Utrecht University

Focus area The Utrecht Platform for Applied Data Science (UPADS)

Starting Document

Peter van der Heijden (editor), Folkert Asselbergs, Joris van Eijnatten, Linda van der Gaag, Rick Grobbee, Steven de Jong, Derek Karssenberg, Chantal Kemner, Daniel Oberski, Marc Rietveld, Arnout van de Rijt, Mirko Schaefer, Floor Scheepers, Alexander Schönhuth, Berend Snel, Marco Spruit, Els Stronks, Sjoerd Verduyn Lunel, Roel Vermeulen

Contents

Chapter 1

Aim of focus area

Historically, all empirical sciences, from the natural sciences to the social sciences and humanities, have used data. Recently, however, scientific data's sheer volume, velocity, and variety, as well as the power of computers and methodology to generate scientific knowledge from the data, have started increasing drastically and in fact are expected to remain increasing over the coming decades.

Commensurately with this growing availability of data, data-analytic tools and computer capabilities, research questions across fields are becoming more complex, leading to commonalities in addressing problems encountered across a large number of fields represented at UU. This similarity of structure of apparently different problems is rapidly leading to the development of a new scientific field to deal with these problems regardless of object of research. We call this new field 'Applied Data Science', as opposed to fundamental research into data science.

For example, researchers analysing high-dimensional gene expression data use many of the same tools as those building automatic object recognition systems; text analysis advances are applied to track the history of ideas in Dutch newspapers, as well as to predict election outcomes using social media; and techniques used in environmental science models of water flow are also applied in studies of ancient settlements by archaeologists. Each of these problems originates from different Faculties within UU (Life Sciences, Natural Sciences, Humanities, Social and Behavioural Science, Geosciences) but uses similar tools, sometimes leading to existing collaborations, but more often than not operating in isolation.

The Utrecht Platform for Applied Data Science focus area aims to break this relative isolation and bring together researchers from all fields who apply Data Science. The aim is to form a community located in Utrecht University in which researchers communicate about their research, keep abreast of developments in Data Science, promote and develop teaching at various levels and, ultimately, collaborate and apply for external funding in national and international interdisciplinary calls such as the NWO or H2020. The impact and need for research that harvests rich data sources has been emphasized in several routes in the Nationale Wetenschapsagenda (the 'National Science Agenda') and in the VSNU memorandum on the Digital Society [4].

The objective is threefold. First, researchers can learn from each other in the ways complexly structured problems can be addressed, including the use of these statistical methodologies. Second, researchers who apply data science will be able to cooperate directly with those who do fundamental research into data science (e.g. computer scientists, statisticians), so that mutual strengths can be explored (see section 2.). Third, the Utrecht Platform for Applied Data Science will deal with cross-cutting problems that also involve external parties. For example, issues with the generalizability of psychological research may also be encountered by Statistics Netherlands; television and media increasingly use the internet to reach out as well as collect large volumes of data on their users; and high-content screening in pharmaceutical research must impute certain amounts of missing values, just as large-scale policy evaluation studies conducted by organisations such as TNO impute large amounts of missing information about citizens. The Utrecht Platform for Applied Data Science focus area will therefore also connect the UU with external parties to encourage the joint solution of similar problems using Data Science.

Chapter 2

Definition

Data Science as a research field has been defined as the extraction of actionable knowledge directly from data through a process of exploration and discovery, or through hypothesis formulation and hypothesis testing [1]. In order to explain the novel perspective of data science, consider the visualisation of the essential skills needed in Data Science shown in Figure 1. According to the standard scientific method, scientists formulate relevant research questions grounded in their domain expertise [red], then develop a corresponding study design which employs the appropriate statistical methods [yellow], after which data are collected and analysed accordingly [orange], to obtain the answers to their research questions.

Data Science also studies data that are not the result from the application of this standard scientific method. In many applications, data are studied that were collected for other purposes and without the primary aim to analyse them. Some data have been collected for administrative purposes, or simply originate from daily activities, such as data from Google or marketing. Data may further be heterogeneous (images, sounds, texts, numbers) and may have to be synchronised or calibrated. Data may also be analysed for different purposes than hypothesis testing. One may think of deriving predictions for the future from the data and of summarising and visualising the data from different perspectives.

Data Science further includes Engineering [blue] as an explicit dimension in addition to analysis. Engineering refers to implementing the data analyses, that is, to designing and developing data science systems, which may range from a simple script to implement an algorithm up to a fully functional analytic system prototype to address a research question.
In other words, data scientists build and/or use analytical software to empirically address questions from data. Here, statistical methods are intertwined with data mining and machine learning techniques as appropriate.

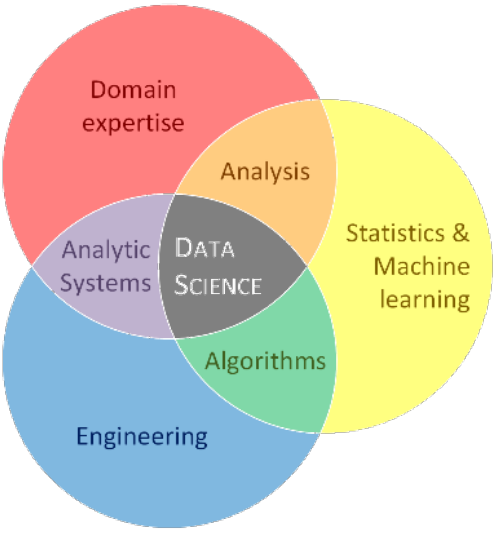


Figure 1: Data science (see [1])

The definition of *applied* data science encompasses all application 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.

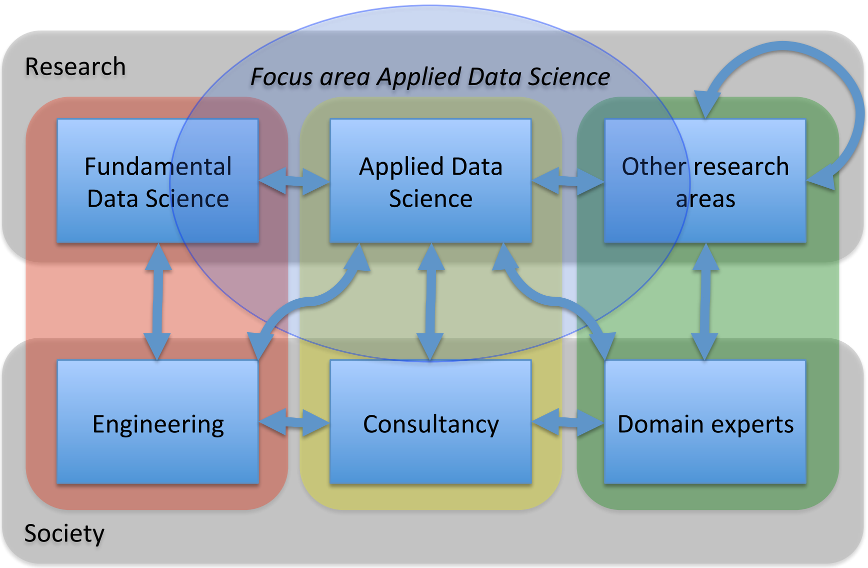
Chapter 3

Participating research groups

In the different scientific domains, expertise exists in specific areas of applied data science. Although present, this knowledge is currently scattered throughout the university. An important aim is therefore to bring researchers from different fields into contact with each other, so that they can share each other's knowledge.

Groups doing fundamental research in data science contribute to the focus area in two ways: first, by aiding the application of existing methods to substantive domains, and, second, by developing novel methods when the substantive domains require them. Regarding the latter: it is likely that bringing new research methodology into a domain will lead to change in the nature of the research questions asked. It is also expected that this development will bring forward research questions that cannot be answered with existing methodology. Here the expertise of groups doing fundamental research is essential.

The new focus area should contribute to the closure of the chain from fundamental data science through applied data science to the application of the achievements of data science within and outside the university. Cf. the following illustration, (Rietveld, 2016, personal communication):



3.1 Domains

Research groups from the following domains will participate:

Health

The emergence of so-called "omics", e.g., genomics, proteomics, metabolomics, in addition to ever more refined and even dynamic ways of capturing the human phenotype, e.g., by imaging, sensor techniques, and movement patterns through wearables is changing the scope and reach of the understanding of health and disease. By unlocking for research existing data assets of electronic health records (EHR) linked to national registries (hospital outcome, disease-specific mortality (CBS), pathology results (PALGA)), and adding these new omics and patient-generated (wearables) data, combined with advanced analytics, discovery and repurposing of drugs/ devices, design and conduct of trials, developing new diagnostics, and algorithms for clinical decision support to improve clinical practice for patient benefit will be accelerated . Differences between individuals in etiology of disorders and responses to medical interventions can be explored in unprecedented detail using real-world patient-centered data paving the way for personalized prevention and treatment. The sheer magnitude and diversity of data on individuals and groups challenge prevailing research approaches and data-analytical techniques. Access and linkage to data obtained in other domains such as geospatial data, social media and environmental data enable completely new and inter-disciplinary opportunities to study the role and interaction of the genome with comprehensive assessments of the environment: the exposome.

Bio-informatics

Life scientists aim at understanding how biomolecules and cells work in both healthy and diseased states, a first and necessary step to be able to develop truly novel therapeutic approaches, to invent more effective medicines and early stage diagnostics strategies, and to create plants with improved qualities. They do that by using a combination of structural biology, imaging, ‟ and ‟-omics” technologies, cell-cell interactions and deep phenotyping, together with computational modelling and bioinformatics that all generate large amount of data that need to be analysed. The relevance of the topic and infrastructure required has been recognized by the KNAW, which has put the Bioscopy proposal from Utrecht on the national infrastructure roadmap. Leading research groups with a strong bioinformatics / data science component have been brought together in the Utrecht Bioinformatics Center (UBC). They are located within Utrecht University, University Medical Center Utrecht, Hubrecht Institute and Princess Máxima Center for Pediatric Oncology.

Geosciences

Automated data collection, in particular by using sensor networks, wearable geolocated sensors, remote sensing, social media and web harvesting, has led to an exponential increase in data on system earth. This is changing research in both natural and social geoscience domains as this big data provides a spatial and temporal coverage and detail, which is orders of magnitude larger than available only a few years ago. As a result, research questions can be answered using new, more powerful, statistical techniques relying on enormous volumes of data, such as deep machine learning and real-time model data integration techniques. At the same time, the use of big data requires new software engineering techniques for integrating data and models at, for instance, high-performance supercomputers. Thus, data science, as a discipline integrating domain expertise, statistical techniques and software engineering is currently one of the corner stones of geoscientific research. At the same time, data science enables linking geosciences to other disciplines, for instance by combining measurements or models of human mobility and environmental data to calculate personal exposures to the environment, essential for health research. Also, access to methodologies from fundamental data science groups allow us to follow research approaches so far out of reach.

Utrecht Digital Humanities Centre (in formation)

The Utrecht Digital Humanities Centre stimulates the use of digital methodologies and digital reflection. Apart from research groups that span the breadth of the Faculty of Humanities, it includes the Digital Humanities Lab (engineering and research support), the Utrecht Data School (data analysis and visualization) and a computer scientist as Research Fellow. The Utrecht Digital Humanities Centre has strong ties with CLARIAH, eScience Center and the e-Humanities Platform of the KNAW.

Media

Media and culture studies critically investigate the epistemic impact of knowledge technologies. As such they already have developed conceptual frameworks for addressing datafication, and increasingly make use of digital methods and cultural analytics with (large) data sets from (social media) web platforms or digitized corpora of media texts. Further more they provide the very much needed domain-specific knowledge for cooperating with data scientists in key sectors such as media industries, creative industries, education, public management and civic participation. They expand the digital humanities research focus with the study of interaction data, digitized corpora, social media platforms in order to analyse datafication of public space, media audiences, publics, and cultural heritage. A distinctive feature is their inherently interdisciplinary approach and the exploratory and often experimental data practices in their research methods as well as the development of novel tools for data research.

Sociology
Social scientists in the field of social network analysis use large-scale data from social media such as Facebook and Twitter to test theories about the patterns of personal relationships, how these patterns change over time, and how they can be explained in terms of demographic characteristics of the individuals involved. Others use large volumes of text from mass media and social media and automated text analysis methods and natural language processing techniques to test theories about group differences -- by gender, race, nationality, news domain and political orientation -- in the quantity and sentiment of media coverage. Other use agent-based computational models -- computer simulations with agents who follow decision rules in their interaction with other agents -- to derive hypotheses from theoretical assumptions when formal mathematical derivation is infeasible.

Dynamics of Youth
In the Utrecht University Strategic theme Dynamics of Youth scholars from all seven faculties work at the forefront of fields like brain research, cognition, language/skill acquisition, identity formation, social interaction, and cultural factors, to integrate critical knowledge on child development. By connecting theoretical and practical perspectives, research in DOY will not only contribute to clarify the mechanisms underlying development, but also to identify the factors that either threaten or promote optimal developmental outcomes. To accomplish our aim of true interdisciplinary science, and to be able to create breakthroughs at the crossroads of disciplines, extensive integration of data of a very diverse nature that result from different expertise domains, is required.

3.2 Fundamental data science groups

The following research groups will participate:

Information science
Large-scale software production is investigated in the Organisation & Information research group of the Software Systems division in the department of Information and Computing Sciences. Applied data science research is mainly performed in the group's Data Science Lab. Aside from its work in learning analytics, the Lab's research mostly focuses on analytic systems, for which already strong collaborations have been established with other UU research groups within the field of Applied Data Science, including the departments of Geriatrics, Psychiatry, and Cell biology at the UMCU, as well as the Utrecht Bioinformatics Center, and Information and Technology Services. Fundamental research topics in these collaborations include natural language processing, knowledge discovery processes, semantic interoperability, analytic infrastructures, and meta-algorithmic modelling [3].

Computer Science
The computing-science research in the field of data science, as conducted in the Department of Information and Computing Sciences, has a focus on algorithmic foundations, methodologies and solutions. Current developments in data science pose many algorithmic challenges in terms of runtime characteristics: while algorithms with a runtime which is polynomial in the size of the input have sufficed in the past, the sheer volume and complexity of current data collections require sub-linear time solutions for access and analysis. From this perspective, the departmental data-science research addresses both traditional and non-traditional data, for both traditional and non-traditional data analytic tasks, such as pattern and pattern-set mining, trajectory discovery, model construction, and retrieval. While typically inspired by real-world applications, the primary goal is to develop algorithms and methodological solutions, which are applicable beyond a specific field of

application; yet, while more generally applicable, the practicability of the developed algorithms and methodologies is illustrated and validated through the motivating applications. Strong collaborations exist with for example the Utrecht Digital Humanities Centre; further collaborations with non-UU partners exist with the Central Veterinary Institute, GD Animal Health, and TNO Innovation for Life among others, mostly for the purpose of developing early-warning systems from data.

Applied Mathematics and Statistics

The research in applied mathematics is devoted to a scientific synthesis of applied analysis, statistics and data analytics. On one side of the synthesis are the traditional scientific methods based on hypothesis building, relying on unifying mathematical models describing links between mechanisms and measured data. On the other side of the synthesis is the advancement of data analytics. The size and number of datasets available for analysis grow exponentially, but current methods for storage, data handling and analysis are not adequate.

Statistical methods aim to make inference on the population based on data from a sample of the population and deals with all aspects of data: collection, analysis, interpretation, presentation, and organization. Within statistics methods have been developed to check the model fit, to deal with measurement error, heterogeneity and missing data, to assess bias etc. Within applied analysis, methods have been developed for model selection, model reductions, and for the analysis of time series using dynamical properties of the signal, effective in the classification and onset of diseases. Mathematicians and statisticians are used to work within various application areas, and collaborations are essential when the aim for method development is to answer relevant questions when applying the methods to the data. Currently, many statistical methods cannot deal adequately with big data yet, and the development of mathematical methods for data analytics is the prime priority of applied mathematics research.

Methodology and Statistics from FSBS

The research of the department of Methodology and Statistics is one of the five central research area of the Faculty of Social and Behavioural Sciences. Research in the Department of Methodology & Statistics covers the development of methods and techniques in the areas of survey data collection, latent variable modelling, multilevel modelling, longitudinal analyses, applied Bayesian statistics and the applicability of these techniques in social and behavioural sciences. There is strong collaboration with Statistics Netherlands. In the focus area this group will investigate methodological issues in statistics and data mining, such as model selection, dealing with measurement error, generalizability of results, data synthesis and data fusion, and solutions for missing data.

Chapter 4

Proposed activities

4.1 Community building

The focus area will build a community of researchers within Utrecht University, and create an ecosystem around this research, that facilitates the interaction between this research with other research groups and with society, stimulates the adaptation of new methodologies and techniques, generates new research, and actively engages in the pursuit of research funding.

Within Utrecht University we will build an active *applied data science centre of excellence* in which data scientists from different disciplines interact with each other, with students and with third-party professionals. The centre will build and maintain an active network with relevant companies and institutions and invite them to articulate questions in the area of data science. It will funnel socially relevant research questions to the data scientists of the UU and help them to translate these questions into viable research projects (ranging from PhD studies to major projects) and to find the required funding.

The centre will then seek student groups and scientists that want to adopt these questions and translate them into research projects. These projects may vary from smaller assignments for student groups to larger PhD-studies. Subsequently, it will help to find and engage possible sponsors for these projects.

The proposed *applied data science centre of excellence* will inspire researchers and students and facilitate collaboration and the exchange of knowledge. There is already the Utrecht Data School, which provides a basic training in data practices for BA students from various disciplines. The Utrecht Data School has developed a professional network of external partners where students apply obtained data practices in conducting commissioned research projects.
The Centre will:

- Have an off-line base on the UU campus, where people can actually meet and relate to each other;
- Offer a relevant and inspiring program of seminars / webinars, workshops, hackathons, excursions and summer schools;
- Facilitate all kinds of online activities and exchange around this program and within smaller groups around specific topics;
- Invite neighbouring research organizations (such as the HU, HKU, Hubrecht, Danone, RIVM, Deltares, TNO, KNMI, Dutch media and software companies) to participate.
- Invite government institutions to participate such as the city of Utrecht and Statistics Netherlands

To give this community credibility the support and active cooperation by top researchers from UU and UMC Utrecht is essential. External organisations are expected to be eager to participate in the community, since this is an easy way to scout new talent. Their involvement is welcome as long as they actively contribute to the community and their input is primarily about data science.
The *applied data science centre of excellence* also plays a role in teaching.

4.2 Teaching

The proposed *applied data science centre of excellence* will stimulate the application of data science within and outside of the university by offering a range of courses both in the regular BSc and MSc education, as well as in post-academic offerings.

Parallel to these courses the graduate schools of Life Sciences (GSLS) and Natural Sciences (GSNS) are developing new profiles and masters in the field of (applied) data science. These courses are very relevant to the Dutch economy, since there is growing need of skilled data scientists. At the department for Media and Culture Studies, a MA track Media, Data and Society is starting in September 2017. Cooperating with key partners from media industries, creative industries and public management this MA track will educate domain-specific experts for tackling data-related challenges in the affiliated sectors. On BA level, the Utrecht Data School provides already a training programme in data practices.

Close cooperation between the graduate schools and the research groups in the focus area will benefit both education and research in applied data science. Research subjects will be discussed in an early stage with students and students will contribute to research by doing projects and relevant internships. External organisations will be extra motivated to be involved in research projects, when this research is combined with relevant courses and internships.

4.3 Governance

The community will be governed by a small board with some of the best researchers in the field of applied data science.

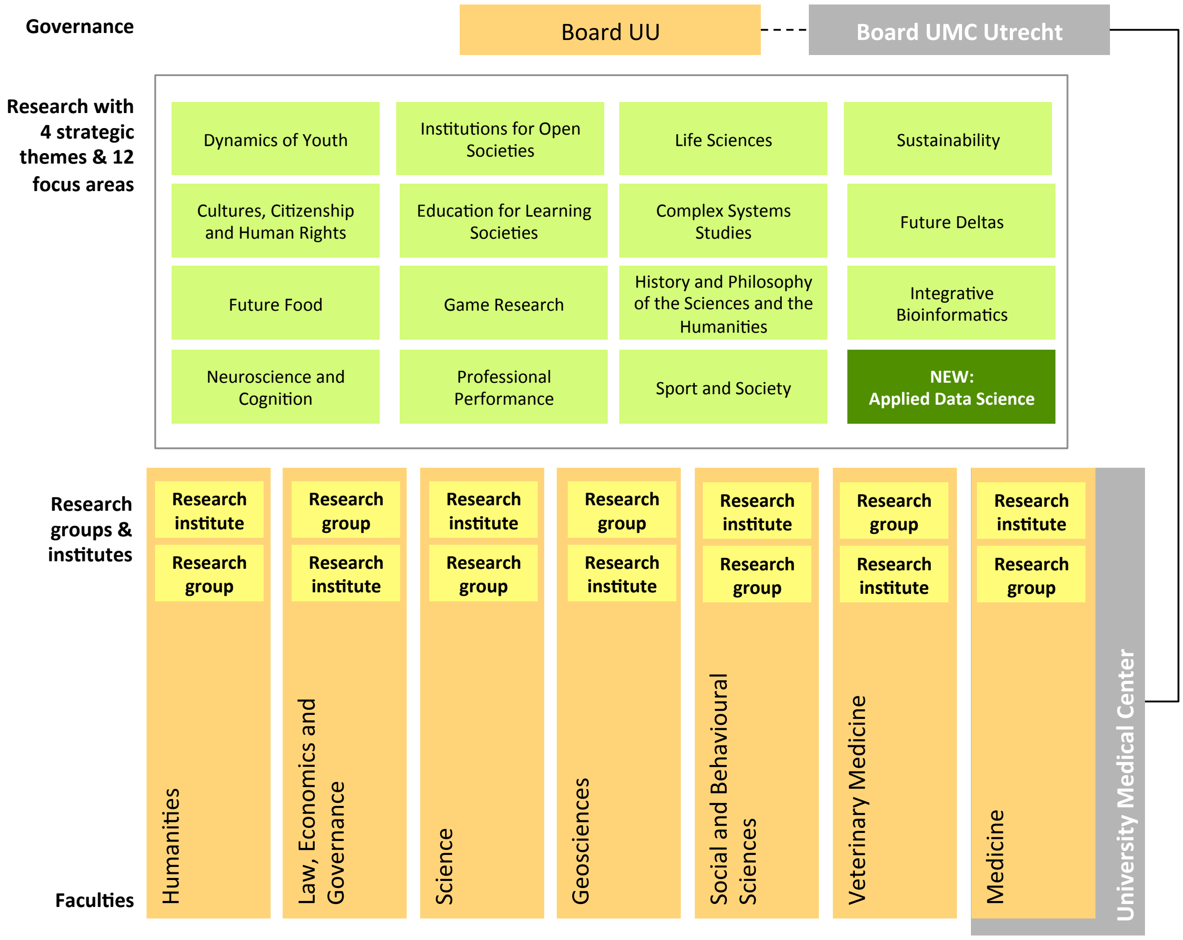
- HU and HKU will be invited to cooperate in the competence centre.
- ITS is willing and able to facilitate the organisation of the community.

Chapter 5

Applied data science in Utrecht University

5.1 Organogram

The focus area the Utrecht Platform for Applied Data Science (UPADS) is a new part of the organogram of Utrecht University:



5.2 Related focus areas

Similar to the *Foundations of Complex Systems* focus area, Applied Data Science approaches societal challenges with quantitative theories and models. In particular, Foundations of Complex Systems focuses on the societal themes *Climate Change, Infectious Diseases Dynamics, Social Network Interactions*, and *Big Data Modelling*. The latter theme allows for particular strong connections with the *Fundamental Data Science* part of the Applied Data Science focus area, which is associated with the topic of *Data Analytics*. The Applied Data Science focus area then aims to apply the appropriate statistical methods and machine learning techniques from fundamental data science research. In effect, the Foundations of Complex Systems focus area thus elegantly contributes to the efforts in the Applied Data Science focus area.

To further describe the connections between Data Science and Complexity, it is convenient to first divide Data Science into two parts: *Data Infrastructure*, which is concerned with ICT, organization, classification, storage and retrieval of data, and *Data Analytics*. The data analytics can be further divided into two subparts: "static" data, for which time is not a parameter, and "dynamic" data, for which time evolution of systems is the focus of interest. The dominant part of the static data analytics is for inference purposes. A classic example of this is the business analytics side of bol.com – they would want to know, e.g., given that out of their customers fifty people have bought the same set of four books, how likely is one customer, who has bought three out of the four books, is to buy the fourth book, so that they can target their advertising of the fourth book to this customer. Similarly, a classic example of dynamic data analytics is given that a blue-tongue disease pathogen is present in a population of sheep, how likely is it to develop into an epidemic in due course of time?

Since Complexity Science in general deals with such key concepts as emergence, transitions, resilience, predictability and control, an obvious bridge already exists therefore between Complex Systems and both "static" and "dynamic" data analytics. This is precisely the purpose of the theme *Big Data Modelling*, which is intended to benefit from the envisaged collaboration with the Applied Data Science focus area. A new bridge between the two focus areas can also be built between the infrastructure level for Data Science and the Complexity Laboratory Utrecht (CLUe), a facility that makes software and data analysis tools available to Complex Systems researchers within the UU. In all likelihood, there exist many algorithms and methodologies that can be used in both fields.

The *Integrative Bioinformatics for Life Sciences and Sustainability* focus has become essential for driving Life Science research, given that the need for bioinformatics has become abundantly clear in the last few years. The research in the focus area aims to provide deep insights into fundamental cellular and genetic processes by sharing knowledge and methods to integrate information at the level of DNA, genes, proteins, metabolites and cellular systems.. With the organization of the Utrecht Bioinformatics Center, Utrecht has developed a strong bioinformatics community, infrastructure and educational program. There are many similarities with the objectives of the Applied Data Science focus. However, bioinformatics concentrates on one specific research domain and also applies technologies from other technical domains such as modelling and mathematics. Bioinformatics should therefore be considered much more of a specialised domain, focused on Life Science research, that incorporates methodologies developed in technical domains such as the Applied Data Science focus area. The Utrecht bioinformatics community also houses researchers with explicit data science expertise such as machine learning (de Ridder, UMCU) or algorithms for genomic data science (Schoenhuth CWI/UU).

Recently, since 2007/8, genomics in particular have been witnessing a data revolution of a kind that was hard to predict before. Novel sequencing technologies ("next-generation sequencing") allow sequencing one's entire DNA in short time and at little cost (see for example "The DNA Data Deluge", http://spectrum.ieee.org/biomedical/devices/the-dna-data-deluge). In the meantime, the DNA that has been sequenced is soon to reach the exa(!)byte mark. The leverage for (personalized/stratified) health research arising from these data masses are enormous, as they, for the first time, allow to gain comprehensive understanding of the fundamental code of life. For example, huge, global-scale cancer genomics studies are under way that hold the promise for individualized, efficient cancer therapies; so are plenty of now much enhanced genome-wide association studies which try to directly link changes in the code with disease risks. At the same time, harnessing these gigantic heaps of sequence fragments poses intriguing research questions in areas of research surrounding modern genomics, such as computer science, mathematics and statistics. Clearly, the data revolution in genomics has affected the other 'omics' as well.

References

[1] Pritzker, P., and May, W. (2015). *NIST Big Data interoperability Framework (NBDIF): Volume 1: Definitions.* NIST Special Publication 1500-1. Final Version 1. National Institute of Standards and Technology.
[2] Demchenko, Y., Manieri, A., and Belloum, A. (2016). *EDISON Data Science Framework: Part 2. Data Science Body of Knowledge (DS-BoK).* Release 1. Horizon2020 project 675419.
[3] Spruit, M., and Jagesar, R. (2016). *Power to the People! Meta-algorithmic modelling in Applied data science.* In: Fred, A. et al. (Eds.), Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (pp. 400–406). November 11-13, 2016 in Porto, Portugal: ScitePress.
[4] Vereniging van Samenwerkende Nederlandse Universiteiten (2016). *De Digitale Samenleving*. Den Haag.

Steering committee

The focus area is governed by a steering committee with the following members:

- Prof.dr. Rick Grobbee (chair)
- Prof.dr. Peter van der Heijden
- Prof.dr. Arno Siebes
- Prof.dr. Marijk van der Wende

Peter van der Heijden is the leader of UPADS.

Contact

For more information contact Peter van der Heijden, P.G.M.vanderHeijden@uu.nl.