bioinformatics Centre (UBC). [HIN]
4. *TAILS: PhD Project Text Analytics Innovations in Life Sciences*. **EUR 40,000** (2016 – 2020). External PhD research on text analytics based innovations from both machine learning and computational linguistics perspectives to better understand their specific added values throughout the broad application domain of personalised medicine. *Financer(s):* Arab Academy of Science, Technology & Maritime Transport (AAST). Financed as a 60% lecturer – 40% researcher contract. [HIN]
5. *SpeechAS*: *Real-time Speech Analytic Systems for HR dialogue support.* **EUR 120,000** (2017-2023). "PhD-IT Rijksoverheid" project on NLP innovations in the Human Resources domain. *Financer(s):* P-Direct. Real-time speech recognition and text classification for real-time dialogue enrichment and chatbot development within a Human Resources context. [IN]

P4: has been leader of a research group

I am the founder and Principle Investigator of the Applied Data Science Lab in the Dept. of Information and Computing Sciences of Utrecht University. I currently supervise 11 PhD students and 1 scientific programmer, as shown in Figure 11. Prof. Sjaak Brinkkemper is the promotor for all PhD students, as I do not have the *ius promovendi*. For 3 PhD students (Armel, Vincent, Wienand) there is a second promotor. For 9 PhD students I am the main co-promotor; for 4 PhD students (Ian, Bilge, Noha, Ingy) I am also the only co-promotor. For Alireza and Friso I act as the second co-promotor.

P5: has (co-)supervised PhD students up to and including their defense

I have completed 3 PhD projects until now, ordered on chronological graduation date:

13 Nov 2012: W. Bekkers, Ph.D.: Situational Process Improvement in Software Product Management.

Willem's dissertation on the Situational Assessment Method (SAM) integrates concepts including Situational Factors, Process Improvement, and Maturity Matrices in the domain of Software Product Management (Funded by Centric IT BV).

Current position: Strategic product manager at Centric IT BV.

13 Jan 2016: M. Meulendijk, Ph.D.: Optimizing medication reviews through decision support: prescribing a better pill to swallow.

Michiel's dissertation investigates the effectiveness and efficiency of decision support in the medication review process. This work has resulted in the development of the STRIP Assistant as a prescriptive polypharmacy platform for primary care (Funded by UMCU/UU).

Current position: postdoc at Leiden UMC.

20 March 2019: S. Syed, Ph.D.: Topic Discovery from Textual Data: Machine Learning and Natural Language Processing for Knowledge Discovery in the Fisheries Domain.

Shaheen's Horizon2020 Marie Skłodowska-Curie (MSC) – ITN - ETN funded dissertation investigates how can we improve the knowledge discovery process using textual data from multiple latent topical perspectives (Funded by Horizon2020).

Current position: Scientific programmer at Computer Science dept. of Arctic University Norway.

Furthermore, 1 PhD student has a defense date set for his approved dissertation:

2 October 2019: V. Menger: Knowledge Discovery in Clinical Psychiatry: Learning from Electronic Patient Records.

Vincent developed the Psychiatry Research Analytics InfraStructurE (PRAISE) to support predicting psychiatric conditions such as agression through patient fingerprinting techniques by combining best practices knowledge with big data analytics explorations, with particular attention to Natural Language Processing (NLP) techniques (Funded by UMCU).

P6: has been responsible for the quality of a research program

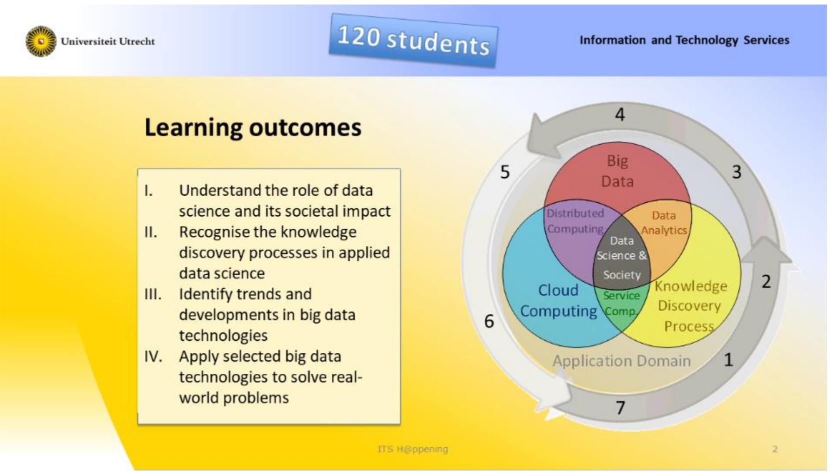
I am the primairily responsible supervisor for all ICS-related research in the research programmes above. The healthcare-specific quality is being safeguarded by my medical collaborators.

I have several quality assurance measures in place. First, I have periodic personal meetings with each PhD student. Second, I organise monthly ADS Lab meetings where PhD students and postdocs exchange and discuss current work issues, so that they know about each other's current research topics. Additionally, this informal colloquium explicitly provides an early feedback platform for yet incomplete research papers. Third, I host bimonthly ADS sessions in the Organisation & Information research colloquium where my PhD students present current work. Fourth, I strictly adhere to a paper-by-paper progress monitoring approach, which is transparant for all parties, although also somewhat frustrating at times due to unfavorable reviewers. Finally, I primarily aim to inspire and motivate my PhD students and allow them quite some freedom in the pursual of the exact research paper topics, in order to optimise their intrinsic motivation, which I believe positively influences the research quality as well.

III. General skills, including outreach and valorization

S1: transfer new research knowledge into education programs

First, I coordinate the master course Data Science & Society, an obligatory course for both Business Informatics (MBI) and Applied Data Science (ADS) profile students, which is structured around my position papers on Applied Data Science as discussed in the introductory lecture and as shown on the main course page (Spruit & Lytras, 2018; Spruit & Jagesar, 2016). Second, I also cover these materials in the introductory lecture of the Data Analytics Informatiekunde bachelor course.



Universiteit Utrecht

120 students

Information and Technology Services

Learning outcomes

I. Understand the role of data science and its societal impact
II. Recognise the knowledge discovery processes in applied data science
III. Identify trends and developments in big data technologies
IV. Apply selected big data technologies to solve real-world problems

ITS H@ppening

2

Figure 15: The conceptual model of the master course Data Science & Society is in close alignment with my research theme of self-service data science (here: service computing). This slide is from a presentation on this course during an ITS H@ppening lunch event at Information Technology Services of Utrecht University on April 1, 2019.

Furthermore, and also in part due to the continuous innovations within my research field, I yearly revise my courses to include a selection of the newly emerging state-of-the-art topics. In fact, I have concluded my lecture series in both courses with a Trending Topics in Data Science for several years now. This regularly results a year later with thesis projects where these topics are being further investigated as student graduation projects. For example, advanced topics such as Deep Reinforcement Learning and Automated Machine Learning were originally introduced in my Trending Topics lecture, and have led to several MBI graduation projects afterwards. Upon successful thesis completion, such techniques can be included into a next edition of the course's lecture series, thereby creating a valuable yet low-entry research-education symbiotic relationship.

S2: develop and teach PhD-level graduate/research/summer courses

Although developing and teaching for PhD students would be great, I have not yet seized an opportunity to pursue such an endeavour. Alternatively, I could join an existing PhD level course related to Applied Data Science and/or Natural Language Processing, for example Kaleidoscope Data Science by prof. Uzay Kaymak (TUE) with whom I actively collaborate.

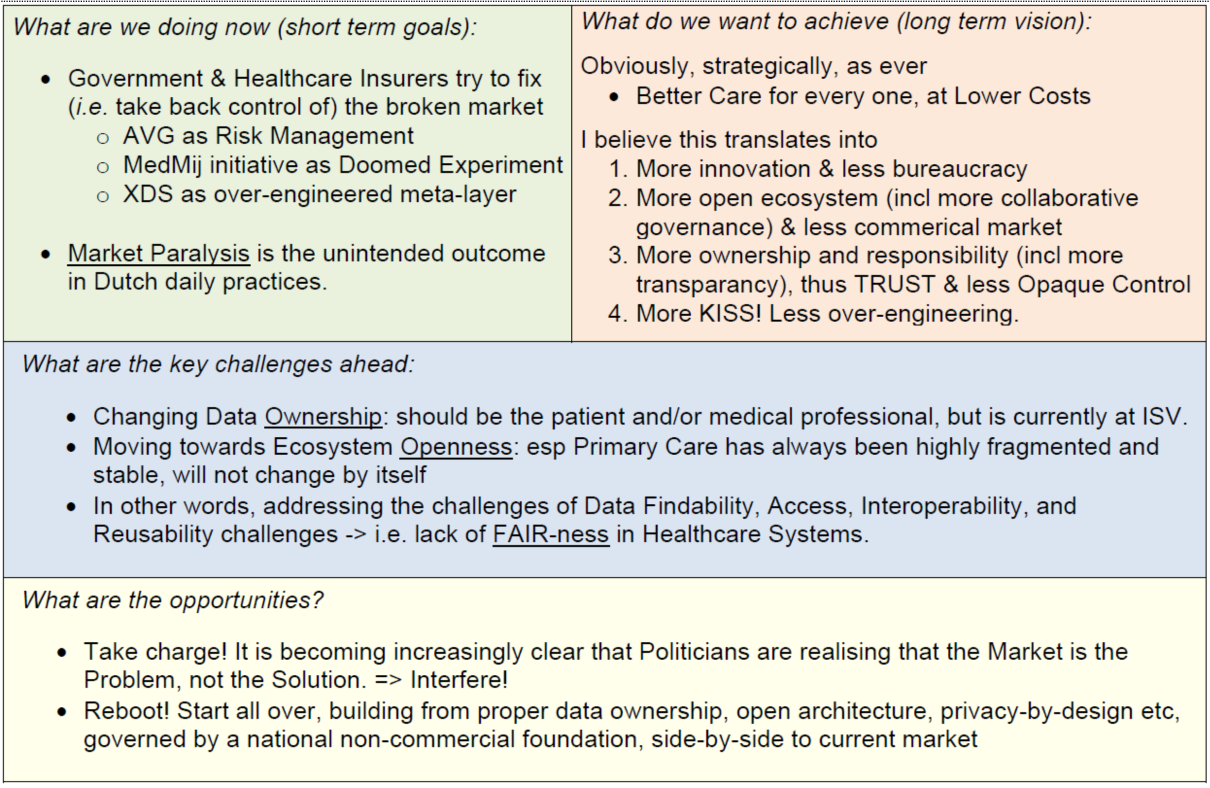
S3: participate in departmental/interfaculty committees, shape research policy

I am a co-organiser of the Applied Data Science (ADS) Special Interest Group (SIG) on Text Mining, providing a hub position to directly influence strategic decisions regarding NLP research at UU and the public and private sectors, by providing advices on ADS SIG funding proposals, workshop topics and potential industrial collaborations. I also participate in the interdisciplinary Governing the Digital Society focus area, and am a General Assembly member of the interdisciplinary Life Sciences EXPOSOME Hub.

From an educational perspective, in 2016 I was the Science faculty representative in the UU working group Applied Data Science, which kickstarted the 90EC Applied Data Science for Health MSc postgraduate programme and the ADS 30EC master's profile for both natural and life sciences students. Due to these efforts, I am now a member of UU's Applied Data Science Education Committee, and programme coordinator of the ADS profile and (former) MSc Postgraduate programme. Before 2016 I was the education manager of the Information Science BSc programme and member of the departmental Board of Examinations. As described in section S1, these strategic education positions indirectly provide opportunities to shape research directions. As an example, I recently participated in the departmental Workgroup "Technical Redesign of the Information Science bachelor programme"; the capabilities that the students will develop, indirectly help determine their research topics, *e.g.* learning Python instead of R will better facilitate NLP research as most advances are implemented using Python libraries. On a final note, in 2013 I was a jury member for the Graduate School of Natural Sciences (GSNS) master thesis award.

Finally, besides organisational and educational roles, PhD research supervision may also influence UU research policies, at least in the case of my PhD student Armel Lefebvre, whose research on Research Data Management is being funded by UU/ITS and has actively influenced ITS policies on several occasions.

S4: actively participate in (inter-)national scientific networks and committees



What are we doing now (short term goals):

- Government & Healthcare Insurers try to fix (*i.e.* take back control of) the broken market
  - AVG as Risk Management
  - MedMij initiative as Doomed Experiment
  - XDS as over-engineered meta-layer
- Market Paralysis is the unintended outcome in Dutch daily practices.

What do we want to achieve (long term vision):

Obviously, strategically, as ever

- Better Care for every one, at Lower Costs

I believe this translates into

1. More innovation & less bureaucracy
2. More open ecosystem (incl more collaborative governance) & less commerical market
3. More ownership and responsibility (incl more transparancy), thus TRUST & less Opaque Control
4. More KISS! Less over-engineering.

What are the key challenges ahead:

- Changing Data Ownership: should be the patient and/or medical professional, but is currently at ISV.
- Moving towards Ecosystem Openness: esp Primary Care has always been highly fragmented and stable, will not change by itself
- In other words, addressing the challenges of Data Findability, Access, Interoperability, and Reusability challenges -> i.e. lack of FAIR-ness in Healthcare Systems.

What are the opportunities?

- Take charge! It is becoming increasingly clear that Politicians are realising that the Market is the Problem, not the Solution. => Interfere!
- Reboot! Start all over, building from proper data ownership, open architecture, privacy-by-design etc, governed by a national non-commercial foundation, side-by-side to current market

Figure 16: My Discussion Canvas for the Round Table Session Healthcare at ICT.OPEN2019.

I represent UU in the Data Science Platform Netherlands (DSPN), a Special Interest Group within the ICT research Platform Netherlands (IPN). Furthermore, I participated as the Data Science expert member in the interdisciplinary New Science Agenda (NWA) taskforce on Prevention (chaired by prof.dr. Rick Grobbé). Additionally, I initiated and remained the UU contact person for the Big Data Alliance until 2017. Recently I was invited for the Round Table Session on Health at ICT.OPEN2019 where I provided policy input to NWO/ZonMW for the Dutch Digitisation strategy, based on a Discussion Canvas as shown in Figure 16. On a related note, I am currently participating in the Evaluation Committee of the NWO Open Competition for Digitalisation SSH.

Finally, I am a member of the National School for Information and Knowledge Systems (SIKS) and Institute for Systems and Technologies of Information, Control and Communication (INSTICC), among others. I have discontinued my memberships of the European Association for Data Science (EuADS) and Association for Information Systems (AIS) because of the “membership fee-benefits” imbalance.

S5: actively participate in (inter-)national societal networks and/or committees

Above all, I have heavily invested in my Utrecht network: in 2013 I participated in the Educational Leadership programme. In 2018 I completed the Academic Leadership for Associate Professors programme, and I will have finalised the Research Leadership Programme at Utrecht University by August 2019 as well. These three programmes have enabled me to further expand and maintain my professional UU network also at a personal level.

Due to my broad Utrecht network, I have been invited twice to act as the ICT expert member of the Hobéon accreditation committees auditing the bachelor programme Informatie en Communicatie Technologie (ICT) and the professional programme Master of Informatics at the Hogeschool Utrecht in 2010 and 2016, respectively.

Next to that, I have initiated many societal organisations for establishing strategic partnerships in my collaboration projects. A good Dutch example can be found in my STRIMP project, where we aim to improve the daily practices of pharmacists and general practioners. This requires establishing relationships with FleiR Pharmacies, LRJG General Practices and XIS information system vendors such as PharmaPartners and Tetra. In other projects I follow a similar strategic approach.

Furthermore, a good international example of engaging with societal networks can be found in the context of my SMESEC IA Horizon2020 project. Apart from the various industrial sector partners, and the Dutch Chamber of Commerce, we are now also actively collaborating with prominent European Standards Developing Organisations (SDOs) such as CEN/CENELEC, ETSI, ECSO, Digital SMEs alliance, Small Business Standards, and the European Software Institute.

Finally, I have been active in several societal organisational boards since 2010, aiming to contribute to a better and more transparant world by utilising my ICT expertise and experience, as listed below. Since 2015 I have been a board member of the mijnIBDcoach foundation in Woerden, where I supervise and advice on all ICT aspects regarding this e-Health solution for patients with Inflammatory Bowel Disease (IBD) which is increasingly being prescribed throughout hospitals within the Netherlands. Before, I was a member of the supervisory board of the County Library of South East Utrecht specializing in ICT innovation, e-participation, and knowledge management. To conclude, as chairman of the Democrats branche in the municipality of Wijdemeren, I have professionalised the organisation through community building and by establishing standardised codification practices to secure the growing body of chiefly tacit knowledge for future political generations (*i.e.* knowledge management).

S6: develop and give presentations on research (programs) for a broad audience

- *Applied Data Science masterclass* (16/04/2019). Rotterdam, ESRI Nederland [5*45 min].
- *Applied Data Science masterclass* (19/12/2019). Rotterdam, ESRI Nederland [5*45 min].
- Towards healthcare business intelligence in long-term care: An explorative case study in the Netherlands (05/03/2015). Dutch Healthcare Authority (NZa), Utrecht, Netherlands [50 min].

- PSGF: The Pricing Strategy Guideline Framework for SaaS Vendors (16/09/2014). How to Price My Saas? Bootcamp, Brussels, Belgium.

S7: developed or participated in various outreach activities to promote science

I have reached out several times in public in the following non-scientific settings: UU, ESRI, NZa, a Belgian SaaS startup Bootcamp, and at several secondary schools in the Utrecht region.

In April 2019 I introduced my master course Data Science & Society during an ITS H@ppening lunch event at the Information Technology Services (ITS) unit of Utrecht University, as shown in Figure 15. The aim was to help the audience better understand the impact of the IT services such as MS Azure that they are supporting and how they enable us to teach students about these technologies.

Second, in 2018-2019 I have co-developed and twice co-presented a one-day (5*45 minutes) masterclass Applied Data Science at ESRI Netherlands for their customers, where we explain them how to reliably conduct the entire knowledge discovery process using various published scientific works. This masterclass has received very positive evaluations.

Other notable outreach activities are invited talks at the Dutch Dutch Healthcare Authority (NZa), the How to Price My Saas? Bootcamp of the Belgian NebuCom/Sirris incubator, and my table host role at the The Hague's iPoort political world café on the theme of "Decentralisatie van overheidstaken naar gemeenten. Is de informatievoorziening al geregeld?".

From a publication perspective, on a few occasions we have written popularised versions of our research to further promote our scientific work to a broader audience, through publication in the professional magazines *Management Accounting & Control* and *Informatie*.

Finally, I have participated thrice in the regional Rector's League to promote information science to secondary school students, with talks titled *Informatiesystemen: van Phishing tot Iron Maiden*, *Informatiesystemen: De wereld achter phishing en Project X*, and *Beter zoeken dan Google*.

S8: led or participated in valorization activities other than teaching or outreach

Figure 17: Core Life Analytics BV, with my PhD student Wienand Omta, signed an agreement on December 19th 2016 for exclusive rights to commercialise HC StratoMineR, the big data analytics platform developed at UU/UMCU.

Next to the Public appearances and Outreach activities listed above, two of my analytic systems research projects have reached a valorisation-relevant status. In terms of Technology Readiness Levels (TRL), the STRIP Assistant has recently reached TRL 6, whereas HC StratoMineR (the analytic system developed throughout Wienand Omta's CESCA PhD project) has already reached TRL 7. Consequently, StratoMineR is now at the heart of the newly founded spin-off Core Life Analytics BV. The milestone event of formally acquiring the license to commercialise the technology was on December 16, 2019, as shown in Figure 17. Because of their recent successes, UU-ICS will now start receiving a significant and yearly percentage of their turnover in concordance with the UU Holding license agreement. I hope to be able to reinvest these royalties to speed up the realisation of my other analytic systems still under development.

From 2013-2015 I have explored together with the UU Holding possible commercialisation possibilities with respect to the STRIP Assistant analytic system. See also Non-academic board positions – Ancillary positions. However, these discussions did not result in a positive outcome.

Finally, one could consider my participation as opposing committee member at the PhD defenses of Stella Pachidi (VU) and Gilbert Silvius (UU) as possibly relevant to this catch-all activity section.

Utrecht, 10 July 2019

Marco Spruit