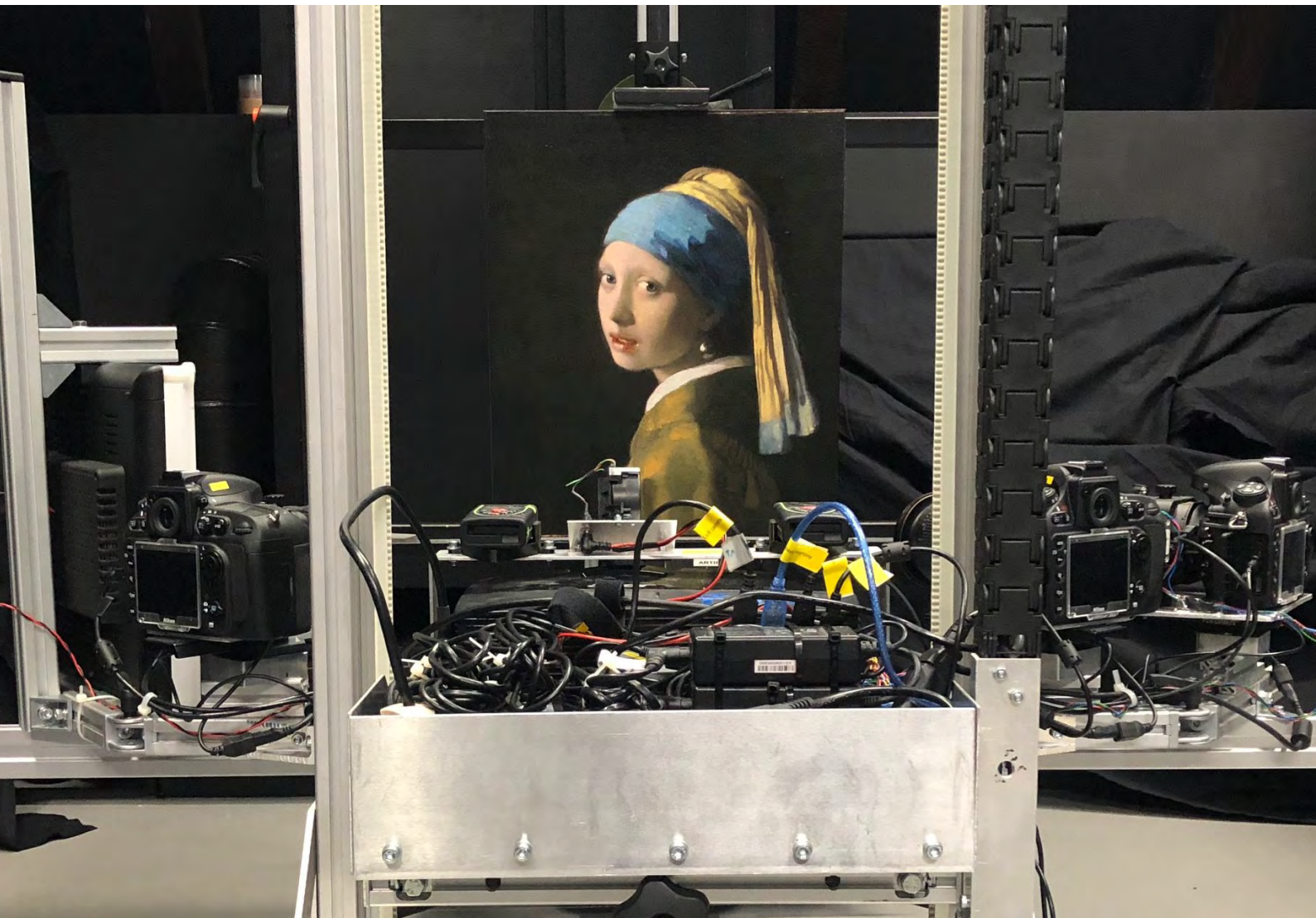


Summary & Case Studies

Self-Assessment Report Computer Science 2015-2020

Delft University of Technology

24 March 2022



Summary

The phenomena of *datafication* and *AI-zation* reflect the increasing tendency to quantify everything through data and to automate the decision-making processes that are also largely based on data. Since these phenomena have entered all segments of our lives and since research in computer science (CS) is at the heart of the technological developments underlying these phenomena, CS as a research field has gained strategic importance. TU Delft Computer Science operates at the forefront of these developments with the aim to help society at large, by enabling it to maximally benefit from these phenomena, while protecting it from potential risks. To that end, inspired and driven by the TU Delft core values of *Diversity, Inclusion, Respect, Engagement, Courage* and *Trust* (DIRECT), our mission includes (1) conducting world class research in selected computer science core areas; (2) maximizing opportunities for societal impact of our research; (3) providing rigorous, research-inspired engineering education in computer science; and (4) contributing to an international academic culture that is open, diverse and inclusive, and that offers openly available knowledge.

We are organized in two departments, *Intelligent Systems* and *Software Technology*, consisting of 5 and 6 *sections* respectively. Sections are small-scale organizational units, typically headed by a full or associate professor and marking a particular CS disciplinary scope. While the departments are separate units, they work closely together in research and education, and collaborate for societal impact. The convergence between the departments in terms of alignment and joint pursuit of strategic and operational goals has even become so strong over recent years that we can speak of an increasingly recognizable CS entity in Delft organizing its research into five main *themes* transcending the departmental and section boundaries: (1) decision support; (2) data management and analytics; (3) software systems engineering; (4) networked and distributed systems; and (5) security and privacy. The themes offer critical mass in order to achieve substantial impact, and each theme involves many researchers with various CS backgrounds and expertise.

Award-winning research in these themes achieved during 2015-2020 include a novel cross-modal (e.g., combining text and images) retrieval method based on adversarial learning; genetic algorithms for the automatic reproduction of software crashes to facilitate automated debugging; and Trustchain, a permission-less tamper-proof data structure for storing transaction records of agents with applications in digital identity. International recognition of our expertise is reflected by numerous leadership roles, e.g., as general or program chairs in numerous flagship conferences, such as AAAI, EuroGraphics, ACM/IEEE ICSE, ACM OOPSLA, ACM RecSys and ACM Multimedia. In the same time period, several staff members also received the highest (inter)national recognition in their fields, such as IEEE Fellow, membership of the Young Academy of the Royal Dutch Academy of Sciences, or the Netherlands Prize for ICT Research. Our scientific reputation also brought us into the consortia of two prestigious NWO Gravitation Projects ("NWO Zwaartekracht") of the Dutch Research Council, Hybrid Intelligence and BRAINSCAPES - the consortia that "*belong to the world top in their field of research or have the potential to do so*".

To maximize societal impact, we embrace eight key *sectors*: transport and logistics, energy, health and well-being, safety and security, finance, online education, creative industry, and smart cities. To enable and support us in making substantial interdisciplinary impact in these sectors, we have built up expertise, a network of collaborators and societal partners, and established the necessary organizational structures. Prominent examples of our impact in these sectors include the *NGS sequencing analysis pipeline* we designed and implemented as part of the NIPT test, which is used routinely by hospitals in several countries; *Cytosplore*, a software system for interactive visual single-cell profiling of the immune system; and *SocialGlass*, a tool suite for integration, enrichment, and sense-making of urban data. Our close ties with society are also reflected in our strategic collaborations with socio-economical partners, such as ING, DSM, Booking.com, Adyen, Ripple, Erasmus Medical Center and Leiden University Medical Center, leading amongst other things to strategic investments in the form of three large industry-funded labs (with ING, DSM and Booking.com) setup in the assessed time period for a duration of five years. Furthermore, we have invested extensive effort in public outreach, explaining and discussing science with a broad audience, and in particular in the context of complex societal debates in the domain of AI and blockchain. Finally, we play a leading role in regional, national and European initiatives, most notably in the Dutch AI Coalition (NLAIC).

In addition to scientific excellence and strong impact in the selected societal sectors, we are committed to (a) meeting the increasing societal need for highly skilled CS experts, (b) development of human capital

in our organization, leading to a new generation of international academic leaders, and (c) advancing the organization and academic culture, with the key pillars of open science, diversity and inclusion.

Regarding (a), we embraced an over 100% increase of our student population, but also aim at securing the highest possible level of their knowledge, skills and academic forming despite scaling up. Therefore, we value a close connection between research and education, and let both MSc and BSc students participate actively in our research. We also formulated an ambitious strategy, the realization of which would enable us to manage this education scale-up efficiently and effectively, leaving sufficient room to our staff for further developing scientific excellence and deploying it for societal impact. Part of this strategy is the growth of our academic staff towards 100 FTE by 2024 to meet the stabilization of the student numbers (due to *numerus fixus*). Between 2015 and 2020, we already achieved a net growth from 54 to 72 faculty members (+33%), with more to come in the upcoming years.

Next to BSc and MSc students, we are committed to delivering highly skilled CS experts at the PhD level. The number of PhD students grew from 105 to 165 (+57%) in the assessed time period, reflecting our ability to successfully acquire research funding in the present landscape. For our PhD students, the Graduate School defines a framework in which they can develop their skills next to conducting their thesis research. We strive towards completion of PhD theses within four years and organize our supervision, official moments of assessment, requirements on the volume and quality of the conducted research, as well as evidence of scientific impact through publications, accordingly.

Regarding (b), development of human capital: as computer science expertise is in high demand across the globe, finding strong new people as well as retaining our current staff proved highly challenging, especially given the high teaching load due to our record student intake. Therefore, acquiring, developing and retaining academic talent has been one of our most important goals. Dedicated actions, such as devising of a *Development Track Plan*, serve to empower each staff member to provide contribution to the organization in his/her own way, based on individual interests, talents and ambitions, and in view of our joint ambition as organization.

In view of (c), our organization, we embrace open science, with a substantial percentage (80% in 2020) of our articles available as open access, and by making numerous software tools and data sets openly available. We are a highly international organization with employees and students from all over the world. We strive to be an inclusive organization, where staff and students feel at home and valued, regardless of their background, age, gender, sexual orientation or functional disability. In terms of female faculty, we realized a net growth from 11 to 14 faculty members. As the number of men employed also increased, the percentage of female faculty stayed stable at around 20%. We consider this too low. We are committed to addressing this, for which we will take a long-term approach with, amongst other means, dedicated budget reserved for continued openings for female faculty in the upcoming years.

We are proud of our scientific successes and societal impact in the core computer science disciplines as well as in interdisciplinary research in our target societal sectors. This is especially so as those were achieved in a period that was transformational for CS@Delft, characterized by substantial growth and development across our organization and activities. We anticipate an even stronger societal demand for our research and expertise in the future. We will therefore continue to initiate, participate and take on a key role in effective and interdisciplinary partnerships at the university (TU Delft AI), regional (LDE), national (ICAI, IPN), and European (ELLIS, CLAIRE) levels. Furthermore, we will continue the growth path for our staff, in order to build up capacity enabling us to further develop our scientific excellence and offer our strongly increased student population the world-class research-intensive education they deserve. To achieve this, we center the next steps in our ongoing transformation around *people*, *organization*, and *profiling* and identify seven key actions for the upcoming years that aim at (1) improving our attractiveness as an employer; (2) improving diversity and inclusion; (3) improving the execution of the PhD program; (4) expanding our staff capacity; (5) aligning our office space with the optimal way of working; (6) articulating the scientific profile; and (7) boosting our scientific and societal impact.

Prof. dr. Alan Hanjalic
Head, Intelligent Systems Department
Delft University of Technology

Prof. dr. Arie van Deursen
Head, Software Technology Department
Delft University of Technology

F.1 Data Science unraveling Cancer

Cancer is a devastating disease manifested by cells acquiring an accumulation of various biological capabilities that make them grow indefinitely resulting in a tumor that eventually metastasizes throughout the body. Various genomic instabilities drive the acquisition of these hallmarks, such as genomic alteration, duplication, and epigenetic changes.

The TU Delft Bioinformatics group led by Marcel Reinders applies data science technologies to find causative variations that are hidden in these Big Data. The group has access to various data acquired from tumors, as well as cell lines and mouse models through, amongst others, its long-standing relationship with the Netherlands Cancer Institute (NKI), institutionalized by the part-time appointment of Lodewyk Wessels at the TU Delft. The well-defined NKI datasets are combined with public repositories, such as, for example, The Cancer Genome Atlas (TCGA), to differentiate patients into groups that have different outcomes or react differently on therapies to contribute to personalized treatment plans.

One of the achievements of the TU Delft Bioinformatics group is the development of a statistical tool (RUBIC, Figure F.1, [1]) that can detect frequent recurrence of copy number aberrations across tumor samples which act as a reliable hallmark of certain cancer driver genes. Our breakthrough was that we changed perspective to detecting recurrent copy number *breaks*, rather than recurrently amplified or deleted *regions*, which we combined with a statistic based on the Euler characteristic.

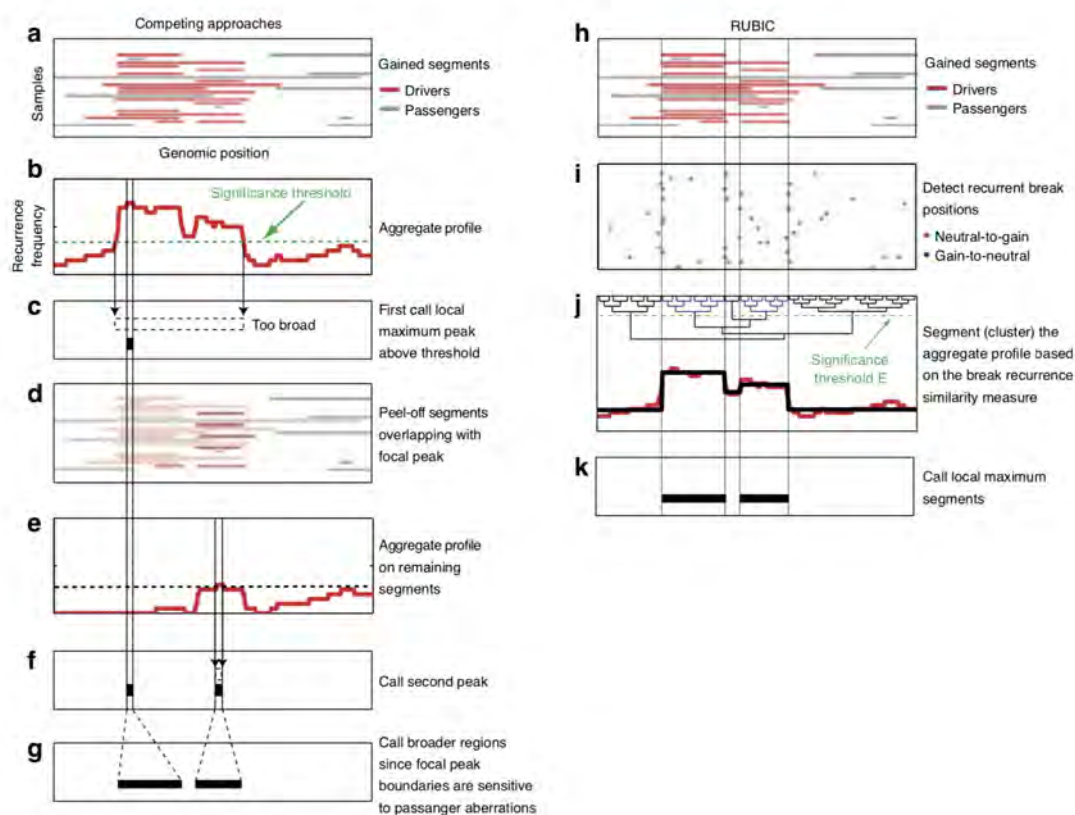


Figure F.1: The steps of RUBIC (Recurrent Unidirectional Break Identification by Clustering) for detection of driver genes (image from [1]).

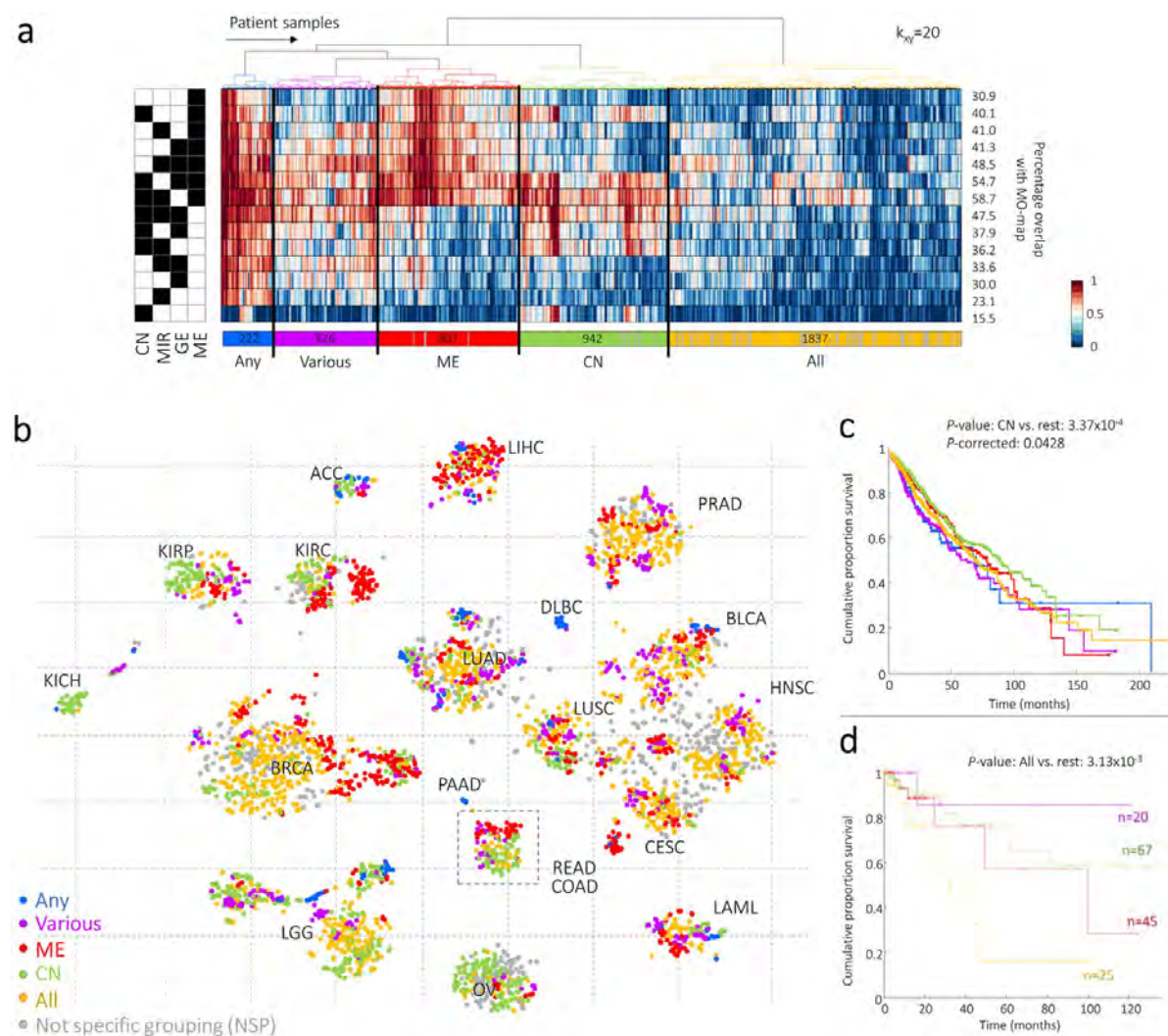


Figure F.2: Key MEREDITH molecular tumor analysis results (image from [2]).

Another innovation by the team of Reinders is the joint analysis of different molecular characteristics of a tumor for which they introduced a novel multi-omic integration approach (MEREDITH, Figure F.2, [2]) supporting visual exploration of the data as well as evaluation of the contribution of the different genome-wide data-types.

The team also proposed a novel computational framework for clinical drug response prediction. This framework employs non-linear manifold learning to capture biological processes active in both pre-clinical models as well as human tumors. Using this non-linear alignment predictors built on cell line response data only can be transferred to human tumors. This framework helps improving personalized therapies as well identification of biomarkers to, for example, resistance of drug therapies.

- [1] Ewald van Dyk, Marlous Hoogstraat, Jelle ten Hoeve, Marcel J.T. Reinders, and Lodewijk F.A. Wessels. "RUBIC identifies driver genes by detecting recurrent DNA copy number breaks". In: *Nature Communications* 7.12159 (2016). Open access: [link](#); #Cites 11, #7 overall, #3 in Life Sciences & Earth Sciences, h5:273
- [2] Erdogan Taskesen, Sjoerd M.H. Huisman, Ahmed Mahfouz, Jesse H. Krijthe, Jeroen de Ridder, Anja van de Stolpe, Erik van den Akker, Wim Verheagh, and Marcel J.T. Reinders. "Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics". In: *Nature Scientific Reports* 6.24949 (2016). Open access: [link](#); #Cites 27, #28 overall, #5 Life Sciences & Earth Sciences, h5:178

F.2 AI for Fintech Research

AI for Fintech Research (AFR)^a is a collaboration between Dutch bank ING and Delft University of Technology. The mission of AFR is to perform world-class research at the intersection of Artificial Intelligence (AI), Data Analytics, and Software Analytics in the context of FinTech. Funded by ING, AFR was launched in January 2020, and will run for a period of five years. AFR is led by Arie van Deursen (TU Delft, scientific director), Luís Cruz (TU Delft, scientific manager), and Elvan Kula (ING, lab manager).

Advances in AI and data analytics are redefining the financial services sector. With 36 million customers, activities in 42 countries, and a total of 50,000 employees of which 15,000 work in Information Technology, software and data are at the heart of ING's business and operations. AI enables ING to predict customer wants and needs, while becoming a safer and more compliant bank. In this context, AFR seeks to develop new AI-driven theories, methods, and tools in large scale data and software analytics.

The core of AFR consists of 10 research tracks, in which 10 PhD students work topics such as software analytics, data integration, fairness in machine learning, model life cycle management, regulatory compliance, AI-based software engineering, and A/B testing. In each track, both TU Delft faculty members and ING engineers are involved, as well as bachelor, master, and PhD students.

As an example, in the *software analytics track*, we analyze software development data collected from hundreds of teams at ING to both understand and improve the efficiency of development processes. We combine static code analysis data with survey data collected from developers at ING to identify factors and their relationships affecting effort estimation and on-time delivery. Based on these factors, we build models to predict whether software deliveries will suffer delay [2], and formulate guidelines on streamlining software project management. This track is a good illustration of the interactive and participative nature of the collaboration between the students in AFR and developers at ING.

In the *data integration track*, we seek to bring together disparate data sources, which is needed in order to be able to employ machine learning methods to data lakes at the scale as in use at ING. To that end, one of AFR's first results is Valentine^b [1]. Valentine offers a range of open source implementations of state of the art data matching algorithms, as well as synthesized data sets with ground truths that we used to experimentally establish the strengths and weaknesses of these algorithms – Figure F.3 shows screen shots of data fabrication, experimental configuration, and result presentation. Valentine illustrates the essential role open science plays in our work (also with industry), as well as the prominent role students play in AFR (besides PhD student Siachamis two master students, Ionescu and Psarakis, participated in this research).

AFR is a member lab of the Innovation Center for Artificial Intelligence (ICAI^c). Launched in 2018, ICAI hosts over 25 labs, each with five or more PhD students, and each a collaboration between one or more Dutch knowledge institutes and industrial or societal partners. AFR benefits from and contributes to this highly successful eco-system of industrial AI labs.

- [1] Christos Koutras, George Siachamis, Andra Ionescu, Kyriakos Psarakis, Jerry Brons, Marios Fragkoulis, Christoph Lofi, Angela Bonifati, and Asterios Katsifodimos. "Valentine: Evaluating Matching Techniques for Dataset Discovery". In: *37th IEEE International Conference on Data Engineering, ICDE 2021*. IEEE, 2021, pp. 468–479. DOI: 10.1109/ICDE51399.2021.00047. Open access: [link](#); #Cites 2, #6 in Databases & Information Systems, h5:56
- [2] Elvan Kula, Eric Greuter, Arie van Deursen, and Georgios Gousios. "Factors Affecting On-Time Delivery in Large-Scale Agile Software Development". In: *IEEE Transactions on Software Engineering* (2021). Open access: [link](#); #Cites –, #6 in Software Systems, h5:52

^a <http://se.ewi.tudelft.nl/ai4fintech/>

^b <https://delftdata.github.io/valentine/>

^c <https://icai.ai>



Figure F.3: Using Valentine for conducting tabular data matching experiments: a) data fabrication; b) configuration of experiments; c) presentation of findings.

F.3 Responsible AI in Psychological Assessment

With digitization increasingly being integrated in society, the technologies researched by CS@Delft are increasingly transferred and applied at scale to real-world use cases. While this is happening, questions and concerns are arising on whether this transfer is happening properly and responsibly.

With multimedia data often representing humans or human information, and multimedia search engines and recommender systems affecting automated decision-making based on such information, these developments are of strong interest to the Multimedia Computing (MMC) Group, especially in Cynthia Liem's team. In its academic activities, the team has been striving to explicitly include interdisciplinary perspectives. With Liem's background in music information retrieval, this interdisciplinary expertise already had been built in connection to the humanities, leading to research lines on inclusive information filtering and validity of measurement procedures. During 2015-2020, the latter research line expanded and solidified in connection to the social sciences, regarding computational approaches to psychological assessment.

Because of the MMC group's earlier track record on audiovisual and affective multimedia content analysis, Liem was approached by colleagues in organizational psychology at the Erasmus School of Social and Behavioural Sciences, who were researching the impact of AI on hiring procedures, and fairness considerations with regard to video resumes. Joint interests led to TU Delft becoming the technical partner in the ERASMUS+ 'Big Data in Psychological Assessment' (BDPA) project. With a prime focus on education innovation for psychology curricula, the BDPA project's outcomes needed to be accessible to psychologists who would be confronted with big data, but would not have extensive computer science training. Joint pilot collaborations on this topic included participation in a benchmarking challenge, presented at a prestigious computer vision conference (ChaLearn Looking at People Competition, CVPR 2017^a).

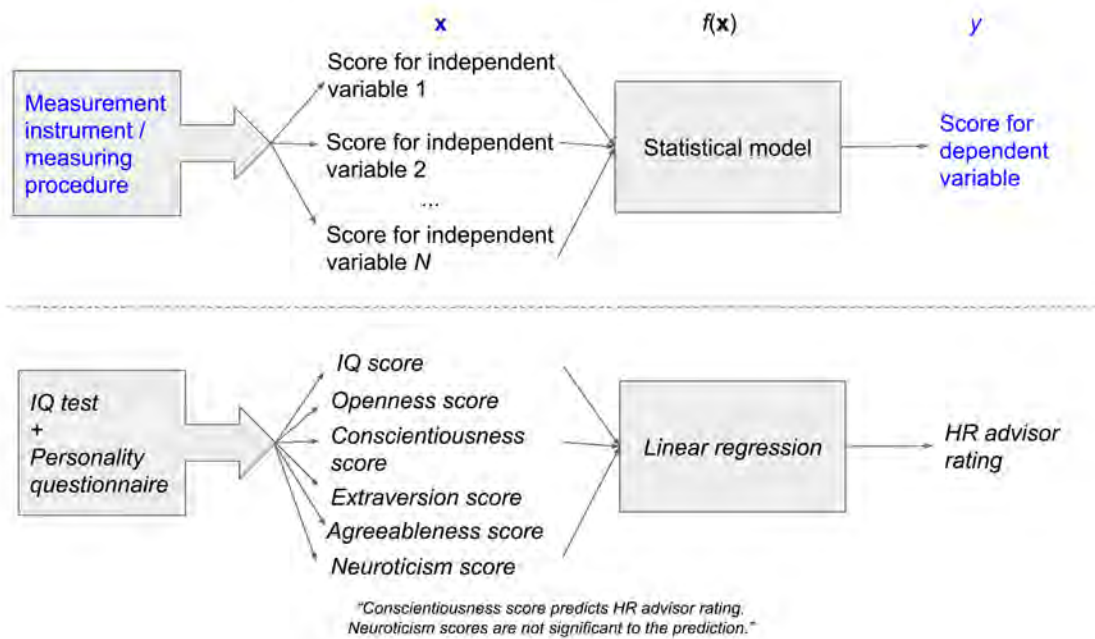
However, the most significant outcome was the realization that psychologists and data scientists consider similar research pipelines, but put their main focus on different parts of these pipelines, also seeking to draw different types of conclusions from them (Figure F.4). This insight, with examples of what realistically can and cannot be researched in data-driven setups, has been documented in a book chapter [2], which is increasingly being integrated in psychology curricula and referenced in courses for HR professionals.

The BDPA project led to both Liem and MMC PhD candidate Andrew Demetriou (a social psychologist by training) gaining considerable experience in educating non-technical audiences about AI fundamentals and applications, and technical audiences about methodological considerations from other disciplines. This expertise has broadly been sought out, e.g. leading to invited talks at the International Convention of Psychological Science, the University of Bucharest and the Dutch-Flemish Network for Selection Research, and participation in a video series^b by Deloitte. It also led to the European Personnel Selection Office (EPSO) remaining in touch with TU Delft as a trusted collaboration partner, paving the way for further research into responsible (multimedia) AI integration in public-interest applications.

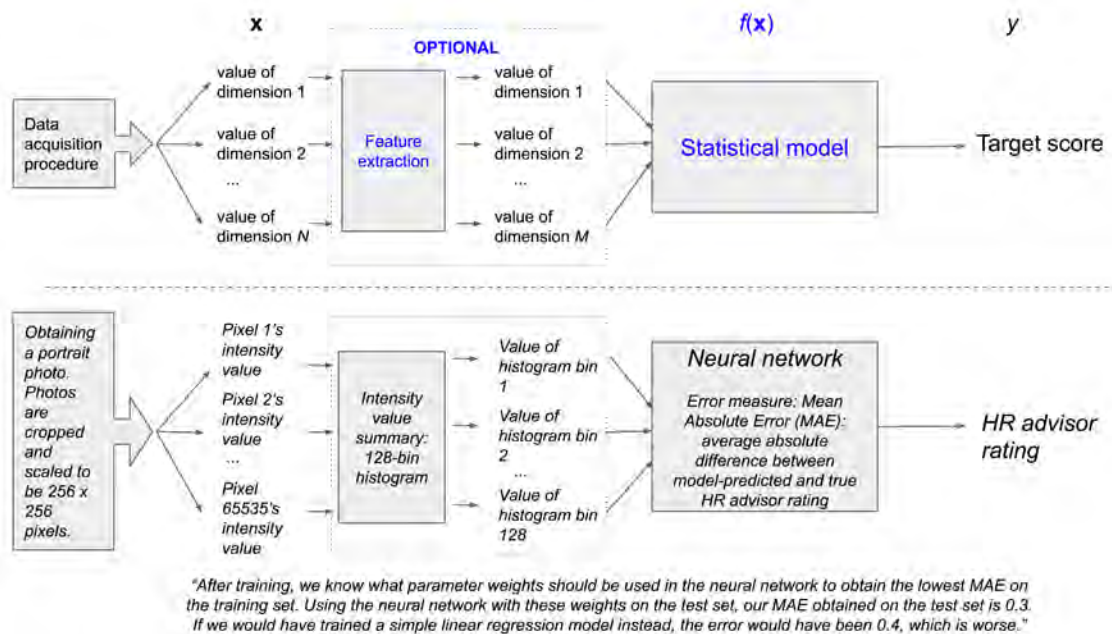
- [1] Cornelius J. König, Andrew M. Demetriou, Philipp Glock, Annemarie M. F. Hiemstra, Dragos Iliescu, Camelia Ionescu, Markus Langer, Cynthia C. S. Liem, Anja Linnenbürger, Rudolf Siegel, and Ilias Vartholomaïos. "Some Advice for Psychologists Who Want to Work With Computer Scientists on Big Data". In: *Personnel Assessment and Decisions* 6 (1 2020). DOI: <https://doi.org/10.25035/pad.2020.01.002>. Open access: [link](#); #Cites 3
- [2] Cynthia C. S. Liem, Markus Langer, Andrew Demetriou, Annemarie M.F. Hiemstra, Sukma Achmadnoer Sukma Wicaksana, Marise Ph. Born, and Cornelis J. König. "Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening". In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. The Springer Series on Challenges in Machine Learning. Springer, 2018, pp. 197–253. ISBN: 978-3-319-98130-7. DOI: 10.1007/978-3-319-98131-4_9. Open access: [link](#); #Cites 21

^a <https://chalearnlap.cvc.uab.cat/dataset/24/description/>

^b <https://www.youtube.com/watch?v=DUv84dG31TM>



(a) Psychology (in an organizational psychology application).



(b) Machine learning (in a computer vision application).

Figure F.4: Prediction pipelines in psychology and machine learning. Abstracted pipelines are given on top, with typical domain-specific research focus areas marked in blue. Simplified examples of how they may be implemented are illustrated at the bottom, together with a typical conclusion as would be drawn in the domain.

F.4 Algorithms for Coordination in a Smart Power System

Caused by the shift to renewable generation, electricity demand needs to be matched to the generation, instead of generating power when it is demanded. For this, the energy system needs to be more smart, scheduling flexible demand depending on weather predictions, taking network capacity into account. This results in algorithmic challenges regarding the autonomous control of components of a large-scale system; components that are adaptable and self-learning. Moreover, there are also significant challenges in supporting strategic decisions regarding such a system under extreme uncertainty, and where different interests play a role. Within the Algorithmics group we contribute to these algorithmic challenges.

A concrete, pioneering example is our work on constrained Markov Decision Processes [2]: De Nijs, now assistant lecturer at Monash University, developed a number of agent-based algorithms that can, for example, be used to control heat pumps such that the capacity of the electricity network is not violated and comfort is optimized. His PhD thesis work, funded by the distribution network company Alliander, concentrates on the fundamental problem of coordinating the use of resources over time with such a capacity limit, while dealing with uncertainty regarding the needed resources (caused by heat loss) as well as the available capacity (caused by other use of the network). The better understanding of this problem has led to a complexity analysis [3], and a number of new algorithms.^a The most efficient method scales well with both the number of households as well as with the decision horizon (number of hours, see Figure F.5), making it possible to coordinate hundreds of heat pumps to stay within the capacity limit of the electricity network, where a standard optimization approach cannot deal with more than ten (see Figure F.6 for a simulation result of this adaptive planning method against others). Results are not just published as papers, but also include code and benchmark data, such as also done for our work on methods which minimize electricity costs, by scheduling demand using price information from the electricity markets.^b

Apart from direct funding from industry, funding for this research typically comes from NWO thematic programs, or via RVO from the ministry of Economic Affairs and Climate. At the time of writing there are five PhD students working on this theme in the Algorithmics group. Within the university this research is connected to other energy-related research, among others through the PowerWeb insitute. Nationally, this type of research on the use of Artificial Intelligence for challenges regarding Energy & Sustainability is connected by a working group of the Dutch AI coalition (NL AIC), which is the organization behind the AiNed proposal, funded by the “Groeifonds”. This coalition is working on defining new programs for AI research and innovation to strengthen the Dutch economy on the long term.

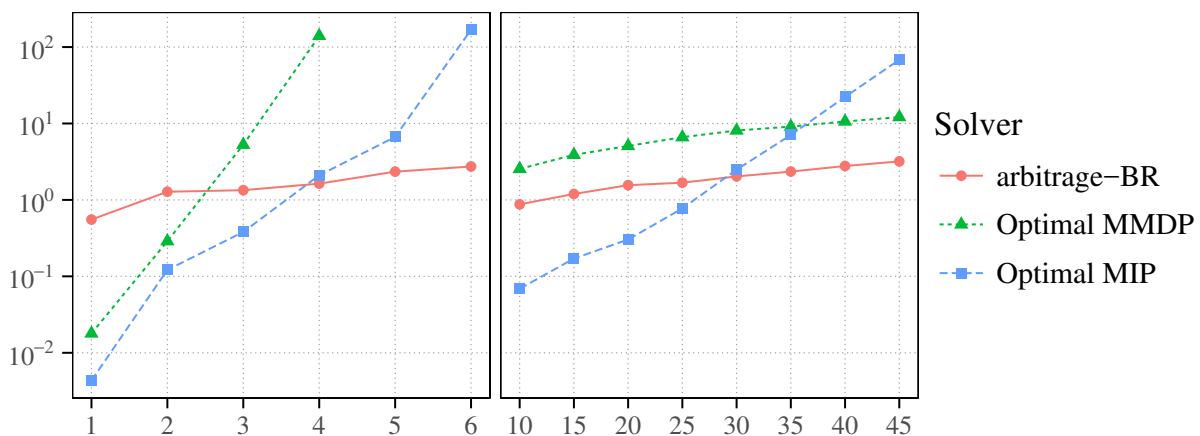


Figure F.5: The runtime of known optimal methods does not scale with the number of agents, while the arbitrage-BR method scales reasonably well in both agents and horizon.

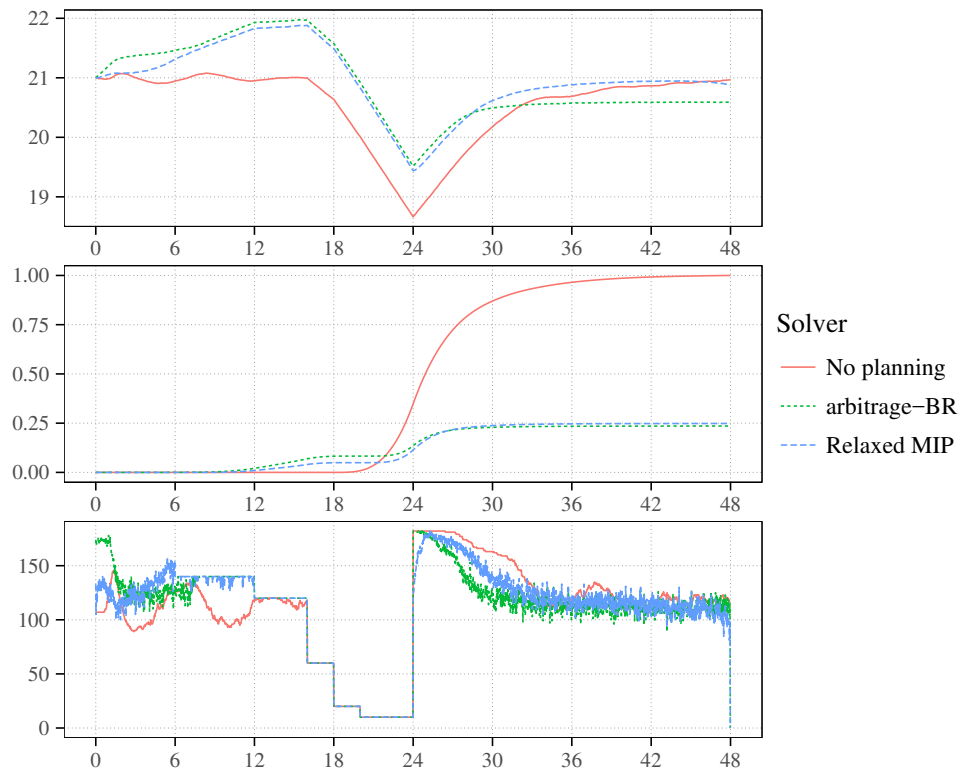


Figure F.6: In this simulation of 182 households with heat pumps the target temperature is 21 degrees for all houses, but a gradual decrease of the available network capacity occurs starting from hour 6, and at hour 20 the minimum capacity is reached and only 10 out of 182 heat pumps are allowed to be switched on. These plots show the average indoor temperature, normalized cumulative penalty, and total load of two different planning methods, compared to not planning at all.

- [1] Germán Morales-España, Álvaro Lorca, and Mathijs de Weerd. “Robust unit commitment with dispatchable wind power”. In: *Electric Power Systems Research* 155 (2018), pp. 58–66. Open access: [link](#); #Cites 44, #3 in Power Systems, h5:59
- [2] Frits de Nijs, Erwin Walraven, Mathijs de Weerd, and Matthijs Spaan. “Constrained Multiagent Markov Decision Processes: a Taxonomy of Problems and Algorithms”. In: *Journal of artificial intelligence research* 70 (Mar. 2021), pp. 955–1001. Open access: [link](#); #Cites 0, #11 in Artificial Intelligence, h5:64
- [3] Mathijs de Weerd, Michael Albert, Vincent Conitzer, and Koos van der Linden. “Complexity of scheduling charging in the smart grid”. In: *Twenty-Seventh International Joint Conference on Artificial Intelligence*. 2018, pp. 4736–4742. Open access: [link](#); #Cites 16, #3 in Artificial Intelligence, h5:95

a <https://github.com/AlgTUDelft/ConstrainedPlanningToolbox>
b <https://github.com/AlgTUDelft/B-FELSA>