Prof.dr. Marco Spruit

Translational Data Science in Population Health

Universiteit
Leiden
The Netherlands

Translational Data Science in Population Health

Inaugural lecture by

Prof.dr. Marco Spruit

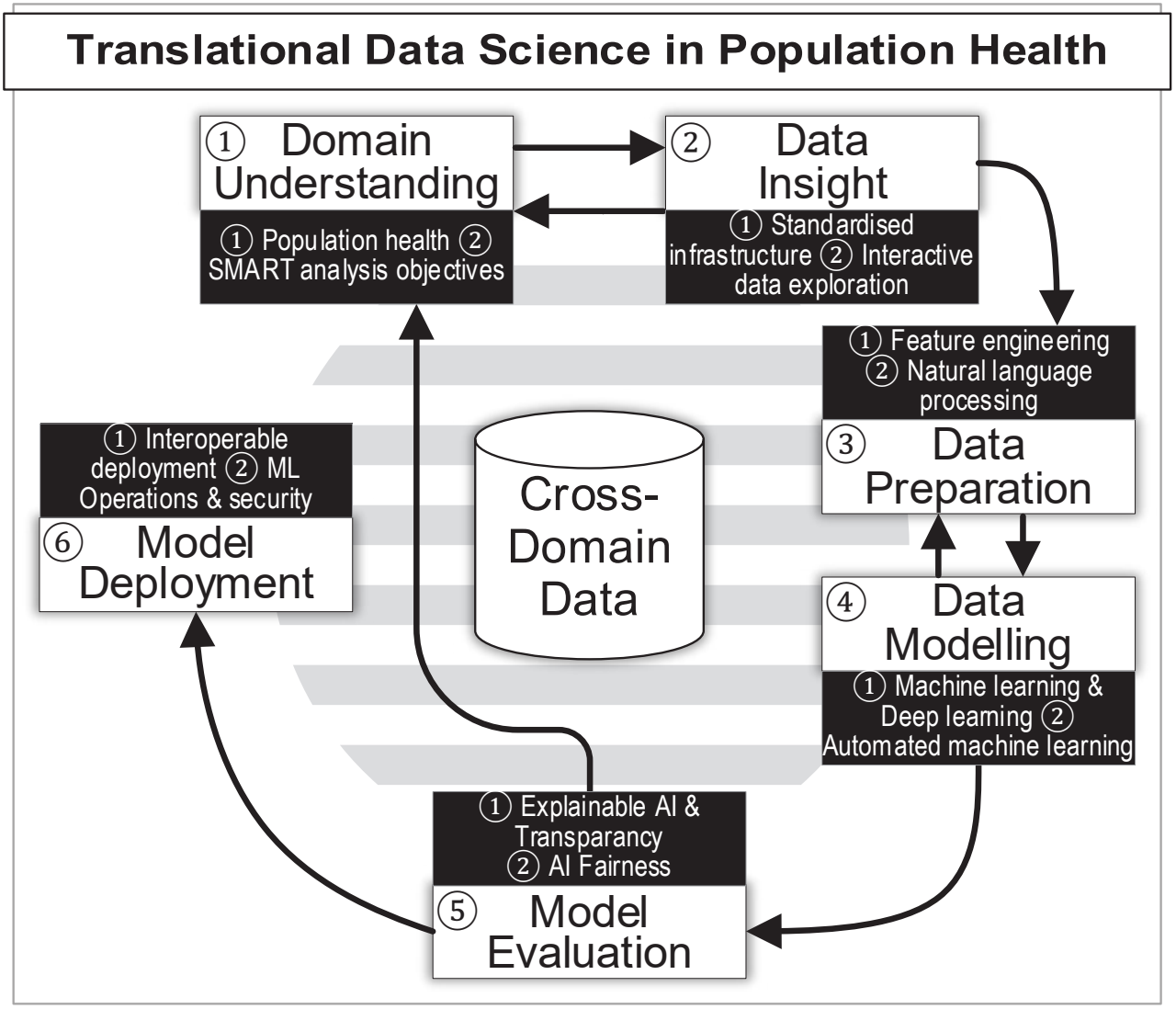
on the acceptance of the position of professor of

Advanced Data Science in Population Health

at Leiden University

on 1 April 2022

Universiteit
Leiden
The Netherlands


Translational Data Science in Population Health
1 Domain Understanding
1 Population health 2 SMART analysis objectives
2 Data Insight
1 Standardised infrastructure 2 Interactive data exploration
1 Feature engineering 2 Natural language processing
3 Data Preparation
4 Data Modelling
1 Machine learning & Deep learning 2 Automated machine learning
5 Model Evaluation
1 Explainable AI & Transparancy 2 AI Fairness
6 Model Deployment
1 Interoperable deployment 2 ML Operations & security
Cross-Domain Data

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

With this public lecture, titled Translational Data Science in Population Health, I officially accept my appointment as full professor at both the Leiden University Medical Center and the Faculty of Science of Leiden University, holding the chair Advanced Data Science in Population Health.

Translational data science: the research area

Over the next 45 minutes, I will introduce translational data science as an independent discipline at Leiden University and in the Dutch scientific landscape. I will explain the *why, how* and *what*. The overarching storyline runs from conceptual science policy to unruly implementation in daily practice.

This story begins in 1945, with Dr. Vannevar Bush. In his position as director of the US Office of Scientific Research and Development, he wrote a highly influential article with the beautiful title *"Science: The Endless Frontier"*. This created a standard policy classification of the nature of any scientific work as either basic or applied.[1]

Until 25 years ago. In 1997, Professor Donald Stokes, then Dean at Princeton University, published an alternative classification tool for scientific research: Pasteur's Quadrant.[2] Pasteur's Quadrant is a policy model in the form of a square with two rows and two columns, which classifies scientific research on the basis of two dimensions instead of just one. The vertical axis represents the already common *"degree of fundamental understanding"*. The horizontal axis shows the *"degree of practical use consideration*". Stokes developed Pasteur's Quadrant to nuance the age-old friction between basic and applied research.

Applying Pasteur's Quadrant of scientific research approaches to the broad range of adjacent data science disciplines, the following three sub-disciplines of data science emerge. First of all, in the upper left quadrant, we find purely fundamental, basic data science, in which one mainly develops new algorithms and data methods in order to perform increasingly better data science, with continuously improving analysis tools. In Bush's words: *"without thinking of practical applications"*. However, it appears that results from this type of scientific research often do not trickle down to society. In fact, the vast majority of this type of research falls *"in oblivion"*... The bottom right of this quadrant of research approaches denotes purely applied data science, where a societal demand-driven approach is used and the aim is to be able to make a tangible contribution to innovation and to tackle societal issues in the shorter term.

Finally, the top right part of Pasteur's Quadrant represents *"basic research that seeks to extend the frontiers of understanding also inspired by considerations of use"*. In other words, translational data science. On the one hand, the translational data science approach, like the purely fundamental approach, seeks a better fundamental understanding of the world around us. On the other hand, like the purely applied approach, it is societally inspired, demand-driven and solution-oriented. In short, translational data science offers the best of both worlds!

Note that translational data science is not a new sub-discipline within the spectrum of data science disciplines. The first academic conference dedicated to precisely this topic took place five years ago in Chicago. It defined translational data science as: *"[...] a discipline that applies data science principles, techniques and technologies to problems in other disciplines [...] and whose results also broaden or deepen our basic understanding of data science"*.[3]

Finally, translational research is traditionally divided into two categories.[4] Applied to the topic of this lecture, "T1"

translational data science research focuses primarily on an effective translation of novel algorithms, data models and techniques in population health. For data science, this results in a better understanding of the exact behaviour of the technique in real-world scenarios. For population health, it provides insight into new intervention possibilities for prevention, diagnosis and treatment of diseases, and for improving health. "T2" translational data science research mainly focuses on translating knowledge gained in T1 into everyday practice through rigorous testing of novel treatments and research knowledge on appropriate patients or populations.

T1, T2... Translational medicine as analogy

The best-known branch of translational science is arguably *translational medicine*. It is concisely described by its well-known motto *"from bench to bedside"*. In this branch of medicine, for example, new knowledge from the research laboratory about the working of the COVID-19 virus ("T1") is translated into a Janssen vaccine in the daily practices of general practitioners and public health services, in order to promote health protection among the appropriate population groups ("T2"). Thanks to the T2 research, we now know, for example, that pregnant women should not receive this Janssen vaccine.

Translational data science: the process

So much for scientific theory. You now know *what* translational data science is, and *where* it fits into the scientific field. You also understand *why* integrating basic and applied research provides added value to both data science and society. Now I will briefly explain *how* the standard process for translational data science is carried out. I will do this using the time-honoured and long-tested scientific method, in four steps. Then I will explain the translational data science process as an extension of this scientific method.

In Step 1 of the scientific method, you formulate a new research question. Note that it requires domain understanding by a subject matter expert to arrive at a relevant research question at all. In Step 2, you design a corresponding research methodology—such as a computational experiment or a clinical study—with which you can answer this question in a reliable and reproducible manner. In Step 3 you collect all the relevant data you need for this. In Step 4, you finally analyse the data from Step 3, using the research plan that you designed in Step 2, to obtain a reliable answer to the research question from Step 1. Often, the newly acquired knowledge raises new follow-up questions, causing you to go back to Step 1 again in order to formulate a new relevant research question based on the domain understanding you have just expanded. And so on. This is the scientific method in a nutshell. It works!

If we now zoom in from scientific research in general to data science research in particular, one more dimension is added, namely that of engineering. With engineering we refer to the technologies and architectures that help implement the data analysis in software scripts or analytical applications.[5] In addition, the toolbox of data analysis techniques has been expanded with techniques from both machine learning and statistics. With these two extensions to the scientific method, we can now describe the cyclical process of translational data science in six standardised phases.

Data scientists in both the commercial sector and science have been using the so-called "Cross-Industry Standard Process for Data Mining" (CRISP-DM) for several decades.[6] This process offers a multi-layered and cyclic step-by-step plan for carrying out each analysis task in each phase of the process in a standardised and already proven manner. This standard process appears to match extremely well with the type of scientific research that we pursue with translational data science. For example, each study is initiated from "practical use considerations" that are translated from problem definitions into specific, measurable, achievable, relevant and

time-bound (SMART) data analysis goals. The "fundamental understanding" of the data-analysis techniques themselves is enhanced, among other things, through qualitative model evaluation with domain experts. For example, a predictive model that performs very well can still be rejected for use in daily practice, due to a lack of transparency, ethical grounding or legal compliance.

With this brief introduction to the step-by-step process of performing translational data science, I have now gently introduced all the pieces of the puzzle. You now know *what* it is, *why* it is important, and *how* to do it. In the remainder of this public lecture, I will provide a step-by-step description of the six phases of the "cross-industry standard process for data analysis", touching on at least two research topics for each phase, along with short practical examples.

Phase 1: Domain Understanding

The first of the six stages in the "cross-industry standard process for data analysis" is to better and explicitly *understand the domain* under investigation. This is because we want to proceed on the basis of *"practical use considerations"*. Therefore, it is crucial to first carefully uncover the intended project objectives of the stakeholders, and to define success criteria for these. On this basis, the feasibility of the objectives can be estimated, given the risks, constraints and available time, money and manpower. Then, we translate the objectives into a technical data analysis problem definition, with measurable performance indicators. In short, we need to devise a SMART plan!

1.1 Population health

My research focuses primarily on the application domain of population health and its data-driven management. This interdisciplinary approach is one of the eight strategic focus areas of the LUMC. On the interdisciplinary Health Campus The Hague, our central mission is to contribute, from various scientific and clinical perspectives, and in cooperation with regional partners, to a longer and healthier life for everyone. Three principles are central to our approach: reducing health inequalities, pursuing a sustainable approach, and utilising a broad, positive health perspective.[7]

Population health management employs a data-driven and innovation-driven focus on prevention and lifestyle. The starting point is the re-use of routinely recorded data on health, medical care and the social domain, insofar as this is technically, legally and ethically acceptable. Moreover, establishing the proper objectives often involves a combination of both qualitative and quantitative data sources. By means of advanced data analysis techniques from the research fields of data science and artificial intelligence (AI), we can use these data sources to identify risk groups and provide the right care in the right place by linking routine medical data to social and public domain data such as living conditions and financial debt problems. At the interdisciplinary Health Campus The Hague, we work together in a so-called *quadruple helix of innovation*[8] from our "Population Health Living Lab" in the Haaglanden region.

1.2 SMART analysis objectives

Thanks to the sustainable collaboration with enterprises, governments and citizens in our quadruple helix, we can usually arrive at the right information needs with confidence. However, if for some reason it proves impossible to formulate one's own information needs accurately, it is also possible to distil objectives semi-automatically from existing organisational data such as internal documents, meeting reports and e-mail conversations.[9] The next step is to translate these information needs based on the available data sources into executable data analysis objectives with explicitly specified minimal performance values.[10] This provides a SMART foundation for the data analysis.

Phase 2: Data Insight
The second of the six phases in the "cross-industry standard process for data analysis" concerns *data insight*. In this phase, the necessary data is first collected and described, such as the number of available data points and the data types present. Of course, a description of the data in itself does not provide sufficient data insight; what is required is an interactive exploration of the available data, in order to personally observe the value variety within the data. It also leads to an assessment of the data quality, as this helps determine the necessary data preparation activities and the quality of the analysis results.

2.1 Standardised infrastructure
Gathering the necessary data sources is more than ever a specialism in itself. Traditionally, the data collection and management process usually consists of extracting and transforming operational data from organisational databases, possibly enriched with publicly available data sources from, for example, Statistics Netherlands (CBS). Depending on the degree of structure present in the collected data and the desired data insights, the infrastructure can be designed as a tabular *database* for numeric values, a *data lake* for texts and images, or a volatile *data stream* for continuously updated messages.[11]

However, in a population health context in which several partners work together in a quadruple helix, an *inter-*organisational infrastructure is required to be able to share data beyond the boundaries of individual organisations. Routinely recorded data on health, medical care and the social domain are subsequently linked in a technically, ethically and legally acceptable way. Such a data infrastructure can provide enriched insights into the health, costs and experiences of the inhabitants of a region and also offers starting points in population health management for both healthcare system players, organisations, professionals and citizens.

ELAN, the Extramural LUMC Academic Network, is such an interdisciplinary cross-domain regional data platform, including an appropriate policy structure, that supports health policy and research. ELAN has been realised as a remote access environment at CBS. For the Haaglanden region, ELAN already links structured medical data, mental health information, social domain context and public health of hundreds of thousands of inhabitants. But, thanks to the unique status of the CBS as laid down in the Dutch 2003 CBS Act, the ELAN data can be further enriched with socio-economic data such as income, education level, employment status, household composition and neighbourhood and district data. On an *individual* level. Nevertheless, the privacy provisions comply with the General Data Protection Regulation (AVG).

Federated learning
Exactly one year ago, our previous cabinet decided to honour the proposal for a national health data infrastructure (Health-RI). With an investment of €69 million from the National Growth Fund, we have now started to make health data more accessible for health research and innovation. An important question here is how to link the various existing architectures, such as ELAN. This is because in the future many data infrastructures will continue to exist side by side, because different analysis tasks simply require different data architectures and solutions. Or to quote the great Maslov: *"It is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail"*.

How does this national plan relate to ELAN? To raise one relevant issue, ELAN's remote access environment does not support text and image data. This is rather unfortunate, since it is known that the many text reports and clinical notes that doctors and nurses write about their patients in their Electronic Health Records (EHRs) provide valuable insights into the health and experiences of their clients. Therefore, in order to share text data securely, new technologies have been

developed in recent years: federated learning and secure multi-party computation.

Privacy-by-design
The emerging fields of federated learning and secure multi-party computation offer an interesting solution direction for inter-organisational linking of unstructured and multimodal data. They allow patterns to be learned from sensitive data from multiple sources in a secure manner, without having to share this data. In federated learning, the analysis task is brought to the data instead of the other way around, so the data does not have to be moved or shared, only the software code needed for the analysis task: Computing Visits Data (CoViDa).[12] Secure multi-party computation, on the other hand, uses cryptography, so that analyses are performed on an encrypted version of the data. Our fundamental right to privacy, as also laid down in the AVG, remains inherently safeguarded. This enhances trust in such technology, and makes it easier to make data available because of this privacy-by-design characteristic.

Another quickly emerging development is the privacy-by-design use of *synthetic* data in data analysis, which guarantees safe handling of sensitive data. A synthetic dataset is a completely newly generated dataset based on the actual source data, that retains the original characteristics, relationships and statistical patterns of the original data. In other words, it is a clone, a digital twin. On the one hand, this results in minimal risk and maximum privacy, and on the other hand, the quality and delivery time of the synthetic data is better and faster than that of the original. So, it's a win-win situation. At the LUMC I am currently leading an initiative to construct a synthetic clone of the unique ELAN dataset, to support more realistic and inspiring data analysis education within the Medicine and Population Health Management programmes, among others.

2.2 Interactive data exploration
Once an integrated data infrastructure is in place, with the proper privacy-enhancing technologies (PETs) for accessing data, as outlined, we are now ready to start looking at the data of interest. The initial exploratory findings on the data analysis objectives help to better understand the exact nature of the data. For example, with regard to data quality: how complete is the data, are there many missing values, is the distribution of possible data values unbalanced? And so on.

Two aspects are particularly important in an interactive data exploration. First, the use of visualisation techniques to clarify data characteristics. For example, histograms to visualise the frequency distribution, or a pair plot to relate pairwise relationships. Research has also shown that interactive graphs improve data understanding.[13] For example, think of the possibility of specifying filters on the data shown, or selecting a sub-area with your mouse to zoom in further.

The other important aspect of an interactive data exploration is to carry it out in a multidisciplinary team with at least one data scientist and one domain expert. In practice new visualisations by the data scientist do indeed result in new hypotheses by the domain expert, because insightful visualisations stimulate creativity. The methodological consequences of this are significant, because such an approach focuses on finding new knowledge and unexpected hypotheses from data, rather than simply performing the carefully defined data analysis tasks from Phase 1.[14] This finding is also reflected in the well-known African proverb: "Alone you go faster, but together we go further".

Phase 3: Data Preparation
We have now arrived at the third of the six phases in the "cross-industry standard process for data analysis": *data preparation*. It is in this phase that the hard and dirty work is done. It is estimated that 80-90% of all data science project activities take place in this data preparation phase. First of all, we select only the data to be analysed, because unfortunately, the marketing mantra *"the more data, the more insight"* does generally not apply in practice. *"The more data, the more*

noise" would be a more accurate mantra. After that, it is often necessary to improve the data quality to a level that is minimally assumed by the chosen analysis techniques. How do we deal with missing values? Do we keep them out of the data selection, do we insert suitable default values or can we estimate missing data with sufficient reliability using a model? The main objective in this phase is to make all the necessary preparations to be able to carry out the subsequent data analysis in the best possible way.

3.1 Feature engineering

The data preparation phase is *the* time to create new, distinguishing features based on already available data. This can vary from deriving the province name from a place name, to automatically extracting the current smoking status of a patient from the journal texts that the general practitioner has entered as freely formulated text in the electronic health record during consultations. Including typos, incomplete sentences and various abbreviations.

Furthermore, all kinds of data transformations are necessary, for example, to convert the data type of a data attribute, such that the desired data analysis technique can be applied. The smoking status "Non-smoker" can be represented with a 0, the smoking status "Smoker" with a 1, "Ex-smoker" with a 2. Similarly, data attributes can be merged, such as "First name" and "Last name" to "Full name". And so on. In short, by now you understand why this often-time-consuming Phase 3 of *data preparation* is also jokingly referred to as *data juggling*!

Data standardisation

Whatever the case, ultimately, we end up with the data to be analysed. Cleansed, enriched, corrected. But this phase is still not complete. Sustainable data science requires that the data is also findable, accessible, interoperable and reusable (FAIR). This is obviously a noble ambition, but also requires extra attention to data standardisation aspects beforehand. How do we ensure, for example, that the data is accessible for reuse without violating AVG legislation? A partial answer to this can be federated learning, but only when the involved parties speak the same language.

Worth mentioning here is the multilingual medical terminology system for and by healthcare professionals (SNOMED). This ontology contains a huge collection of about 370,000 medical terms and their synonyms. Healthcare providers can use these terms to unambiguously record all kinds of healthcare information, including complaints, symptoms and diagnoses. Data that is consolidated using this ontology is, therefore, very suitable for exchange and reuse.[15] In addition, for the digital exchange of healthcare data within and between healthcare institutions, a 'fiery' healthcare exchange language has also been developed (FHIR). There is even a shiny universal exchange language for data model definitions (ONNX), so that prediction models can be reused in every system. Finally, for the large-scale analysis of medical outcomes, a real "Odyssey" has recently been initiated, in analogy with Homer's epic poem.[16]

The development of a cross-domain data infrastructure, as I explained in Phase 2, therefore requires a high degree of data standardisation in order to facilitate population health and research in a sustainable manner.

3.2 Natural language processing

I just mentioned as one of the *data preparation* steps the possibility to extract the current smoking status of a patient automatically from a free text field within the patient's electronic health record. I would like to elaborate on this, as it is a relevant example of natural language processing in data science. By the way, natural language processing is also a prominent research area in Artificial Intelligence (AI). In AI, the ultimate goal is to make machines exhibit human behaviour. However, from my data science perspective, natural language processing techniques are primarily interesting for understanding the content of unstructured texts, with the aim of further enriching the already available structured data, so that we can achieve the most complete data analysis insights.

When I myself studied computational linguistics at the University of Amsterdam in the early 1990s, the curriculum consisted mainly of studying and programming various linguistic theories in the form of so-called rewrite rules. For example, the very simple English sentence *"My daughter listens"* is rewritten as a sequence of a noun phrase *"My daughter"* plus a verb phrase *"listens"*. The noun phrase *"my daughter"*, in turn, is made up of a possessive pronoun and a noun. This sentence structure of rewrite rules can also be visualised as a hierarchical tree structure with branches, where the leaves are the actual words of the sentence. This all sounds very intuitive, logical and is based on language theory. Nevertheless, during each course it became apparent that these ingenious grammars of rewrite rules could not adequately represent the unruly language of everyday life. That was so frustrating... Our language simply appears to be too dynamic, too unpredictable for computers to understand in *our* way.

However, one course in the curriculum took a completely different approach, namely that of stochastic language models. Instead of logical rewrite rules, this approach uses probabilities to calculate the grammaticality of sentences. It is mathematics with language! No rewrite rules representing semantic structures, but high-dimensional vector spaces, in line with the so-called *distributional hypothesis*.[17] This distributional hypothesis states that the meaning of a word derives from the way the word is embedded or "distributed" in our everyday language usage. When a word is projected as a vector in a high-dimensional space, the nearby vectors in this space represent the specific context, and thus also the word's meaning. This alternative, almost magical approach to natural language processing caught my attention, and in 1995 I graduated with a thesis on a self-learning artificial neural network for selective dissemination of information in text data streams.[18] I am very honoured that my supervisor at the time, Professor Scholtes, is also present here today, 27 years later!

Over the past decade, this stochastic approach to natural language processing has become predominant in both scientific theory and practical implementations. The progress made over the past 30 years has frankly exceeded my wildest expectations. Nevertheless, we are not there yet. Computers still do not understand our natural language. However, we have already managed to get computers to have simple conversations with us along the lines of *"Okay Google... where is the Academiegebouw?"*.

In recent years, however, awareness has grown that the stochastic approach to natural language processing should be integrated with the time-honoured logical approach. In order to integrate distributed word embeddings with symbolic rewrite rules. This is not a new idea, by the way. Ever since the 1980s, researchers have been trying to understand our own human language capacity by connecting the sub-symbolic, stochastic and the symbolic, logical approaches to natural language processing.[19] In the coming years, I would like to contribute to the integration of these two seemingly complementary language processing approaches, from my use-inspired objective of enriched data understanding, to help realise more meaningful data analysis results.

Specifically, my interests include natural language processing tasks such as open information extraction, topic modelling and language marker detection. For example, an important healthcare application of information extraction on clinical notes in electronic health records is the de-identification of all personally identifiable information, such as name, address, telephone number, date of birth, and so on.[20] This is not only ethically highly desirable, but also legally required under the AVG. Topic modelling is a popular language processing technique to automatically extract latent (hidden) topics from clinical notes, among others. For example, a patient report may be 30% about a specific chronic condition and 70% about the patient's anxiety.[21]

Finally, I am also interested in detecting a person's mental state based on the way in which they tell their personal story, by analysing their choice of words, sentence structure and prosody, among others. A person's use of language reflects their identity. My assumption is that there are indeed language markers, that are comparable to biomarkers which indicate the presence of someone's biological properties. Through language markers in someone's language utterances, we could gain insight into that person's mental state. For example, if someone uses the first-person singular form "I" excessively in their communication, this, in combination with excessive use of words concerning home and motion, could be a language marker for autism. It is these kinds of fascinating research questions that currently have my primary interest in natural language processing techniques within the health and wellbeing domain.[22]

Phase 4: Data Modelling

After these three phases with several sub-topics each, we have finally reached the fourth of the six phases in the "cross-industry standard process for data analysis": *data modelling*. It is in this phase that the true data scientist shines! The data analysis objectives were already drawn up in Phase 1. Now the time has come to first select and run the most suitable algorithms for each objective. This is far from trivial, as there are many hundreds if not thousands of unique algorithms.

Selecting the right modelling technique and algorithm starts with determining the type of task of the desired data analysis. For example, consider predicting someone's weight in kilograms (Y) based solely on that person's height in centimetres (X). Here, the goal is to predict a numerical value. Another common task is classification, where we want to predict the correct category. For example, instead of predicting someone's weight in kilograms as the outcome, we would classify someone's weight into a predetermined category, such as: underweight, healthy weight or overweight.

Apart from regression and classification, data science includes many other analysis tasks such as clustering, association rule mining, anomaly detection and time series analysis. In addition, data analysis tasks can be descriptive, predictive or prescriptive in nature. For sure, data science revolves around a very rich and comprehensive toolbox!

4.1 Machine learning

In data science, we mainly study and use *data modelling* techniques from the research field of machine learning. In short, machine learning is a technique for allowing software and algorithms to improve themselves *autonomously* by analysing and recognising patterns in data. This is often done with the aim of making accurate predictions, such as predicting someone's weight in kilograms based on their height, age and waist circumference.

Machine learning resembles, but is not the same as, statistics. Whereas machine learning is mainly concerned with uncovering generalisable and accurate predictive patterns, statistics above all aims to infer population-wide conclusions from a sample.[23] In addition, these two adjacent disciplines have a significantly different culture and language. Programmers versus methodologists. False positives versus type-I errors. Python versus R. It is precisely for this reason that in translational data science we want to "recognise and reward" both cultures in order to get the most complete answers to our data analysis goals.[24]

As for the combined toolbox for data modelling, it now contains many hundreds of machine learning techniques and statistical tests. For example, there are algorithms for techniques that structure data based on decision trees (C4.5), association rules (Apriori), network connections (PageRank), a forest of decision trees (Random Forest) or an ensemble of weak predictors (AdaBoost). And so on and on.

In addition, most algorithms have specific configuration settings to function optimally. However, this aspect receives far too little attention. Hence my plea for meta-algorithmic modelling, which aims to document the best practices for the optimal use of algorithms in data science in a standardised manner.[25] Given a dataset with only 250 data points and 10% missing data, which classification algorithm can I best select, and how do I then determine the optimal corresponding configuration?

Deep learning

This brings me to the most popular area within machine learning: *deep* learning. In short, deep learning is an improved version of the technique used in my Master's thesis 27 years ago. The analogy of these neural networks which simulate the behaviour of the human brain still fires the imagination of many. Like our human brain, deep neural networks can learn complex patterns from large amounts of data.

There are countless examples of artificial intelligence and data science: self-driving cars, facial recognition in photographs, automatic text translations, smart game consoles, virtual conversation partners... Deep learning models, like our human brain, are robust in nature and capable of great accomplishments, but we still do not fully understand how both our brain and deep learning work. In short, the price of the top performance of deep learning techniques is a lack of transparency.

In addition, the very best performing deep learning models also have an improbably large number of parameters. For example, the number of parameters in recent deep language models is now approaching the number of synapses in our human brain: 100 *trillion*. It is also estimated that it has cost the major technology companies around €100 million to create this enormous language model! Unfortunately, I was unsuccessful in negotiating such a budget at Leiden University.

What's next? Fortunately, for many data analysis tasks it is not necessary to develop a completely new deep learning model. For many tasks, pre-trained models are already available, which we can optimise using *transfer learning*. With transfer learning, a pre-trained data model is customised so that the already modelled knowledge is preserved in the model, but at the same time new data analysis tasks can be represented efficiently and effectively.[26] This approach also offers the possibility of adding symbolic, grammatical information to such deep learning probabilistic models.

4.2 Automated machine learning

Perhaps your head is spinning by now? After all, my field of translational data science combines natural language processing, federated learning, machine learning, deep learning, transfer learning, and so on. Fortunately, there is a technological solution for results-driven translational data scientists: *automated* machine learning! As the name suggests, automated machine learning takes care of as much as possible of the data analysis process work for us. First of all, automated machine learning can greatly shorten the often very time-consuming *data preparation* in Phase 3. For example, the data quality can be improved fully automatically by calculating missing values, by converting data types in advance and by automatically revealing new, distinguishing characteristics of the data. Manual data *juggling* becomes automatic data *magic*.

However, the most ground-breaking aspect of automated machine learning lies in the automatic determination of the best machine learning algorithm, including its optimal settings, during *data modelling* in Phase 4. Given a dataset and the desired outcome measure, such as the degree of explained variance, the optimal algorithm and configuration are determined fully automatically. Additionally, it is important to urge the currently prevailing culture to move with these developments. For example, it appears that Dutch hospital physicians are very interested in automated machine learning for their clinical research, at least.... as long as it uses technique

X, because (start quote:) *"otherwise the anonymous reviewers won't understand it and will reject the manuscript. Not because it's bad, just because they don't understand it"* (end quote).[27]

It is in part for this reason that I have set myself the goal to introduce automated machine learning more broadly to medical centres in order to democratise data science and accelerate innovation in the healthcare sector. To put our money where our mouth is, we are currently developing a new Master's course entitled *"translational data science"* for students with a medical background that will be designed around the possibilities of automated machine learning. After all, there is no better environment than Leiden University to realise this plan, as LIACS colleagues such as Professor Hoos, are world-renowned experts in the field of automated machine learning.

Phase 5: Model Evaluation

We have now reached the last two phases in the "cross-industry standard process for data analysis": *model evaluation* and *model implementation*. Given the time, I will discuss these last two phases more briefly. However, this does not mean that they are less important!

Take Phase 5 in the "cross-industry standard process for data analysis": *model evaluation*. This is the moment of truth to which all previous activities have led. Have we found a meaningful answer to our question? Have we gained new insight? It is important to realise that a data scientist alone cannot provide an adequate answer. Meaningful interpretation of the results requires domain expertise. In this *model evaluation* step, for example, we assess the extent to which the data model meets the objectives from Phase 1, and the predefined success criteria of the stakeholders. And we try to understand the situations in which the best model falls short.

5.1 Explainable AI

Because of the importance of being able to explain the model results, which are after all a prerequisite for acceptance by the end users, it is no surprise that making model results more explainable has become a very active field of research in recent years. Think, for example, of predicting the risk of aggression incidents in psychiatric patients, based on the many daily clinical notes taken by both the doctor in charge and the nurses about the current state of wellbeing of their patients. In earlier research we developed a prediction model using natural language processing that, in principle, can predict better than the doctors themselves whether a patient has an increased risk of a future aggression incident.[28] Nevertheless, this model is not yet used in clinical practice, mainly because the rationale of the deep learning-based prediction model remains opaque. This results, rightly so, in a lack of trust on the part of end users such as doctors. After all, they remain ultimately responsible for the decision taken and must also be able to communicate it to the patient and colleagues.

Interactive visualisations can offer a solution here, as in the previously discussed interactive data exploration in Phase 2. For example, through dimension-reduction techniques, data points can be translated from a high-dimensional space into a two-dimensional visualisation, while maintaining the relationships between them. Or by clarifying the influence of certain data points by removing them from the data, and visualising the difference before and after. Or by visualising the nearby environment of a few random data points, in order to get a representative impression. There are even promising techniques that feed the complete prediction model into a new to-be-generated distinctive prediction model, so that the influence of specific style characteristics can then be compared interactively.[29,30]

5.2 AI Fairness

In addition to the need for explainability of a forecasting model, there is also the need for the fairness of its outcomes. In recent years, it has become clear that prediction models used in daily practice often carry considerable bias. Incidentally, this is not necessarily a bad thing, or sometimes even unavoidable,

unless its users are unaware of it.[31] Consider, for example, the recent Dutch childcare benefits scandal, in which the Dutch Tax Administration subsequently had to admit that the characteristics of nationality and second nationality, among others, were fed into the Tax Administration algorithms as risk-increasing factors, without the interested parties being able to know this.

It goes without saying that in population health, too, the fairness of the prediction models used is an important aspect. For example, we recently investigated the validity of a prediction model for the administration of sedatives to psychiatric patients.[32] Here it appeared that gender had an undesired influence: women were administered more sedatives during the first three days, solely on the basis of their gender. However, with recently developed bias compensation techniques, such as the *Prejudice Remover* method, we can neutralise such undesired biases and make the prediction model work more just and inclusively.

To answer such relevant research questions about the explainability and fairness of artificial intelligence and data science in healthcare and wellbeing, at Leiden University we join our multidisciplinary forces in an ELSA laboratory for Healthy Society and Artificial Intelligence, where researchers study the influence and impact of ethical, legal and social aspects in conjunction. Also, the "Guideline and quality criteria for AI-based prediction models in healthcare" was recently published to assist healthcare professionals in the correct use of predictive AI-driven models. Finally, the European Parliament has also recently proposed the Artificial Intelligence Act in a timely attempt to safeguard the internal market by creating conditions for the development and use of reliable artificial intelligence in the European Union.

Phase 6: Model Implementation

Phase 6, the final phase in the "cross-industry standard process for data analysis", is *model implementation*. It is this sub-area in which the outcomes of data science research, or all the steps I have discussed so far, are appropriately introduced into population health practices. Of course, this requires the necessary documentation that supports the many decisions made in the previous steps, so that the model outcomes can be adopted with sufficient trust. But from a data science perspective, we are above all interested in how the prediction model can be used in practice. Because... large-scale research has recently shown that worldwide only 13% of all developed data models are actually implemented in practice.[33] Just think of the accurate risk prediction model for aggression incidents among psychiatric patients that I just discussed. Then, after publication of the model, the question naturally arises of how the model will be managed.

6.1 Model publication

There are many methods for implementing a data model for use in daily practice. An elegant and safe way is, for example, to publish all the characteristics of the data model in a standard exchange format. In other words, the data itself, with which the model was created, is not included in the standard exchange format. Another strategy goes one step further: in addition to the data model, a custom-built end-user application is also delivered, so that the data model can be applied self-contained in every practice.

The highest achievable goal is arguably to publish a usable data model directly within an inter-organisational information system such as the electronic health records system. The doctor can then use the data model directly within his standard working environment. However, in all honesty, this is very rarely achievable, at least in the Netherlands. My own implementation experiences with the STRIPA system for better prescribing of medications also seem to indicate a system error in the current Dutch healthcare system.[34]

In practice, we mainly see a completely different publication approach, namely to simply publish the complete source code

that produces the final data model, so that the data model can be reconstructed on the target computers. This is a very accessible and popular option for data scientists, but for end users in healthcare practice, it is often too impractical.

6.2 ML Operations & security

In any case, apart from the publication strategy for the data model, the question remains how this model can stand the test of time. Because... standing still is going backwards. The world changes, so the data describing the world changes with it. In short, even a data model needs periodic maintenance to remain relevant. Often, the use of a prediction model in daily practice yields new insights that were not anticipated during the development process, resulting in new data analysis objectives in a subsequent iteration of the "cross-industry standard process for data analysis".

Finally, because of the use of healthcare models, which are often constructed on the basis of sensitive data, there is an explicit role for information security as an integral part of process management, in particular cybersecurity. On the one hand, it is important to control the inherent risks of the constantly intertwined process cycles from internal model development to external practice implementation.[35] On the other hand, proactive periodic checks should be carried out to minimise, for example, the risk of re-identification of personal data due to an incomplete anonymisation process. After all, in 1997, it turned out that the medical data of Massachusetts governor Weld in the United States could be extracted from an anonymised insurance dataset using re-identification techniques.

Summary

To summarise, I first introduced my research field of translational data science and explained *why* integrating basic and applied research provides added value for both data science and society, and population health in particular. I then explained *how* the standard process for translational data science works, as a technical extension of the scientific method. Finally, I have described the six phases of this "cross-industry standard process for data analysis" by highlighting some research areas for each of the six phases so that you can also understand what translational activities entail.

Acknowledgements

Thanks to all who contributed to the realisation of my appointment. First of all, I would like to thank the Executive Board of Leiden University, the Executive Board of LUMC, and the department heads of PHEG and LIACS for the confidence they have placed in me. Professors Numans and Plaat, dear Mattijs and Aske, I am a happy man.

Many people have played a role in my scientific education over the years. Obviously, I cannot name them all here. I would also like to thank you, everyone who is present here, very much for coming, including those who are now watching via the livestream. It is wonderful that you are all here!

I would now like to address my personal Top 3 Influencers. First of all, my college friend Edwin Brinkhuis. Dear Edwin, thanks to your cunning trickery a PhD position came onto my radar. You saw that I was ready for a new, more profound challenge. Even though I didn't know it myself yet. I am forever grateful to you for that.

Secondly, Professor Barbiers. Dear Sjef, under your calm but focused guidance at the Meertens Institute I have been able to further develop myself as a data scientist in four years with tremendous freedom of action. You were also the first to mention that my delayed start as a scientist might well speed up the remainder of my academic career.

Thirdly, Professor Brinkkemper. Dear Sjaak, during my twelve years in your research group at Utrecht University I have been able to experience all aspects of academic work. You have always guided me to get the best out of myself and

were notably the first to encourage me to apply for a full professorship. By doing so, you gave me your trust, for which I am very grateful to this day.

Finally, the front row... You are of a completely different order. Dear dad, thank you for subtly and successfully convincing me to finish secondary school at the age of 16. Dear Karin, my sister, thank you for always being there for me. Dearest Jet, my wife, your life motto says it all: "The best is JET to come". For 18 years now, you have been colouring my life, from buzz to chaos, we are on an adventure! Together with our dearest Fien. You are my smiley, my sweetheart, you make me *kaulo* happy.

Ik heb gezegd.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

20 Menger, V., Scheepers, F., Wijk, L. van, & Spruit, M.

(2018). DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telematics and Informatics, 35(*4), Patient Centric Healthcare, 727–736.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Prof.dr. Marco Spruit

2020 –	Professor Advanced Data Science in Population Health, Leiden University
2019 – 2020	Associate professor Applied Data Science, Utrecht University
2007 – 2018	Assistant professor Information Science, Utrecht University
2003 – 2007	Ph.D. researcher Computational Linguistics, University of Amsterdam
1997 – 2006	Independent product software vendor, Wizzer & Insertable Objects
1995 – 2001	Editor Personal Computer Magazine, VNU Business Publications B.V.
1995 – 1997	Big Data system developer, Dutch Military Intelligence and Security Service
1993 – 1995	Application programmer, ZyLAB Europe B.V.
1989 – 1995	Doctorandus, Computational Linguistics, University of Amsterdam
1990 – 1991	Propaedeuse Musicology, University of Amsterdam
1988 – 1989	Propaedeuse Dutch Language and Literature, University of Amsterdam

Marco Spruit is Professor Advanced Data Science in Population Health at the department of Public Health & Primary Care (PHEG) of the Faculty of Medicine (LUMC) and the Leiden Institute of Advanced Computer Science (LIACS) at the Faculty of Science (FWN) of Leiden University in the Netherlands. He is interested both in translating new algorithms to novel health applications as in implementing new insights from these novel applications into daily practices.

Marco's strategic research objective is to establish an authoritative national infrastructure for Dutch Natural Language Processing and Machine Learning to *democratise* Data Science. He focuses in particular on the Population Health and Wellbeing domain in his Translational Data Science Lab.

Marco leads the research line Translational Data Science in Population Health at the Health Campus The Hague. This research line has three themes. First, in *Data Engineering* he investigates the further consolidation, standardisation and enrichment of the Extramural LUMC Academic Network (ELAN) data infrastructure, in line with national initiatives and in collaboration with his PHEG colleagues. Second, in *Data Analytics* he investigates Natural Language Processing and Machine Learning techniques for their suitability to answer current and novel types of translational research questions, especially from a democratising Data Science perspective, in collaboration with his LIACS colleagues. Third, in *e-Health Implementation* Marco designs and implements Data Science interventions through e-Health software solutions within the region in close collaboration with the Campus partners.

Until 2020 Marco worked as associate professor in the Natural Language Processing research group at the department of Information and Computing Sciences at Utrecht University, where he notably conducted numerous European-funded studies (OPERAM, SAF21, SMESEC, GEIGER, OPTICA) and nationally funded research projects (STRIMP, COVIDA). He participated in various leadership programmes and obtained academic qualifications such the Senior Research Qualification, Senior Teaching Qualification, and *Ius Promovendi*. From 2007-2018 he was an assistant professor Information Science, acting as the Information Science and Applied Data Science programmes manager for several years, among others.

From 2003-2007 Marco worked as a Ph.D. researcher in the Language Variation group of the Meertens Institute at the intersection of syntactic variation and dialectometry as a linguistic data scientist. In 2005 he notably received an Association for Literary and Linguistic Computing bursary award for his scientific work. Before 2003 he was active in industry for ten years as a Natural Language Processing and Big Data engineer at ZyLAB Europe B.V. and the Dutch Military Intelligence and Security Service, among others. In 1995 he graduated in Computational Linguistics at the University of Amsterdam.

Universiteit
Leiden
The Netherlands