Lfd.Nr. IKT/01-2014

Conquering Data in Austria

Technologie-Roadmap für das Programm IKT der Zukunft: Daten durchdringen - Intelligente Systeme

Dr. Helmut Berger, Dr. Michael Dittenbach, Dr. Marita Haas

max.recall information systems GmbH
Künstlergasse 11/1, A- 1150 Wien

Dr. Ralf Bierig, Dr. Allan Hanbury, Dr. Mihai Lupu, Dr. Florina Piroi

Technische Universität Wien
Institut für Softwaretechnik und Interaktive Systeme
Favoritenstraße 9-11/188, A-1040 Wien

Wien, Jänner 2014

Bundesministerium für Verkehr, Innovation und Technologie

Acknowledgements

This study was commissioned and funded by the Austrian Research Promotion Agency (FFG) and the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT).

We thank the following for their valuable input: members of the project advisory board, participants in the three workshops held in Salzburg, Graz and Vienna, the key note speakers, the people who responded to the online survey, and those that sent comments on and corrections to the initial Position Paper. A very special thanks is due to those Austrian companies that generously shared their valuable expertise and industry perspective on the future data challenges with us.

Contact

For more information, contact the project leader:

Dr. Helmut Berger
max.recall information systems GmbH
Künstlergasse 11/1
A-1150 Vienna, Austria
phone: +43 1 2369786
e-mail: h.berger@max-recall.com

About the Authors

Helmut Berger is Co-founder and CEO of max.recall GmbH. He completed his doctoral studies at the Vienna University of Technology in 2003. His research areas include: Semantic Information Systems and Content Analytics. He has substantial project management experience and more than 60 publications.
Ralf Bierig is a researcher at the Vienna University of Technology. He completed his doctoral studies at the Robert Gordon University, UK in 2008, and has postdoctoral experience in the USA and Denmark. His research areas include: Interactive Search and Multimodal Search. He has more than 20 publications.
Michael Dittenbach is Co-founder of and Information Access Engineer at max.recall GmbH. He completed his doctoral studies at the Vienna University of Technology in 2003. His research areas include Neuro Computing, Content Analytics and Information Retrieval. He has substantial project management experience and more than 60 publications.
Marita Haas is Gender Research Consultant at max.recall GmbH. She completed her doctoral studies in Economics and Social Sciences in 2006. Her research areas include: Biography Research, Female Biographies, Life Stories, Gender and Work & Life Balance. She has over 20 publications.
Allan Hanbury is a senior researcher at the Vienna University of Technology. He completed his doctoral studies in Applied Mathematics at the Mines ParisTech, France in 2002, and was granted the habilitation in Computer Science from the Vienna University of Technology in 2008. His research areas include: Vertical Search, Multimodal Search and IR Evaluation. He leads large international research projects as well as national research projects. He has over 130 publications.
Mihai Lupu is a researcher at the Vienna University of Technology. He completed his doctoral studies with the Singapore-MIT Alliance in 2008. His research areas include Vertical Search, Multimodal Search and IR Evaluation. He has more than 25 publications.
Florina Piroi is a researcher at the Vienna University of Technology. She completed her doctoral studies at Johannes Kepler University Linz in 2004. Her research areas include Vertical Search and IR Evaluation. She has over 20 publications.

Copyright

Executive Summary

A comprehensive roadmap study on Intelligent Data Analytics technologies is provided. This study, commissioned by the Austrian Research Promotion Agency (Österreichische Forschungsförderungsgesellschaft, FFG) and the Austrian Federal Ministry for Transport, Innovation and Technology (Bundesministerium für Verkehr, Innovation und Technologie, BMVIT), provides objectives for the short-, medium- and long-term focus (year 2025) of the FFG funding programme *ICT of the Future: Conquering Data - Intelligent Systems* (IKT der Zukunft: Daten durchdringen – Intelligente Systeme). The results presented in this work arose out of a mix of approaches that included an exhaustive literature review, interactions with stakeholders through an online survey, workshop discussions, and structured expert interviews. This technology roadmap brings the technology perspective and the perspective of the area's stakeholders (public, research, industry) together. It identifies the requirements for new ICT in this area and presents a selection of expected developments, requirements, and guidelines in the ICT field.

Surveying the Intelligent Data Analytics field, we have analysed the relevant methods and techniques and categorised them into four (interacting) groups: *Search and Analysis*, *Semantic Processing*, *Cognitive Systems and Prediction*, and *Visualisation and Interaction*.

The coverage of Data Analytics applications, on which Austrian companies, research institutes, and universities focus, has a rather wide range. These application areas were reviewed with respect to how they currently handle data and how they make use of Intelligent Data Analytics. Healthcare, Energy and Utilities, eScience as well as Manufacturing and Logistics were identified to be the most important application domains in Austria.

The most important challenges in Intelligent Data Analytics were summarised by aggregating the different stakeholders' viewpoints on data. These challenges range from Privacy, Security and Data Ownership over algorithmic and technological shortcomings to shortages in the supply of qualified personnel.

During this study, a comprehensive landscape of Austrian competences in Intelligent Data Analytics was compiled. This competence landscape covers Austrian research institutes, universities, universities of applied sciences as well as commercial service providers operating in Austria. Austrian strengths are in the areas of statistics, algorithmic efficiency, machine learning, computer vision and Semantic Web.

Based on the analysis, nine roadmap objectives that span over the short, medium and long term are made. These objectives cover three primary areas: *Technology*, *Coordination*, and *Human Resources*.

The first four objectives cover technological topics that aim at i) the advance of the current data integration and data fusion capabilities, ii) at the increase in algorithm efficiency, iii) at turning raw data into actionable information, and iv) at automating the knowledge workers' processes. These engineering-focused objectives require dedicated R&D funding, which will,

on the mid to long term future, result in novel, Austrian-made lead technologies in the area of Intelligent Data Analytics.

The next three objectives focus on measures supporting the stakeholders' capabilities to innovate and extend their competitive position. These measures aim at improving Austria's visibility, integration, and attractivity in the international ICT research and development context. They are coordination-oriented objectives that require investment in order to build an Austrian Data-Services Ecosystem. The Ecosystem will make data accessible and interoperable in order to generate greater economic value. Further objectives involve the elaboration of a legal framework for dealing with data, and the launch of various initiatives—including a dedicated "Austrian Data Technologies Institute"—which will strengthen the networking of and know-how exchange between Austrian and international stakeholders in the field.

The remaining two objectives cover the area of Human Resources and aim at addressing the urgent need for highly qualified personnel in data technologies. They advocate investment in novel education programmes that assist in creating polymath thinkers capable to cope with the requirements emerging from (Big) Data Analytics. The second of these two objectives presents actions to improve the gender and diversity awareness in the field of Intelligent Data Analytics.

Potential lighthouse projects are presented as a route to achieving some of these objectives. These include a broad impact lighthouse, the Data-Services Ecosystem, which allows cross-fertilisation of technologies between application domains. Furthermore, application-specific lighthouses, which channel the development work towards solving challenges in a specific domain of application, are described. The suggested application domains for application-specific lighthouses are manufacturing, energy, healthcare and digital humanities.

In summary, Intelligent Data Analytics has the potential to greatly benefit the Austrian society and economy. It is essential for a successful innovation economy to provide the ecosystem in which data-centred innovation and technology transfer can take place. There are still many challenges to overcome from both a technological and societal point of view before Austria is ready to take full advantage of this opportunity.

Kurzdarstellung

In der vorliegenden Studie wird die österreichische Technologie-Roadmap für den Bereich "Intelligent Data Analytics" beschrieben. Diese Studie wurde von der Österreichischen Forschungsförderungsgesellschaft FFG und dem Bundesministerium für Verkehr, Innovation und Technologie, BMVIT in Auftrag gegeben, und liefert einen Empfehlungskatalog für die kurz-, mittel- bis langfristige (2025) Ausrichtung des FFG Förderprogramms *IKT der Zukunft: Daten durchdringen - Intelligente Systeme*. Die Ergebnisse dieser Studie wurden mittels Methodenmix bestehend aus einer umfassenden Literaturrecherche, einer Online-Umfrage, Brainstormings im Zuge dreier Workshops sowie strukturierten Experteninterviews erhoben. Diese Technologie-Roadmap beleuchtet und vereint die unterschiedlichen Perspektiven der Akteure aus Forschung, Industrie, und der öffentlichen Hand. Einerseits zeigt die Studie den Bedarf an neuer IKT in diesem Bereich auf, und andererseits bereitet die Studie auf die zu erwartenden Entwicklungen, Anforderungen und Richtlinien vor.

Nach eingehender Studie des Gebiets, wurden eine Vielzahl relevanter Methoden und Verfahren zur intelligenten Datenanalyse identifiziert und in die folgenden vier, interagierenden Gruppen eingeteilt: *Suche und Analyse*, *Semantische Verarbeitung*, *Kognitive Systeme und Vorhersagen* sowie *Visualisierung und Interaktion*.

Die Breite der Anwendungsgebiete, die von österreichischen Unternehmen, Forschungsinstituten und Universitäten in diesem Bereich adressiert wird, ist beachtlich. Auf Basis einer detaillierten Analyse wie und in welchem Umfang Akteure in diesen Anwendungsgebieten Technologien nutzen bzw. zu nutzen planen, wurden der Gesundheitsbereich, Energie, e-Science sowie Produktion und Logistik als die wichtigsten Anwendungsgebiete on Österreich identifiziert.

Durch die Konsolidierung der Blickwinkel auf das Gebiet "Daten" und der umfassenden Erhebung der Standpunkte der unterschiedlichen Akteure, konnten die wichtigsten Herausforderungen in der intelligenten Datenanalyse ermittelt werden. Die Themen reichen in diesem Zusammenhang von Datenschutz, Datensicherheit und Datenbesitz über algorithmische und technologische Herausforderungen bis hin zur Mangelware "qualifiziertes Personal".

Im Zuge dieser Studie wurden die Kompetenzen der österreichischen Akteure auf dem Gebiet der intelligenten Datenanalyse erhoben, und in Form einer Kompetenzlandschaft abgebildet. Diese Kompetenzlandschaft deckt österreichische Forschungseinrichtungen, Universitäten, Fachhochschulen und in Österreich operierende Dienstleister ab. Österreichische Stärken liegen dabei im Bereich Statistik, effiziente Algorithmen, Machine Learning, Computer Vision und Semantic Web.

Als ein Ergebnis unserer Analysen geben wir neun Empfehlungen, die drei Hauptaspekte – Technologie, Koordination und Personal — adressieren. Vier dieser neun Empfehlungen betreffen technologische Herausforderungen und damit die Erforschung und Entwicklung von i) Verfahren die die Integration und Fusionierung von Daten vorantreiben, ii) innovativen

Ansätzen, die die Effizienz von eingesetzten Algorithmen erhöhen, iii) Technologien die Rohdaten in verwertbare Informationen umwandeln und zum Erkenntnisgewinn beitragen, und iv) Systemen, die zu einer weitgehenden Automatisierung von "Wissensarbeit" beitragen. Um mittel- bis langfristig herausragende Technologien "Made in Austria" im Bereich der intelligenten Datenanalyse zu produzieren, bedarf es zielgerichteter Förderung im Bereich dieser technologischen Herausforderungen.

Drei der neuen Empfehlungen fokussieren auf Maßnahmen welche die Innovationskraft und Wettbewerbsposition österreichischer Unternehmen stärken. Darüber hinaus zielen diese Empfehlungen darauf ab, die Sichtbarkeit und Attraktivität Österreichs sowie die Vernetzung der Akteure in einem internationalen Kontext zu verbessern. Um diese Koordinationsmaßnahmen zu voranzutreiben sind Investitionen in ein österreichisches Ökosystem für daten-basierte Innovationen zu setzen. Dieses Ökosystem macht Dienste und Daten zugänglich und interoperabel, und hat das Potential großen wirtschaftlichen Mehrwert zu erzeugen. Ein weiteres Ziel stellt die Erarbeitung der rechtlichen Rahmenbedingungen für den Umgang mit Daten dar.

Die beiden letzen Empfehlungen adressieren das Thema Personal und zielen darauf ab, den dringenden Bedarf an hoch-qualifiziertem Personal im Bereich der Datentechnologien zu befriedigen. Dazu sind neue Bildungsprogramme gefragt die dabei unterstützen "universelle Denker" anstelle von Fachexperten auszubilden, um die Anforderungen, die sich rund um das Gebiet (Big) Datenanalyse ergeben, zu bewältigen. Zusätzlich wird empfohlen ein dezidiertes "Austrian Data Technology Institute" ins Leben zu rufen, dass einerseits Spitzenforschung und andererseits die Vernetzung und den Know-How-Austausch zwischen österreichischen und internationalen Akteuren im Bereich der Datentechnologien ermöglicht. Zusätzlich werden Maßnahmen zur Verbesserung des Bewusstseins für Gender und Diversity im Bereich der intelligenten Datenanalyse geliefert.

Zur Umsetzung dieser Empfehlungen werden mehrere, potenzielle Leuchtturmprojekte vorgestellt. Dazu zählt ein breitenwirksames Leuchtturmprojekt – das Data-Services Ökosystem – dass die gegenseitige Befruchtung von Technologien und Anwendungsgebieten ermöglicht. Zusätzlich werden konkrete, anwendungsspezifische Leuchtturmprojekte vorgestellt, um die Herausforderungen in spezifischen Anwendungsgebieten wie etwa in Produktion und Energie, dem Gesundheitswesen und Digital Humanities zu adressieren.

Innovative Technologien zur intelligenten Datenanalyse besitzen das Potential nachhaltig Mehrwert für die österreichische Wirtschaft und Gesellschaft zu generieren. Für eine erfolgreiche Innovationswirtschaft ist es wichtig ein Ökosystem zu schaffen, welches den Raum für daten-basierte Innovation bietet und Technologietransfer unterstützt. Es gilt jedoch noch viele Herausforderungen – sowohl aus technologischer als auch aus gesellschaftlicher Sicht – zu überwinden, bevor Österreich den vollen Nutzen aus diesen Entwicklungen ziehen kann.

Contents

Chapter 1

Introduction

As the citizens of this digital world we generate more than 200 exabytes of data each year. This is equivalent to 20 million Libraries of Congress [41]. According to Intel, in 2012 each Internet minute sees 100,000 tweets, 277,000 Facebook logins, 204 million email exchanges, and more than 2 million search queries fired [128]. Data artifacts are now primarily digital and the need to digitise as a separate process is increasingly a phenomenon of the past as many devices (e.g. cameras) produce digital information right out of the box, tagged with additional information such as geo-coordinates or social connections. The increasingly wide use of sensors of all kinds leads to a flood of machine-generated information, while improvements in sensor technologies mean that this information is usually at a higher spatial or temporal resolution than was possible before. It is expected that as the *Internet of Things* gains traction, previously data-silent devices and objects will also begin contributing data. Looking at the scale at which data is being created, it is beyond the scope of a human's capability to process this data and hence there is a clear need for automated information processing and analysis [41]. More than this, automated processing and analysis can in general extract more value from data than is possible with manual analysis. As automated methods are becoming widely employed, there is even a risk associated with an organisation choosing not to employ such methods.

There is no *dearth* of data for today's enterprises. On the contrary, they are mired in data and quite deeply at that. Energy providers now receive energy consumption values from Smart Meters every 15 minutes instead of once or twice a year. Manufacturing machines generate ever more detailed logs of their actions. Hotels are rated online on multiple websites in multiple languages. Increasing amounts of medical information are stored electronically, while multi-dimensional medical imaging is becoming more common. Surveillance cameras are proliferating. Satellite images are increasing in resolution. Increasing amounts of data are being made available as Open Data. Today, therefore, the focus is on discovery, integration, consolidation, exploitation and analysis of this overwhelming information [41]. Paramount is the question of how all this (big) data should be analysed and put to work. Collecting data is not an end but a means for doing something that hopefully proves to be beneficial for the data owner, the business and potentially the society at large. However, this task should be approached carefully. The words of John Tukey are still applicable: *The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data* [132]. Even worse, it is possible to extract false or misleading correlations or relations from the data.

The increasing use of technologies to collect data and the ever improving capabilities to process this data efficiently have transformed our society and keep on doing so. They trigger an entire array of new questions and interesting challenges about how society should handle this new opportunity and the technology attached. Views on data have been dramatically transformed in just a few years. On the one hand, people are comfortable with storing large quantities of personal data remotely and are willing to provide information about their behavior on a regular and large scale. On the other hand, there is rising concern about data ownership, privacy and the dangers of data being intercepted and potentially misused [30].

This document presents a technology roadmap for *Conquering Data: Intelligent Systems* (Daten durchdringen: Intelligente Systeme) for Austria, more specifically for the Austrian Research Promotion Agency (Österreichische Forschungsförderungsgesellschaft, FFG) funding programme *ICT of the Future* (IKT der Zukunft)[1]. It is the result of a study[2] commissioned by the FFG and the Austrian Federal Ministry for Transport, Innovation and Technology (Bundesministerium für Verkehr, Innovation und Technologie, BMVIT). The study commenced in mid-June 2013 and was completed in January 2014. Its insights, results and objectives are based on a comprehensive literature review and various interactions with stakeholders through an online survey, discussion in three interactive workshops and eight interviews with representatives from Austrian companies and the public sector. In summary, the study proposes a nine objectives spanning over the short, medium and long term in order to strengthen Austria's innovation capacities in the *data* domain. Furthermore, potential lighthouse projects are presented as a route to attaining some of these objectives.

The importance of this topic can also be seen internationally, with the consultation process to create a Public-Private Partnership in Big Data currently underway in Europe[3], work by the National Institute of Standards and Technology (NIST) Big Data Public Working Group[4], and the creation of the Smart Data Innovation Lab[5] in Germany.

1.1 Objectives of Austrian ICT of the Future Programme

The ICT of the Future programme funds ICT research in the areas of Systems of Systems, Trusted Systems, Intelligent Systems und Interoperable Systems. The objectives of this programme[6] are the following:

1. **Develop lead technologies:** Increase both the quantity and the quality of ICT-research and development that can achieve and sustain technological leadership; Enable the exploration of new ICT research topics and application fields;

2. **Achieve lead positions in competitive markets:** Strengthen the capability of firms to innovate, support firms in establishing and extending their competitive position;

3. **Establish and extend a lead position as a location for research:** Secure and improve Austria's visibility, interlinkedness and attractivity in the international context in the area of ICT research and development;

[1] https://www.ffg.at/iktderzukunft
[2] http://www.conqueringdata.at
[3] http://europa.eu/rapid/press-release_SPEECH-13-893_en.htm
[4] http://bigdatawg.nist.gov
[5] http://www.sdil.de/de/
[6] http://www.bmvit.gv.at/innovation/ikt/ikt.html

4. **Produce highly qualified personnel:** Train and attract lead researchers; improve the availability of a sufficient number of trained researchers as the backbone of excellent ICT-research and development.

According to the call text, the objectives of this *Conquering Data: Intelligent Systems* technology roadmap are:

- Bring together the perspectives of technology and industry stakeholders in this area;
- Identify the requirements for new ICT in this area;
- Prepare for the expected developments, requirements and guidelines in the ICT field.

The questions to be answered by the study are:

- Concretise understanding of the ICT of the future topic "Conquering Data";
- Identify the research priorities for this topic in order to solve societal concerns and industrial challenges in Austria under consideration of the objectives of the ICT of the Future programme;
- Network the Stakeholders;
- Identify future need for action and potential in the area "Conquering Data" also with regard to possible application areas and under consideration of the interfaces to the other areas of the ICT of the Future programme;
- Propose concrete and realisable recommendations (e.g. recommendation for measures such as lighthouse projects);
- Identify new necessary business models and consider the development of human resources in the form of e.g. education of young researchers with strong interdisciplinary knowledge;
- Give recommendations for the short, medium and long term (2015, 2020, 2025) and visualise the results in a roadmap.

1.2 Strengths and Weaknesses in Austria

One of the questions that was asked to stakeholders during the workshops and interviews was to identify strengths and weaknesses of Austria in the area of Conquering Data. In this section, we give a summary of the responses, to serve as a motivation for discussions and solutions in the remainder of the document. The elicitation and understanding of Austria's strengths and weaknesses in the area of Conquering Data is important for identifying both the opportunities and limitations to face.

According to the stakeholders, the networking and thus the knowledge exchange between Austrian stakeholders working in areas related to Conquering Data is poorly developed. On the one hand a competence landscape (or index) of Austrian stakeholders in research and industry was requested and on the other hand the lack of transparency was criticised. Companies lack information about legal matters and competences of Austrian research organisations and universities while, on the other hand, academia is missing information about available data,

service providers, hardware, etc. It is also frequently the case that potentially useful (research) results are not effectively disseminated to potential adopters of the results.

Besides the lack of transparency in terms of resources, it was also criticised that both industry and research tend to be reluctant about disclosing data. As a consequence of this missing openness, it is difficult to obtain real-world data for research. This, however, does not only apply to industry and research but also to government data. Despite some established initiatives, as for example the Open Data program run by the City of Vienna, experts see a need for more Open Data offerings in Austria and a coordinated approach adhering to a nationwide Open Data master plan. On the contrary, Open Data providers criticise the lack of take up of publicly available data by research and innovative SMEs.

The absence of clear legal guidelines and frameworks is regarded as a major obstacle in any dealing with data. This, however, is not just an unsolved subject in Austria — we are facing a global issue that on the one hand slows down the take-up of new technologies and business models and on the other hand negatively impacts the trust in this kind of business and analytical technologies.

Above all, stakeholders from different research disciplines as well as industry agree that a culture and mentality of experimentation prevails in Austria. On the up-side the variety of funding opportunities in Austria is a highlight. Even though numerous funding opportunities exist the lack of continuity in terms of providing funding from basic through applied research to start-up support was criticised. Thus, the final step of transforming research and experiments into actual innovation is rarely taken.

One strength, according to the discussions during the workshops, may emerge from the specific size of Austria. It was stressed that Austria has the perfect size (in terms of population, markets, industries, etc.) to become the optimal "testbed" for innovation and technology testing at large. This is further supported by the geographical proximity of Austria's major cities.

In terms of research and technological competences the stakeholders see weaknesses in the areas of Recommendation Systems, Computational Linguistics as well as Inference. On the other hand, Austrian expertise is believed to have a wide range of well established capabilities in the fields of Search and Analysis, Reasoning, Semantic Web and Processing, Data Mining and Data Warehousing, Event Detection, Data Preservation, Security, Parallelisation and High-Performance Computing, Visual Computing, Computer Vision, Simulation and Visualisation, Medical (Bio) Informatics, and Music Information Retrieval. When taking a look at the application domains, Austria's strengths are seen to lie in the healthcare domain, the tourism and event/congress sector, eGovernment as well as the field of next-generation production.

Finally, the marketing of Austrian competences in the field is considered to be unsatisfactory and the awareness of the general public about the benefits of data-based innovation is minimal and mostly negatively biased by the fear of data misuse and general distrust in technology.

1.3 Overview of the Document

We begin in Chapter 2 with a description of the methodology used in the study. This is followed, in Chapter 3, by a definition of Intelligent Data Analytics, the domain of research and development working toward conquering data, and a list of the techniques that fall into this domain. In Chapter 4, we summarise some application areas in which Intelligent Data Analytics has made an impact or has the potential to make an impact. The open issues

and challenges for Intelligent Data Analytics, particularly those in which additional research and development are required, are listed and discussed in Chapter 5. After an overview of the Intelligent Data Analytics competence landscape in Austria (Chapter 6), we present, in Chapter 7, the proposed Austrian Intelligent Data Analytics Roadmap. Lighthouse projects to provide an impulse to achieving some of the objectives are described in Chapter 8. The roadmap objectives and proposed lighthouse projects are summarised in Chapter 9.

Chapter 2

Methodology

The methodology used in this roadmap study follows a triangulation approach and combines qualitative and quantitative methods to gain an in-depth look at expert's opinions and attitudes about data and the future data challenges. Mixed-method designs are believed to give qualitative data collection a strong analytical foundation. These approaches have gained wide acceptance in the last decades (cf. [127, 66]). When investigating the same topic from different viewpoints, triangulation seeks convergence of empirical results from various stakeholders, their ideas and opinions [70, 69].

More specifically, expert interviews with key people from Austrian companies were combined with an online survey, addressing stakeholders in industry and research. Additionally, workshops with senior participants from research and academia were organised. The research design was supported by an extensive literature research and the compilation of a position paper serving as the *point of departure and reference* during the study and its various empirical investigations. Furthermore, an international advisory board was set up. The board members consulted with the study authors during all phases of investigation. An overview about the relation of the different phases is given in Figure 2.1.

2.1 Advisory Board

During the start-up phase of the study, an international advisory board was set up. The objective was to involve experts with different backgrounds in order to obtain an interdisciplinary mix of people advising and guiding the study during all phases. Furthermore, we aimed at including members of diverse age, gender and national backgrounds in the advisory board. Finally, the advisory board consisted of 12 people — five female and seven male members from research and industry. The members were either working at renowned national or international universities or research institutes, or at small to large players from industry operating in or with data. Their backgrounds and expertise range from Big Data, Digital Networks, Visualisation, over Intelligent Information and Semantic Processing to Security and Law. The members of the Advisory Board are listed in Table 2.1.

2.2 Online Survey

Surveys are a preferred approach for collecting data from a large number of participants. The objective of this survey was, on the one hand, to sharpen and concretise the understanding of

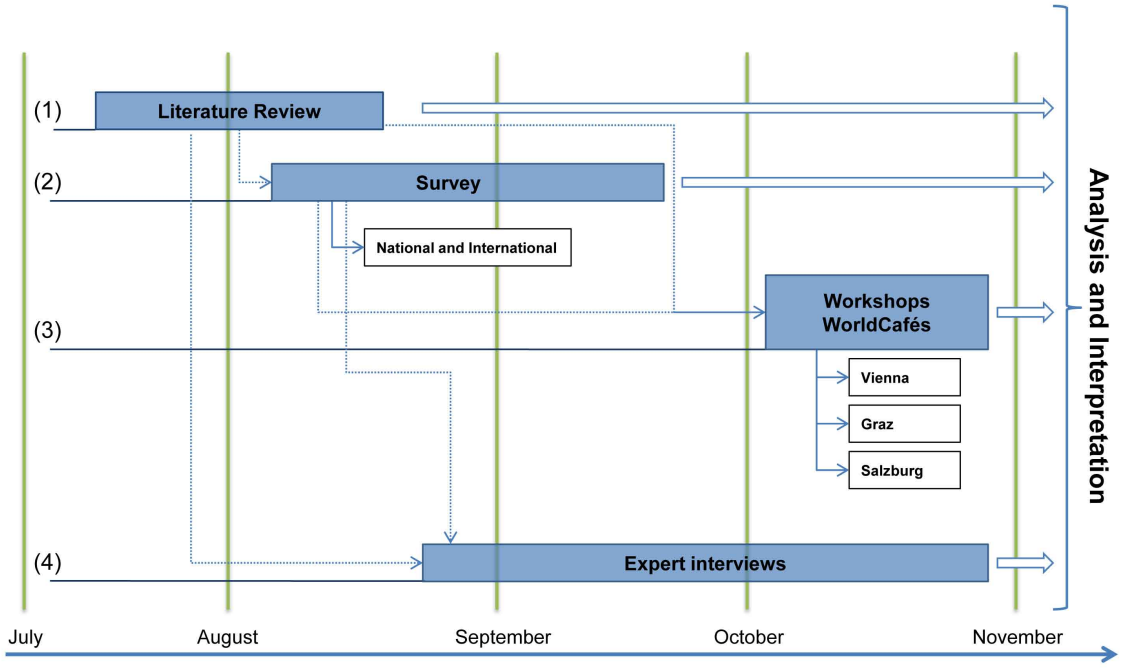


Figure 2.1: *Overview of the study methods.*

the "conquering data" area. On the other hand, the aim was to identify the future challenges in this area. For a detailed overview of the questionnaire see Appendix A.

Following a strategy of active sampling, the identification of Austrian stakeholders in this field formed the the starting point for further research steps: First, Austrian industry and research institutions were systematically analysed according to their research fields. These research fields served as a basis for the compilation of a competence landscape in *data research*

Wolfgang Nimführ	Information Agenda Consultant	IBM Austria	AT	Big Data Industry Solutions, Big Data Leader AT and CH
Stefanie Lindstaedt	Managing and Scientific Director	Know-Center GmbH	AT	Managing Director of Digital Networked Data
Silvia Miksch	Associate Professor	Vienna University of Technology	AT	Data visualisation expert
Seth Grimes	Industry Analyst	AltaPlana	US	Content analytics expert
Manfred Mitterholzer	Department Head	APA IT GmbH	AT	Expert Information Search
Janet Smart	Senior Research Fellow	University of Oxford	UK	Big Data expert
Edgar Weippl	Research Director	SBA Research	AT	IT security expert
Doris Ipsmiller	CEO	M2N GmbH	AT	Expert in Intelligent Information Processing
Dietmar Jahnel	Associate Professor	University of Salzburg	AT	Expert for Constitutional and Administrative Law
Anna Fensel	Head of Research Unit	STI Innsbruck	AT	Expert in Semantic Processing
Andreas Rauber	Associate Professor	Vienna University of Technology	AT	Information management and preservation expert
Alexander Löser	Associate Professor	Beuth University of Applied Sciences Berlin	DE	Big Data expert

Table 2.1: *Members of the Advisory Board.*

(cf. Chapter 6). Second, contact persons from these institutions were identified and asked about their opinions, attitudes, feedback and participation during the roadmapping process.

The final list of experts comprised 271 contacts that were contacted twice for filling in the questionnaire. The survey was online from the beginning of September 2013 until mid-October 2013. 105 people opened the survey link resulting in a response rate of 39%. However, several of them turned down the questionnaire or decided not to complete it after one or two questions. In this particular case, the general advantages of online surveys, such as truthfulness, better statistical variation or the improved data analysis processes (e.g. [24, 44]), unfortunately collide with the problems of controllability, time-capacities of participants and the fact that, while someone opens a questionnaire, he or she might be doing other tasks at the same time that increase drop-out rates (e.g. [27]). While a response rate of around 40% reflects an average rate of participation with online surveys [24], the drop-out rate in our case may be attributed to the rather complex matter. In the end, a total of 56 people completed the questionnaire of which four were female and 52 were male. They were mostly Austrians (96%) and the majority of them were working as researchers or academics. About a fifth (21.4%) of all responses came from industry, the larger part worked for academic (55.4%) or non-industry research organisations (33.9%)[1].

2.3 Workshops – World Cafés

Bringing together experts with diverse backgrounds through World Cafés is believed to allow the best compromise between different opinions and to foster development of joint ideas, motivated from different viewpoints. Originating from management and organisational change research, World Cafés are designed as open spaces to discuss and reflect on new ideas [18, 57]. The use of a café-style social context allows the sharing of information in an equitable and non-threatening manner[2].

For this study we created such inter-subjective spaces by equipping the meeting rooms with tables allowing small groups of people to sit down and to have a coffee. At every table the discussion topics were disclosed in terms of a *menu*. The table moderator invited participants to share their thoughts and ideas on the respective topic (see Appendix C for a list of the discussion topics). After 20 minutes, the participants had to change tables and form new groups of discussants. Each table moderator informed the new group of people about the previous discussions and invited them to connect to these previous thoughts.

All participants were motivated to put down keywords or important thoughts directly on the tables, which were covered with paper (for some workshop impressions see the pictures in Figure 2.2). As a result of the small groups and the informal atmosphere, nearly all participants started talking and expressing their opinions [18, 57].

Our World Café workshops took place in three different Austrian cities, namely Salzburg, Graz and Vienna, giving stakeholders from different regions, universities, research centres and companies the possibility to participate. Each workshop lasted a full day and featured a keynote, a presentation of the study objectives and preliminary results, a joint lunch as well as three World Café sessions each of which lasting approx. one hour. We invited senior researchers from the list of stakeholders and contacted them via email, followed by a telephone

[1] Multiple assignments possible.

[2] An overview of the method and examples for the application of World cafés can be found at http://www.theworldcafe.com or http://www.all-in-one-spirit.de/res/res_wc.htm

Figure 2.2: *Some workshop impressions.*

reminder and/or a personal email with the invitation to participate in the workshops. We purposefully invited female stakeholders to participate in the World Cafés. The meetings were scheduled in the morning and finished in the afternoon to not collide with family and care obligations of parents.

Especially in this phase of the study, marginalisation issues of women were discussed and reflected upon. While former explanations about the low percentage of women in science and technology ranged from natural aptitudes, poor attitude towards science, missing role models, a specific pedagogy of science classes etc. (for an overview and the systematic criticism on these factors see [28]), it has been widely accepted that structural and not individual impediments hinder women from a career development equal to men [60, 47]. This includes a masculine understanding of the profession and its career models [26, 47], male networks that are disadvantageous to women in informal selection, promotion processes [32] and also daily, informal interaction [63, 53]. To make sure that we take gender issues into account throughout all phases of the roadmapping process, two expert meetings with Dr. Brigitte Ratzer, head of the Center for Promotion of Women and Gender Studies at Vienna University of Technology were arranged. The meeting's objectives were to a) reflect on method and research design and b) to discuss our findings in relation to gender aspects.

In total, 61 experts participated in the workshops, 11 of them were women. 52% of the participants came from universities, 37% work for a non-university research institutes and 11% are engaged with industry research. In Salzburg, a member of the advisory board attended the workshop, in Vienna four out of 12 advisory board members were present.

2.4 Expert interviews

Experts are people of the target group who are consulted because of their expert knowledge and/or expert role. To learn more about the Austrian industry challenges in data, CEOs and C-level managers, including managers of research and data departments, were identified as the ideal candidates for our analysis. This group of people has gained practical and empirical experiences highly relevant to our research [29]. Furthermore, anonymity, availability as well as the preparedness to talk to researchers played a role in selecting interview partners [65].

Focus was put on a heterogeneous sample [108] in order to identify commonalities among diverse cases. Following the principles of theoretical sampling [64] we further extended the sample as long as new findings turned out. In total eight expert interviews were conducted. Five out of these eight were carried out with Austrian companies. The remaining three were

public/governmental organisations. The companies operated in life sciences, information technology, market research & statistics, healthcare and finance[3] One of the interviewed experts was female heading a department of about 50 staff members.

The expert interviews followed an explorative and open approach: the interviewer takes an expert position such that the discussions took place at eye-level. We used a semi-structured interview guideline [95] where questions are pre-defined (for the interview guideline see Appendix B). Nevertheless, structure and the organisation of the interview guideline was up to the interviewer during every expert talk. Interviewees contributed through their professional knowledge and position in the respective organisation in a way that key topics for business and decision makers could be identified.

Following the principles of global analysis, a multi-stage analysis [59] to identify the most relevant topics and challenges was applied. The interviews lasted between one and 2.5 hours. If agreed by the interviewee the interview was recorded. Each interview was documented and summarised via comprehensive meeting minutes. Every single interview was interpreted according to its own business needs, the main challenges which institutions face as well as their own position in the data analytics domain. In a second step, cross-case comparisons, relating the single cases and challenges to each other, were performed [96].

2.5 Analysis

In the end, findings of all empirical steps were combined and cross-referenced. The process, however, followed a qualitative research structure using results from one phase in the following research steps, thus analysis and interpretation is always intertwined with sampling and direct feedback of stakeholders [64].

[3] More information about interview partners and respective organisations cannot be displayed because of anonymity concerns expressed by the experts.

Chapter 3

Conquering Data with Intelligent Data Analytics

The art of Conquering Data with Intelligent Systems includes those areas of research and development in *Intelligent Data Analytics*, the area including Data Analytics and Intelligent Systems, that focus on computational, mathematical, statistical, cognitive, and algorithmic techniques for modeling high dimensional data with the ultimate goal of extracting meaning and actionable information from (raw) data of any format including text, audio, video or machine-generated data [38]. This requires methods such as learning, inference, prediction, knowledge modelling and representation, knowledge discovery and visualisation that are applicable to both small and large volumes of mostly dynamic data sets collected and integrated from multiple sources, across multiple modalities. These methods and techniques trigger the need for assessment and evaluation: automated and by humans. Intelligent Data Analytics enables (semi-)automated hypothesis generation, event correlation, and anomaly detection and helps in explaining phenomena and inferring results that would otherwise remain hidden [81]. Intelligent Data Analytics is a cornerstone in modern Big Data Analytics.

We now focus on the techniques common to Intelligent Data Analytics. We have divided these techniques into four (interacting) groups: **Search and Analysis**, **Semantic Processing**, **Cognitive Systems and Prediction**, and **Visualisation and Interaction**. The first three groups are the initially defined sub-themes of the theme *Conquering Data: Intelligent Systems* in the FFG funding program ICT of the Future. The fourth group was identified by the study authors as an important component of Conquering Data. Two over-arching topics are applicable to all four groups: **algorithmic efficiency** and **evaluation and benchmarking**. The former concerns developing algorithms to process larger amounts of data in a shorter time, while the latter deals with the quantitative evaluation of algorithm and system performance. Figure 3.1 presents a graphical overview of these groups along with the application domains from Chapter 4. The figure emphasizes that the same Intelligent Data Analytics techniques are applicable to multiple application domains. The goal of using these techniques is to generate increased value for the application domains through the analysis of the application domain data.

In each of the following sections, the techniques for each group are listed and briefly defined. Unless otherwise indicated, the definitions are taken from Wikipedia, representing a broad consensus on the concept meaning. The list is refined and grouped from the list presented in [91], but also based on an analysis of the research and development in data analysis currently

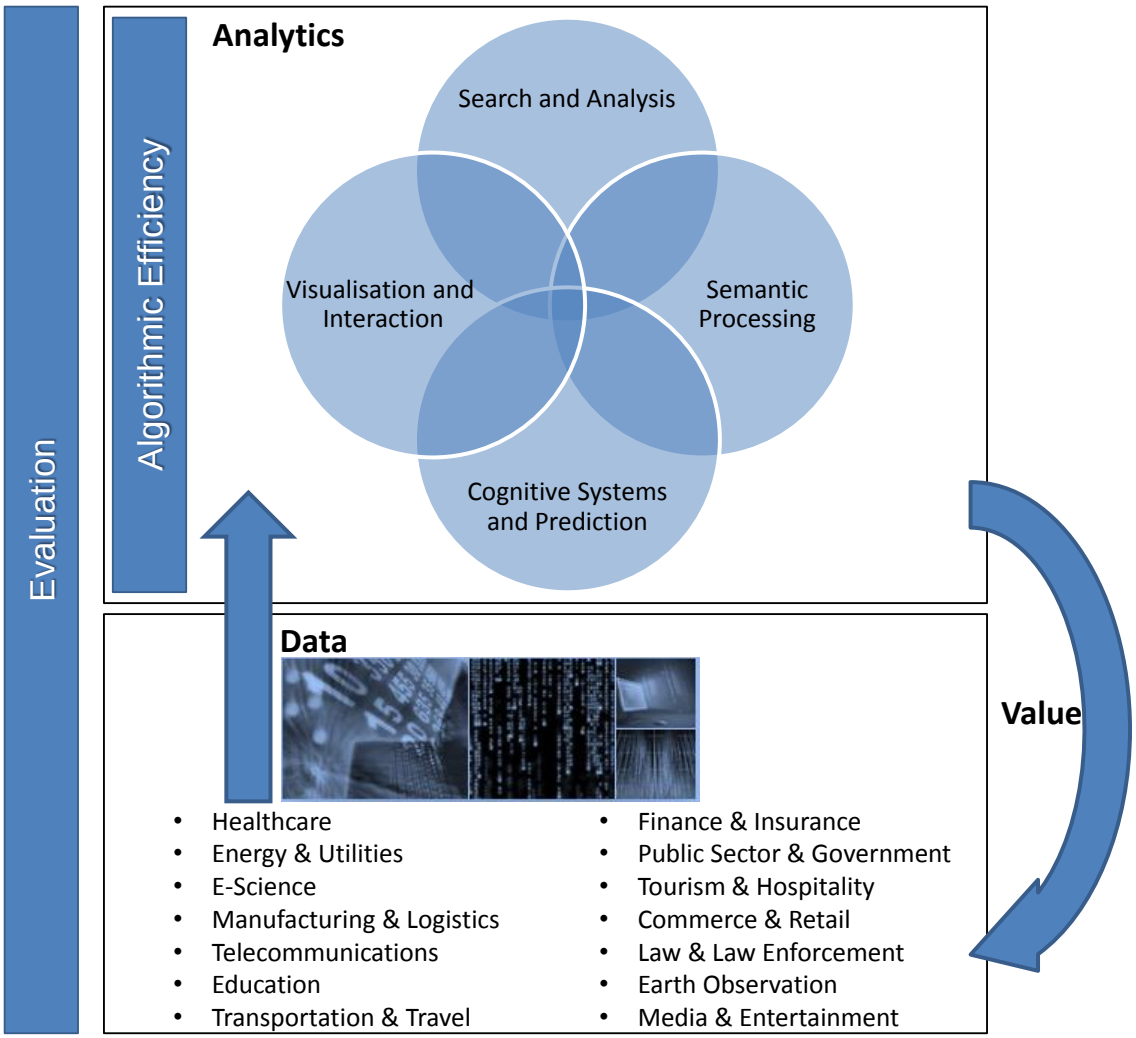


Figure 3.1: *Relationship of the groups of Intelligent Data Analytics techniques to application domains.*

undertaken in Austria. As is unavoidable with such a taxonomy, the various techniques have different scopes, with some, such as *statistics*, encompassing a wide field, while others, such as *natural language processing*, are much narrower in scope. There are, of course, overlaps in the scopes. Also, some techniques are used by other techniques, e.g. reasoning could be used by a decision support system. Figure 3.2 summarises the techniques within each group.

The techniques and their grouping form the basis for the analysis of the Austrian Intelligent Data Analytics landscape presented in Section 6.1. Hardware developments can contribute to speeding up processing and reducing energy consumption — this topic is not included in this analysis as it falls under a different funding stream.

3.1 Search and Analysis

Search and Analysis is the domain of searching and analyzing multimodal data (text, image, audio and voice, video, ...), and the aggregation and fusion of such multimodal data streams.

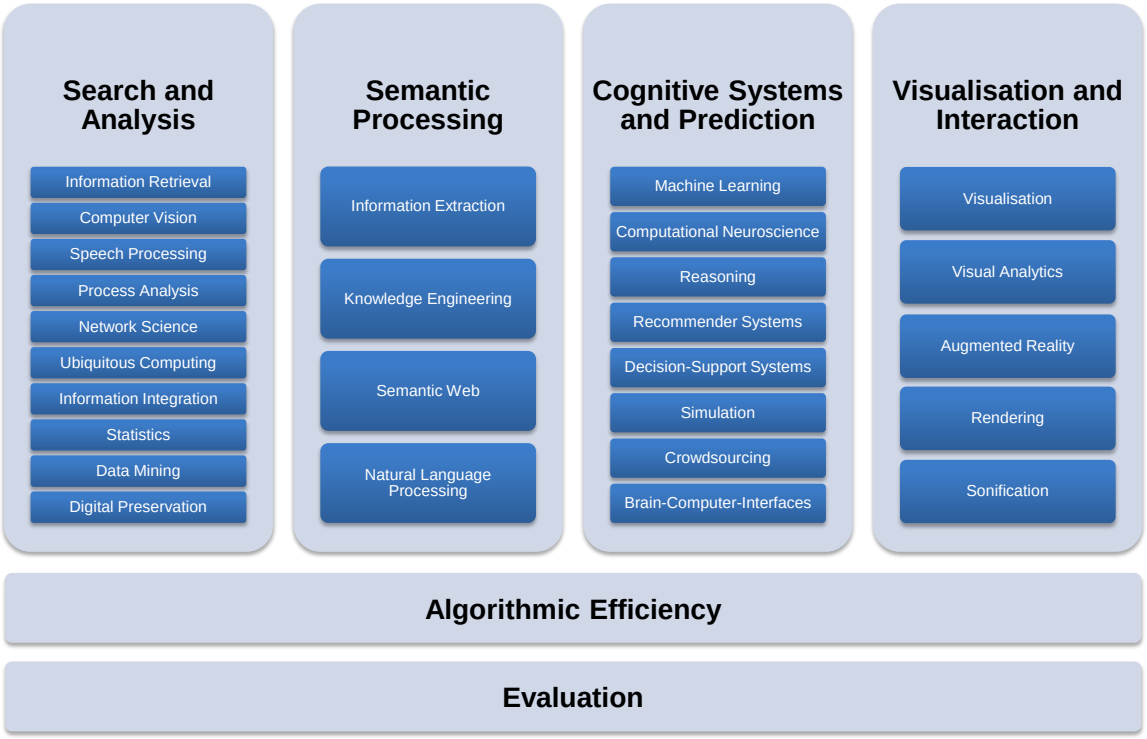


Figure 3.2: *Intelligent Data Analytics techniques classified by group.*

The results of the analysis range from preparing the data for further semantic processing or as input for cognitive systems, over discovering interesting patterns or relationships in the data, to making the data more easily accessible.

Search (or **Information Retrieval**) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing. The features used in search algorithms are generally specific to the modality being indexed, leading to **text search**, **image search**, **music search** and **video/multimedia search**. Web search is currently the most visible application of search. *Context-sensitive search* is becoming increasingly important, especially with the ubiquity of mobile search.

Computer vision is a field that includes methods for acquiring, processing and analyzing images and, in general, high-dimensional data from the real world in order to produce numerical or symbolic information. It includes a wide range of sub-fields: image analysis, stereo vision (creating depth from pairs of images), 3D vision (dealing with data acquired in 3D), document analysis, recognition of objects in images and videos, and tracking of objects in videos.

Speech processing includes the acquisition, manipulation, storage, transfer and output of digital speech signals. It is a special application of **audio signal processing**, which is part of the broad field of **digital signal processing**.

Process analysis involves analysing and comparing processes or workflows (e.g. business processes).

Network science is an interdisciplinary academic field which studies complex networks such as telecommunication networks, computer networks, biological networks, and social networks, leading to predictive models of these phenomena.

In **ubiquitous computing** (or **pervasive computing**), computing is made to appear in any device, in any location, and in any format. **Sensor networks**, spatially distributed autonomous sensors, are a specific case of ubiquitous computing.

Information integration is the merging of information from heterogeneous sources with differing conceptual and contextual representations. **Information fusion** involves the combination of information with the aim of reducing uncertainty.

Statistics is the study of the collection, organisation, analysis and interpretation of data, including the design of surveys and experiments. Statistical techniques typically allow an estimation of the significance of relations between variables.

Data mining is the computational process of discovering previously unknown or "interesting" patterns, interesting relationships (e.g. association rule mining) or groups (e.g. cluster analysis) in data sets.

Digital preservation is the series of managed activities (planning, resource allocation, and application of preservation methods and technologies) necessary to ensure continued access to digital materials for as long as necessary.

3.2 Semantic Processing

Semantic Processing adds structure and "meaning" to data, whilst making it accessible for machine learning methods and large scale automatic processing.

Information extraction is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents or other content.

Knowledge engineering is the process of integrating knowledge into computer systems in order to solve complex problems normally requiring a high level of human expertise [54]. **Knowledge representation** (KR) aims at representing knowledge in symbols to facilitate inferencing from those knowledge elements, creating new elements of knowledge. **Ontologies** are a very commonly used knowledge representation, and are constructed and maintained by **Ontology Engineering**. **Knowledge integration** is the process of synthesizing multiple knowledge representations into a common representation.

The Semantic Web aims at the creation of a "web of data" by encouraging the inclusion of semantic content in web pages, allowing the information to be interpreted by machines so that they can perform more of the tedious work involved in finding, combining, and acting upon information on the web. **Linked Data** describes a method of publishing structured data so that it can be interlinked and become more useful.

Natural Language Processing (NLP) uses algorithms to analyze human natural language [91]. It also includes the areas of **Natural Language Understanding** (NLU), **sentiment analysis** and **Natural Language Generation** (NLG), where the latter is the generation of natural language from a machine representation system such as a knowledge base or a logical form. **Machine Translation** is an area covering both NLU and NLG. **Speech analytics** aims at analysing recorded speech to gather information, including information not directly encoded in the words such as the emotional character of the speaker.

3.3 Cognitive Systems and Prediction

Cognitive Systems transform data into knowledge structures and act on it in ways inspired by the human mind and intellect. *Prediction techniques* learn or model a relationship between input data and output variables on existing data, and then predict the value of the output variables when given a new set of input data. Prediction techniques have been placed into this group, as they are often inspired by cognitive systems, but are also used in analysis, semantic processing and visualisation.

In the discussions with the workshop participants it has been underlined that, depending on their background, the notion 'cognitive system' has different meanings to different persons. Psychologists usually refer to a cognitive system as a mental system of an individual — a system of interrelated beliefs, ideas, knowledge, and how an individual understands and reacts to the surrounding world. The computer scientist's view on cognitive systems, which is also the view on which this section focuses, is one in terms of artificial intelligence systems that try to incorporate the way humans think and react to external events according to their own, previously gathered knowledge and experiences.

Machine learning concerns the construction and study of systems that can learn from data. Machine learning is related to **pattern recognition** and makes use of algorithms such as **neural networks**, **ensemble learning** and **genetic algorithms**.

Computational neuroscience is the study of brain function in terms of the information processing properties of the structures that make up the nervous system [37]. It is distinct from machine learning in so far that it emphasises descriptions of functional and biologically realistic neurons (and neural systems) and their physiology and dynamics.

Reasoning uses deductive logic and inference on machine-readable descriptions of content (e.g. in the Semantic Web) to allow computers to perform automated conclusion generation and gathering of information.

Recommender systems predict the rating or preference that a user would give to an item (e.g. a book), using a model built from the characteristics of an item (content-based approaches) and/or the user's social environment (collaborative filtering approaches) [118]. They can also be based on behavioural modelling.

A **decision-support system** is an interactive system intended to help decision makers compile useful information from a mix of raw data, documents, and personal knowledge, or business models to identify and solve problems and make decisions.

A **simulation** aims to model the behaviour of a particular (complex) system, and then "runs" the model to explore, forecast, predict and gain new insights.

Crowdsourcing creates a large cognitive system by using a group of online people to obtain data or complete a task .

A **brain-computer-interface** (BCI) is a direct communication pathway between the brain and an external device. The availability of inexpensive BCIs opens the possibility to obtain new information from people interacting with a computer, while leading to the generation of large amounts of data.

3.4 Visualisation and Interaction

Visualisation creates a human-accessible interface to Data Analytics. It supports exploring and understanding data more easily (especially when data volumes or the number of dimensions

grow) and helps discovering patterns, thus making data more accessible for specialists, various user groups, and the general public. Non-visual interaction paradigms are also included in this group.

Visualisation is any technique for creating images, diagrams, or animations to communicate a message. Areas of interest for Data Analytics include **scientific visualisation**, concerned with the presentation of interactive or animated digital images to scientists who interpret potentially huge quantities of laboratory or simulation data or the results from sensors out in the field; **information visualisation**, the use of interactive, sensory representations, typically visual, of abstract data to reinforce cognition; and **knowledge visualisation**, the use of visual representations to transfer knowledge, including insights, experiences, attitudes, values, expectations, perspectives, opinions, and predictions, between at least two persons.[1]

Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces [130].

Augmented Reality is a means to (visually) represent data in connection with its real-world context, making it visible in real-time. Technological developments (e.g. smartphones, Google Glass) are pushing its widespread application. The increasing diversity and capability of **touch and gestural interfaces** allow the end user more flexibility in interacting with visualisations.

Rendering is the process of generating an image from a model by means of computer programs. Of particular interest for Data Analytics is volume rendering, a set of techniques used to display a 2D projection of a 3D discretely sampled data set.

Sonification is the use of non-speech audio to convey information or perceptualise data [85]. Auditory perception has advantages in temporal, amplitude, and frequency resolution as an alternative or complement to visualisation techniques.

3.5 Algorithmic Efficiency

Algorithmic efficiency refers to developing efficient algorithms that are able to process larger amounts of data in a shorter time. The techniques and methods listed above can usually be implemented by a variety of algorithms. Algorithmic efficiency can be enhanced through the use of the following techniques:

A **computer cluster** consists of a set of loosely connected or tightly connected computers that work together so that in many respects they can be viewed as a single system. **Grid computing** is distinguished from cluster computing by grids tending to be more loosely coupled, heterogeneous, and geographically dispersed.

Parallel algorithms carry out several operations simultaneously on parallel processors or on multiple processors. MapReduce [43] is an example of a commonly used framework for processing parallelisable problems.

Optimisation is the process of modifying a software system to make some aspect of it work more efficiently or use fewer resources. Through an investment of effort by knowledgeable people, it is often possible to increase the performance of algorithms significantly.

For certain types of operations, **specialised processors**, such as Graphics Processing Units (GPUs), can be far more efficient then general-purpose CPUs.

[1] Further definitions of the latter three concepts are available here: http://www.infovis-wiki.net/index.php?title=Category:Glossary

Algorithms can be benchmarked in order to determine their performance quantitatively. This usually involves measuring their time and space efficiency, but can also involve measures such as power consumption or cost of execution (e.g. if the algorithm is to be run on a pay-per-use cloud infrastructure).

3.6 Evaluation

There are usually many ways to solve a given problem in the area of Intelligent Data Analytics, through various combinations of the techniques and methods from the "toolbox" described above. In order to estimate the suitability of a solution for a specific task, an evaluation should be performed, where the result of the evaluation should aid in the choice of the best solution. This evaluation can be done at a number of levels. For tasks for which sufficient examples of input data and expected outcomes are available, various solutions can be evaluated on a subset of the data, measuring how well the solutions can produce the required outcomes using one or more of a number of common metrics (such as sensitivity, specificity, precision, recall...). A task suitable for this type of evaluation is the detection of fraudulent transactions, for which sufficient examples of fraudulent and non-fraudulent transactions exist, and where the task is to return a prediction, which can be easily compared to the known status of example transactions.

In the academic community, evaluations form the basis of evaluation campaigns, also known as challenges, benchmarks or competitions. Such events are common, for example, in Information Retrieval [136] and Machine Learning [115]. In these events, the organisers make data and associated tasks available, and solutions are submitted, usually by academic researchers and, occasionally, by industrial research labs. The main aim of these events is an in-depth analysis of the behaviour of the submitted algorithms in order to advance scientific knowledge while winning plays a subordinate role.

One of the longest running evaluation campaigns is TREC, the Text REtrieval Conference, which is organised by the National Institute of Standard and Technology (NIST) in the USA, and has been running since 1992. A recent independent study of the economic impact of TREC [120] came to the conclusion that "for every $1 that NIST and its partners invested in TREC, at least $3.35 to $5.07 in benefits accrued to [Information Retrieval] researchers," as more efficient and effective experimentation could be done due to, amongst other benefits, the availability of shared datasets and evaluation methodologies. At present, work is being done to update the way in which TREC evaluation campaigns are carried out, allowing, for example, the use of larger datasets and real-time evaluation independent of a cycle of workshops.

For solutions that require interaction with the end user, a more user-centred evaluation needs to be performed. Different amounts of user input can be collected depending on the design of the experiment and the available resources. A minimal user input can be gained through an artificially constrained result presentation, for example requesting that a user selects the group of search results most useful to a specific query. Such evaluations can also be conducted by means of crowd-sourcing. More detailed feedback can be obtained from requesting users to carry out controlled tasks in the laboratory, although such experiments require a correspondingly higher investment of time on the part of both the researchers and end users. For such experiments it is necessary to ensure the *internal validity* through careful experimental design, meaning that the conclusions of the study are warranted, but also the *external validity* needs to be guaranteed, meaning that the conclusions can be generalised

beyond the users taking part in the experiment. Usually, a trade-off between internal and external validity is required. In order to have *ecological validity*, the experiment should be carried out in a way that approximates the real world. A common way of achieving this is through field observation by observing the end users using a solution in their standard work environment.

Chapter 4

Intelligent Data Analytics Applications

Austrian companies, research institutes and universities focus on a range of application areas. For this report, we reviewed many of these areas with respect to how they currently handle and use Intelligent Data Analytics. Healthcare, Energie and Utilities, E-Sciences as well as Manufacturing and Logistics were identified to be the most important application domains in Austria. However, to provide a comprehensive and holistic overview to the reader, we took an international viewpoint. Where possible, we complemented this view with an Austrian perspective on the application areas.

A recent McKinsey report [93] estimates that potential global economic value per year that might be expected from widespread use of open data combined with proprietary data in seven areas of the economy ranges from $3.2 trillion to $5.4 trillion, where the seven areas are education, transportation, consumer products, electricity, oil and gas, healthcare and consumer finance. This economic value is generated by enabling better decision making, providing insights for customised products and services, or exposing anomalies in performance data that lead to better processes.

In the online survey we have asked which of the application domains are important (multiple selections from the list of application domains were possible). The percentage of respondents that selected each of the domains is shown in Figure 4.1. Further application areas were also identified from free text responses to the survey and from the workshops, and are covered in this chapter. The order of application domains in this chapter is based on the survey results, from most important to least important, followed by the additional application areas identified. In the workshops, the question of important application domains in Austria was asked in a less structured way. In the responses, Austria's strengths were seen to lie in the healthcare domain, the tourism and event/congress sector, eGovernment as well as the field of next-generation production.

4.1 Healthcare

The healthcare sector consists of many stakeholders, including the pharmaceutical and medical products industries, healthcare providers, health insurers and patients. Each stakeholder generates pools of data, which have typically remained disconnected [91, p. 37]. For example, in Austria there are significant differences in how the various federal states deal with patient

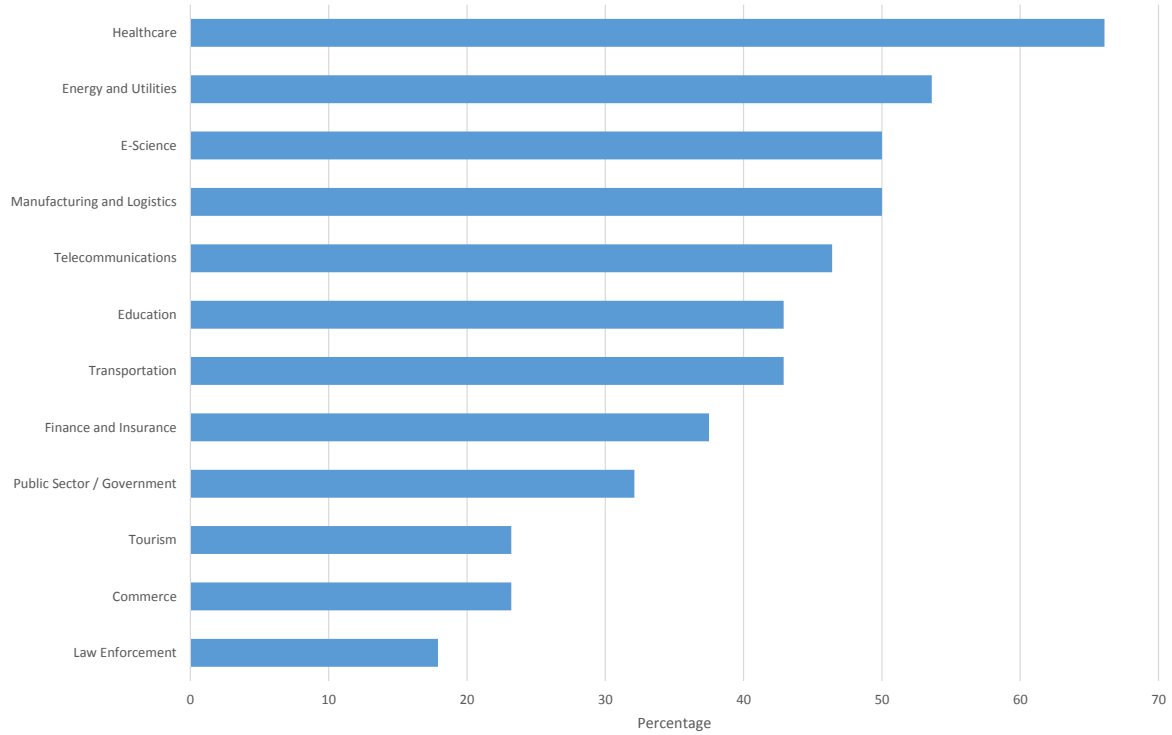


Figure 4.1: *Importance of the application domains based on the survey responses.*

data centralisation. In Styria, over 90% of the patient data is centralised, but in Vienna many disconnected data silos exist (health insurance agencies, hospital patient data, etc.) [112].

The amount of information to analyse in the health sector is growing rapidly. Medical imaging devices produce data ever more rapidly and at increasing resolution. It is estimated that medical images of all kinds will soon amount to 30% of all data storage [8]. The cost of sequencing a human genome has already dropped below US$1,000, and the increasing ease of extracting further -omics data (proteomics, metabolomics, epigenomics...) will also increase the amount of data to be processed.

There is a substantial opportunity to create value if these pools of data can be digitised, combined and used effectively [91, p. 37], especially through so-called secondary use of this information (uses beyond providing direct healthcare). The following are some of the ways in which the use Intelligent Data Analytics in healthcare can lead to significant savings in health care expenditure [91, p. 44]: *Comparative Effectiveness Research* predicts which treatments work best for which patients based on analysis of extensive patient and outcome data; *Clinical Decision Support Systems* can automatically mine the literature to suggest courses of action based on a patient's record; *Remote Patient Monitoring* for chronically ill patients reduces the frequency of hospital visits; *Predictive Modeling* can lead to more efficient and effective drug development by making predictions of optimal drug R&D paths based on aggregated research data; and *Personalised Medicine* takes a patient's genetic information into account to provide tailored treatment.

Difficulties to overcome, while implementing effective solutions in the healthcare sector include the fact that many medical records are still either handwritten, or in digital formats

that are not useful, such as scanned versions of handwritten records [105]. Furthermore, many issues with privacy and security related to the secondary use of health information need to be solved, although the opinion has been expressed that it is in fact "in some cases unethical, to store [population-based data] without installing mechanisms to allow access and publication in appropriate and useful ways" [56]. Currently, the pharmaceutical industry is under pressure to release all of the experimental data obtained during the clinical trials of medication that they have conducted. In Austria, privacy of patient data is of utmost importance and a culture of data privacy has to be developed, also by the use of reliable pseudonymisation processes [112].

Workshop attendees as well as the online survey respondents consider the healthcare domain to be one of the currently most important domains where data management and analytics play an important role. The biomedical domain is also seen as one of the strengths in the Austrian research scene.

4.2 Energy and Utilities

The energy and utility industry, applying latest smart metering and smart grid technologies, now has the capability to record information about consumption and production of energy in much higher resolution. Whereas meter recordings used to be done manually on a monthly (or even less frequently) basis, modern meters now record constantly on a 15-minute basis. 96 million readings per day, for every 100 million meters, results in a 3,000 fold increase in data volume and adds a big data component to the energy and utility business. As a result, it is now possible to better understand the usage pattern of customers, to find out how well they respond to price changes, to better segment them and, with this information, provide services to help customers understand their own usage pattern better and help them save energy. In addition, energy grids become more "intelligent" and adaptive to fast changes and (locally) increased and reduced power demands.

Ecova, an energy and sustainability management company, evaluated its Energy Data Warehouse and revealed interesting trends in its clients' energy consumption, which were published in a recent report [74]. In just four years, between 2008 to 2012, energy consumption across the US dropped by nearly 9%. They also found that water prices increased by 30%. This report is just one example of how Intelligent Data Analytics already starts informing us about the large-scale transformations in our society.

In Austria, the quality of the measured energy data is usually high, but data on the clients is rather low quality. The possibility to acquire further client data to validate and complement existing data in the energy and utilities companies is currently considered to be too expensive. Nevertheless, data analytics is used to predict client churns or to market new energy rates.Workshop participants considered the energy economics [122], especially the green and renewable energy use and generation, to play an important future role in Austria. The online survey respondents see energy and utilities to be a very important application domain in the short term.

4.3 eScience

Science is moving into a new, data-intensive computing era that has been called the *fourth paradigm for science* [77], brought about both by the increasing capacity to generate data and to process data. Terabyte-sized data sets are now common in earth and space sciences, physics,

bio-informatics and genomics [94]. Furthermore, the digitisation of extensive cultural heritage allows analyses at an unprecedented scale in the humanities and social sciences [35, 73], leading to the practice of *Digital Humanities* or *Digital History*.

The European Commission Report *Riding the Wave* [8] lists the requirements and challenges in creating e-Infrastructures that will allow the practice of eScience, and a European Commission Recommendation on e-Infrastructures was made in 2012[1]. E-Infrastructures should contain a generic set of tools that support the full range of data activities: capture, data validation, curation, analysis, and ultimately permanent archiving [77]. The e-Infrastructure should allow interoperability between data from different sources, facilitate access by researchers to the data, but also control access to sensitive data. Incentives should be created to encourage researchers to contribute data to the e-Infrastructure, while financial models are required to ensure the sustainability of e-Infrastructures. Finally, amateur scientists or citizen scientists have already made significant contributions to scientific data collection and data analysis (e.g. solving protein folding puzzles and scanning through astronomical images), and methods to involve them in eScience would be constructive. The European Union Framework 7 e-Infrastructure program supported the creation of e-Infrastructures and projects have been funded in research areas including physics, astronomy, earth sciences, biology, seismology and agriculture. A basic e-Infrastructure to allow archiving of scientific publications and data is currently being created in Austria[2].

Beyond the opportunities offered by e-Infrastructures, scientific communication is also changing. Although publishers have adopted technologies such as the web and pdf documents, it is largely true that "the current scholarly communication system is nothing but a scanned copy of the paper-based system" [134]. Due to the volume of scientific papers published, augmenting the papers with machine-readable metadata encoding key findings and the literature citations could allow efficient data mining to, for example, identify promising avenues of research. For detailed descriptions of experiments, the myExperiment platform allows sharing of computational work-flow descriptions. A proposal for an ICT-based system, incorporating these concepts in order to revolutionise scholarly communication and hence create an innovation accelerator is presented in [135].

39% of the online survey respondents see this application domain as an important one for the medium term, and another 28% consider it important for the short term.

4.4 Manufacturing and Logistics

Manufacturing traditionally generates large data sets. This phenomenon has developed even further in recent years to having collected almost 2 exabytes of data in 2010 alone [91] and is continuously growing at the speed of RFID tags sold[3]. Data is mostly used to ensure quality and efficiency in the production process.

Based on these large data sets, data analytics has also become a long-standing tradition in manufacturing, especially with the current move toward the fourth industrial revolution,

[1] See Recommendation 5 in the European Commission Recommendation of 17.7.2012 on access to and preservation of scientific information (http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf).

[2] http://www.bibliothekstagung2013.at/doc/abstracts/Vortrag_Budroni.pdf

[3] RFID tag sales are projected to rise from 12 million pieces to 209 billion between 2011 and 2021 [91].

Industry 4.0.[4] Harding, in [75], summarises research in analytics[5] that dates back to the mid-1980s and distinguishes the areas of Manufacturing Systems, Fault Detection, Engineering Design, Quality, Decision Support, Customer Relationship Management (CRM), Maintenance, Scheduling, Layout Design, Concurrent Engineering, Shop Floor Control, Resource Planning and Material Properties.

These areas remain relevant, although categories and names may have changed. More importantly, Data Analytics has to be reviewed in the new light of big data with more information available and more dynamic ways of combining and processing them. IT systems have to manage data that is increasingly more complex and interactive as manufacturers start to better associate and integrate data from different sources, systems and formats from areas such as computer-aided design and collaborative product development management. Data is also increasingly shared and integrated across organisational boundaries [91]. The McKinsey report forecasts the biggest applications of (Big) Data Analytics to be in R&D, supply chain, and production functions and therefore closes the cycle with Harding's summary with a modern touch. What is new is the more collaborative use and the social aspect of data that can enhance the entire product life cycle by better integrating the customer in the design process and by combining into a better whole.

At the workshops organized as part of the study, the manufacturing and the logistics sectors were regarded by the participants as areas where intelligent handling of data will play a major role in the Austrian economic future. There is, however, the issue of data 'lying around', that is data collected in the manufacturing and logistic processes, for example as part of the Industrial Internet, for which no use is envisioned by the companies that generate it. It is hoped that Intelligent Data Analytics will be able to help companies extract and visualize knowledge from their data. The (Austrian) industry's missing openness and transparency regarding the handling of the data they have collected or to which access is granted is seen by the workshop participants as a weakness of this domain. The respondents of the online survey consider manufacturing and logistics to become a relevant domain in the short (46% of the participants) to medium term (33%).

4.5 Telecommunications

Telecommunication providers traditionally work with large data sets. With five billion mobile phones in active use in 2010 [91], one can only imagine how much data is stored and available on a daily basis. Every time a customer makes a call, texts somebody or uses the Internet, there is an activity log. Even when the phone is just left alone, information about approximate positioning to nearby cell antennas is collected, not to mention the much more precise GPS information. Information was rich even before Big Data was a term and a topic. Hence it comes as no surprise that telecommunication providers have largely recognised the need for Data Analytics at the centre of their business model and participate more strongly in the movement than other businesses. Unlike other types of businesses, they focus on the real-time aspects of data, and to a lesser extent on the large volume of the data, probably because large data volumes were part of their business model long before Big Data existed. In [91], it is found that communication businesses focus on the customer as the main objective for their

[4] http://en.wikipedia.org/wiki/Industry_4.0

[5] In the paper Data Analytics is referred to as data mining.

big data analytics efforts, but that they are at a pilot stage and use their internal, pre-existing data sources.

The respondents of the survey consider the telecommunications application domain to be most relevant in the short term (42% of the participants) and in the medium term (32%).

4.6 Education

The education sector is another potential beneficiary of Intelligent Data Analytics in the near future. Educational Data Mining [21] uses data analysis focused on developing methods to explore data from educational settings to better understand students and their learning environments. Methods are applied from data mining and machine learning, statistics, information visualisation, and computational modeling. The theme is still largely a research topic that aims to model students — especially individual differences, domain knowledge, collaboration behavior and pedagogical support (as for example administered by learning software). There is initial evidence of integrating these methods with existing educational systems such as Moodle [119] with potential for a dramatic change in the educational landscape of the future. Online survey respondees considered education to be most important on the medium term (43%)

4.7 Transportation and Travel

The travel and transportation industry faces challenges and opportunities as the economy moves into an information age [39]. Travel and transportation companies are facing many of the same challenges and opportunities as other business segments in terms of managing risk, enhancing the customer experience, and ensuring operational excellence. The need to balance cost, product/service quality/safety, and customer service is particularly important for travel and transportation companies because these businesses are undergoing a fundamental shift. For these "service businesses" the importance of quality and customer satisfaction has never been questioned, but they are seeing a change from "product-related services" to "information-related services".

As the industry evolves and becomes more complex, the amount of data to be handled, in particular real-time and near real-time data, is growing. New technical, organisational, process, and decision management frameworks will be required to cope with the volume of data generated across the industry [49]. This includes applications such as: price optimisation for the transportation and travel industry commodities (e.g. airplane seats and price [50]); offer personalisation based on customer purchasing history and preferences; efficient booking and travel management for corporations and organisations; human resource optimisation (e.g. matching call centre operators to customers based on personality attributes); financial performance management and the evaluation of capital investments, including carefully monitoring employee and customer satisfaction and promoting customer loyalty. In Austria, predictive data analysis methods are increasingly used for transportation incident and malfunction management. The employment of these predictive methods is in an initial stage and it is expected that it will take about five more years until they will give reliable results.Data Analytics is also used on sensor data reads to control the traffic on Austrian highways, to control the lorry traffic and the traffic in cities.

Data Analytics of primarily large volume, unstructured data plays a vital role in delivering a more efficient and tailored travel experience with benefits to both travel companies and travelers alike. These benefits range from better decision support, new products and services over better customer relationships, to cheaper and faster data processing.

Parts of the travel and transportation industry have been using information technology for decades; a consequence of this is that key data is often fragmented across multiple functions and units. Integrating this information is difficult and often fraught with privacy issues. Furthermore, the real-time IT architectures used by many travel industry companies cannot run on Hadoop or other open-source environments; they are called TPF for Transaction Processing Facility and were developed by IBM in the 1960s and 70s, and have been refined ever since [42].

The participants to the online survey considered the transportation sector to be equally important on both medium (36%) and short terms (36%).

4.8 Finance and Insurance

Financial institutions have large volumes of custom data at their disposal, most of it from storing the details of every transaction performed within the bank's business or information systems [133]. Many traditional systems store this data for years, for example on tapes, without being able to analyse it in some automatic way. The transition to current distributed file systems like HDFS[6] is, however, a factor that permits cost effective analysis of past data, be it for fraud detection or customer-tailored product offers. Financial institutions have already recognised the competitive advantage they can gain when making use of the information they store about their customers. A recent IBM study [133] identified a 97% increase in the number of financial companies that have gained competitive advantages using Data Analytics in the last two years.

Fraud detection is a flagship of Big Data Analytics in the finance and insurance industry. Adding more layers of security to the financial transactions, like additional verification requests, or temporary account blocks is one avenue to take towards better financial crime prevention. These measures, though, may alienate customers, even when they do accept various amounts of surveillance of their financial activities in exchange for some degree of peace of mind. This emphasizes the importance of employing complex algorithms to detect frauds as fraudsters are more data and technical savvy and, at the same time, access to financial products is diversified to new channels (via smart phones, computers, branches).

At the global level, PayPal was one of the first to use complex fraud detection algorithms [129]. In Austria, paysafecard,[7] for example, successfully introduced in the previous year the use of Big Data Analytics solutions in order to be able to respond in real-time to transaction requests and to recognise patterns of fraud attempts [22].

For commerce institutions and banks, with plenty of customer data (derived from, e.g. credit card transactions), and insurance companies, with less frequent customer interactions [31], predictive models based on data analysis enrich the customer experience and improve the communication relevancy between financial institutions and customers. Combining internal and external data along with sentiment assessment on social networks, institutions aim to create a 360° customer view.

[6] Hadoop Distributed File Systems, http://en.wikipedia.org/wiki/Apache_Hadoop
[7] http://en.wikipedia.org/wiki/Paysafecard

Financial institutions make use of their internal data (collected by the institution) and, in recent years, perform real-time data analysis on social networks to watch news about companies, to evaluate prices and opinion trends, or to do damage control, as well as the data being fed into predictive models of economic forecasts or trading data. Continuously changing regulatory and compliance requirements lead to the need of better analysis algorithms that move towards better risk reporting and improved transparency. In Austria, such regulations are issued by the Austrian Financial Market Authority[8] (FMA) and by the Austrian National Bank[9] (ÖNB), which also overview the Austrian finance sector.

The importance of this application domain was seen as highest on the short term by 32% of the online survey participants. Another 26% see it as important on the medium term, and 19% see it as important for the long term. The Austrian banking sector is behaving conservatively, employed IT solutions often lagging behind existing novel solutions. It is wished that the solutions provided by research teams and companies have solid European references before they are translated into practice. Data related to finance, published voluntarily on social networks is currently not used. Generally, the existing data is of good quality, though the different architectures in use pose a challenge in integrating the data stored in different repositories.At the management level, the importance of intelligent data handling is often overlooked.

4.9 Public Sector and Government Administration

The benefits of using advanced analytics in the government and public sector are increased efficiency and transparency. Two classic examples of government services are tax and labor agencies. The main activities of tax agencies include managing submissions, organizing examinations, administering collections, and providing services to the taxpayer. Labor agencies perform market and consumer analysis, as well as provision and management of employment services and benefits. Optimizing these processes with interconnected data sets and more powerful data analytic tools has advantages for both citizens and governments. The citizens' advantages include shorter paper trails (data does not have to be re-stated), which lead to fewer errors and faster results. The government can collect taxes more efficiently and minimise the so-called "tax gap"—the difference between what taxpayers owe the government and what they pay voluntarily. For labor agencies, powerful data analytics tools can help speed up retrieving relevant offers and matching them with the profile of the job seeker. In all these processes special care must be taken to ensure that the citizens' private data is protected when integrating and interconnecting different data sets.

At the European level, the public sector could reduce administrative costs by 15 to 20 percent and create an equivalent of €150 billion to €300 billion in additional value [91, p. 54]. A recent report of the British Policy Exchange initiative [141] projects that £16 to £33 billion of extra value could be generated per year for Britain when using Data Analytics correctly.

Administrative data in the public sector is primarily textual or numerical. This means that they generally deal with smaller data sets than other branches such as the health sector described in Section 4.1. The McKinsey study [91, p. 56] found that the increase in digital data creation in the public sector administration is also due to the many successful e-government initiatives from the last 15 years. Public sector agencies, however, either do not see the use

[8] http://www.fma.gv.at
[9] http://www.oenb.at

of making their data public or often are inefficient in publicizing it and communicating it internally to colleagues, citizens and businesses. Inconsistent data formats and input protocols create additional difficulties. Digital data is not moved electronically but with old technology (e.g. fax, CDs, post). Strict policies and additional legal restrictions often also prevent sharing and using data for advanced analysis.

In Austria, essential steps were taken towards a more efficient public sector. However, numerous island solutions exit in the different governmental areas, with every department having its own databases and solutions [111]. Data security and availability is of highest interest in all these solutions, as well as the 'right to forget', where it is important to guarantee that data is indeed deleted [111]. A project that was already started in 2001 had as the main objective the introduction of the electronic filing (ELAK–der elektronische Akt) to replace paper based filing and archiving in all Austrian ministries [101, 5]. The system shortens the paper processing times by 10-15%, has over 9,500 users and is used not only at the federal level but also at the provincial government level and other Austrian institutions [13]. Bundesrechenzentrum GmbH has been awarded the 2012 EuroCloud Europe Award for its E-Gov Portal Services. The cities of Vienna, Linz and Graz provide much of their data as part of the Open Government Initiative[10] so that companies can develop applications that ease daily life. It is, though, general knowledge that governments generally lack data analysis, cleaning and integrating capabilities. In a six country comparison (Austria, Sweden, Switzerland, Germany, United Kingdom, and USA), Austria ranks first in the use of eGovernment services (with 65%) [86], but there is a call for more security and trust in the light of negative trends in the users' satisfaction with these services.

This application domain is seen to be more important on the medium term by the survey respondees (41%).

4.10 Tourism and Hospitality

Tourists leave digital traces on the Web and through interactions with mobile technologies. The resulting data is not only massive but also multidimensional (e.g. movements through space and time) and requires new approaches for storage, access, and analytics [71]. This increasing amount of structured and unstructured data and the availability of Data Analytics technologies are changing the theory and practice of hospitality and tourism businesses. Companies are using Data Analytics technologies to anticipate customer needs, rewrite how they meet customer expectations, redefine customer engagement, and achieve new levels of customer satisfaction. In the end this creates a new basis for the award of customer loyalty. The hospitality and tourism business is about "selling experiences" and about "expectations". If fundamental expectations are not being met, it detracts from what the experience could be and negates the type of delight that a customer could have. In a commoditised marketplace, differentiation is about being able to build on basic transportation, accommodation, and destination services to offer a variety of personalised customer interactions. Current developments in Data Analytics make an individual high-class experience possible on a mass market basis.

Companies in the hospitality and tourism domain are required to look outside their enterprises for critical information; to base operations on both internal and external data resources; to leverage data as they relate to customers moving through space and time; and not so much to measure customer satisfaction and respond to complaints as to design, implement

[10] http://data.gv.at/

and assess entire customer travel experiences. An example of how tourism statistics combined with active digital footprints (like social network postings) may reflect an image of the tourist dispersion among sites in Austria is described in [84]. Workshop attendees considered that the companies in the tourism and hospitality currently suffer penalties from not being more active in using Intelligent Data Analytics.

The future may belong to those firms best able to shape and deliver the consumer travel experience. In doing so, advantage may go to companies with the longest history in offering hospitality and tourism services. Identifying preferences and affinities may become as important to travel and hospitality companies seeking customer loyalty as being able to provide these services themselves [80].

Respondents to the online survey estimated that this application area is important on the medium term (28%). We note here that another 28% of the survey participants did not know how to answer our question about this area's importance on the short, medium or long term.

4.11 Commerce and Retail

Business Intelligence analyses data collected by an organisation with the aim of transforming it into useful and actionable information. While business intelligence activities have existed for many decades, the way of analysing the data has been changing recently due to the transformation from having to actively collect data to being faced by a flood of potentially relevant data. *Competitive intelligence* is related to business intelligence, focusing on the environment of operation, in particular on the competitors, but also on customer requirements in general or the economic environment.

One of the first areas that took advantage of Data Analytics at a large scale was the stock market, where high speed algorithmic trading has made some fortunes, although the impact of such trading on stock market crashes is under debate [7].

Data Analytics is having a significant effect on retail [91]. Through online available data, consumers have improved means to compare products in terms of features and prices, often in real time. This results in increased transparency, and allows the customers to put pressure on prices. Available data is growing also on the retailer side, by recording customer transactions and operations, product tracking, and customer behaviour and sentiment. Retailers make increasingly informed decisions by mining customer data. For example, online retailers can use recommender systems to offer customers products matched to their tastes. In a real-world shop, consumer behaviour may be tracked by video-surveillance systems, allowing the retailer to improve store layout, product mix and shelf positioning. Other examples of Data Analytics in the retail sector are the use of mobile device generated data to display profile-based advertisements targeted to nearby customers, the use of weather and social media information to adjust offer and pricing to trends, product quality and freshness tracking by supply chain monitoring.

Companies with large R&D investments apply Data Analytics methods to scientific publications and, in particular, patents to create state-of-the-art technology landscapes to support decisions to invest in research and product development.

The Commerce and Retail application domain was considered important on both short and medium terms (32% and 34%, respectively) by the online survey participants.

4.12 Law and Law Enforcement

This area is related to the government and public sector as described in Section 4.9. Traditionally, police departments collect millions of service calls, archives thousands of reports in combination with months of audio and video recordings. Data Analytics helps to explore and explain crime statistics much better when combined with additional data sources, like locations of interesting objects (housing, businesses, ATMs and local events). In an experiment, the police department in Santa Cruz, California, integrated data collected from 5,000 crimes dating back to 2006 and used it for predictive crime forecast [104]. An algorithm predicted the features of certain crimes (e.g. location, time) based on past crimes, and deployed officers before the crimes were supposed to happen. Even though some might feel reminded of Hollywood's 'Minority Report', the first round of experimentation lowered new crime by up to 11%. The use of Data Analytics to prevent crime, also called Predictive Policing, is gaining some attention in the US. In [102] the crime recording process is enhanced with data mining techniques that assist the crime prediction process, e.g. clustering, classification, deviation detection, social network analysis, entity extraction, association rule mining, string comparison and sequential pattern mining.

Such techniques can help to better understand and differentiate the districts of a town. Police officers can be assigned more efficiently based on such findings to improve crime prevention. Concurrent data analysis can be the decision support for strengthening or reducing the force when necessary in a timely manner. It transforms the re-assignment of officers into a dynamic and data-driven process in comparison to one that is based entirely on the perception and the experience of seniors. This allows policing to become more accountable, prevents power abuse and improves the living quality of entire neighbourhoods on an informed basis.

Financial crime detection is an area that benefits from Data Analytics, when, for instance, statistical and predictive methods are applied on transactions [98], where financial fraud investigators have to comb through extensive financial and bank transaction reports to find suspicious activities (see also Section 4.8).

Due to the significant increase in the amount of written information that must be searched in legal proceedings, "it is becoming prohibitively expensive for lawyers even to search through information" [109]. This is leading to advances in the field of *eDiscovery*, which are being noticed in the US court system, for example, in the case Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251 (D. Md. 2008), the court proceedings mention that "there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search," and mention work done in the TREC (Text REtrieval Conference) evaluation campaign.

In Austria concrete steps to better use technology in the law domain were carried out in 2013 as it has been decided that courts of law must use the electronic file system ELAK (der elektronische Akt) starting with April 2013 [10].

Regarding the existing laws, especially those that define data privacy, data use, etc., the workshop participants believe that law makers are lagging too much behind the current technological developments. Online survey participants considered the law and law enforcement application domain to be most important on the medium term (36%). A high percentage of respondees – 30% – did not know whether this domain will be more important on the short, medium, or long term.

4.13 Earth observation

Earth observation satellites generate many Terabytes of data per day, which is complemented by data from other platforms such as sensors on airplanes. This data is routinely processed in applications such as weather forecasting, wild fire observation, earthquake and volcano research and land use analysis. There is also a large potential for the development of commercial applications based on earth observation data. However, the development of these applications has been held back by the complexity of gaining access to the data, which has usually been captured at great expense to the organisation operating the satellite, and hence requires extensive negotiation about usage and payment conditions between the company doing the application development and the data owner. Recent initiatives have used the Cloud paradigm to simplify access to the data. The SuperSites Exploitation Platform running on the Helix Nebula Science Cloud[11] provides a common platform and data for geo-hazard scientists. The company CloudEO[12] provides a full Cloud-based ecosystem for earth observation data, on which content providers can make data available, developers can create applications using the data, and end users can access the applications, with the fees paid by end users fairly divided between the data and application providers.

4.14 Agriculture

Farming has changed drastically in the last 50 years. Smartphones, portable computers, RFID tags on livestock, and other environmental sensors are nowadays used to collect agricultural information. This information is then consolidated with, for example, climate data and market conditions, provided by public and private institutions, to help farmers do their business planning [25]. Opportunities of Intelligent Data Analytics in agriculture include higher efficiency, reduced machine operation costs, better crop productivity, healthier livestock with impacts on food safety and environment, and better activity planning using data provided by weather and earth observation systems[13]. In the end, it is believed that this growing body of data will ultimately free producers of much of the day-to-day guesswork associated with farming [89].

4.15 Media and Entertainment

In [91, p. 10], this sector is listed as being less capable of capturing the value of big data by lacking relevant skills and having a less data-driven mindset. Interestingly, at the same time, this area is one of the most IT intensive, with semantic technologies being used to structure contents (e.g. BBC's semantically structured websites [83]). Collaborative filtering and recommender systems are also used to find similar media objects (songs, video clips, articles). Others [106] recognise the benefits the entertainment industry has had from the digitisation of media, too. The main barriers in making the most out of the entertainment data are the siloed organisational structures, disconnected cross-functional and multi-business unit processes [106]. The strategies taken to increase the companies' competitiveness use Data Analytics and Big Data algorithms to manage customer experience, consolidate the hardware

[11] http://www.helix-nebula.eu/index.php/helix-nebula-use-cases/uc3.html
[12] http://www.cloudeo-ag.de/
[13] see e.g. http://www.deere.com/wps/dcom/en_US/campaigns/ag_turf/farmsight/farmsight.page

systems management to address the multitude of devices, services and networks currently in use, or develop new, on-demand delivery business and consumer products.

4.16 Further Application Areas

We briefly discuss some additional application areas, in order to give a wider picture of the capabilities and opportunities that come with using Data Analytics on large quantities of data.

Sports. Since many years, sensors in racing cars are used to read tire pressure and temperature, oil temperature, brakes, etc. To give drivers an edge on winning a race real-time data analysis is employed to take decisions on the driver's course of action during a race [82]. Sport teams turn to Big Data to evaluate players, analyse strategies or past games.

Gaming. The competitive and growing gaming industry is driven by two key factors today: lifting the top line through better understanding of the customer and efficiently managing the bottom line through operational excellence. Superior data-driven applications at the front line of the business, both on the customer-facing aspects as well as internal operations, can provide a competitive advantage to gaming companies and help them meet their revenue growth and operational efficiency goals. This can be enabled by fast and rich analysis of large volumes of data that gaming companies, both online and land-based, gather about their customers and their operations. These data volumes are exponentially growing as companies employ more sophisticated tracking of their customers, capturing each individual interaction at the most granular level. Such detailed data related to the customer's behavior and preferences, if analysed quickly, can provide invaluable customer insights that can help gaming companies effectively segment their customers, identify high potential customers, and execute customer relationship strategies that would help maximise their revenue. Similarly, the ability to analyze fresh operational data efficiently can reduce costs and plug revenue leakage caused by issues such as game and payment fraud [6]. Through loyalty programs companies understand the behavior patterns borne out in hundreds of thousands of interactions per day [78].

Real estate. Commercial or residential, publicly owned or private, real estate is a beneficiary of the use of Data Analytics. Facility managers use past utility readings from digital sensors to analyse and predict future energy consumption, plan their estate improvements, etc. Real estate brokers can create better, personalised marketing offers based on real-estate search engine logs, past sale data, neighbourhood crime rates, public transport connectivity. Building inspection and construction authorities can also tap into Data Analytics algorithms and effectively use their rich body of real estate documents on which development plan decisions are based. Other stakeholders in this domain are banks (marketing mortgages), insurance companies, investors and housing companies.

Defence and Intelligence. Modern defence systems must nowadays collect more information than ever. Defence and intelligence agencies are in need of efficient real-time Big Data solutions to correctly identify and respond to possible threats. The nature of the new threats is changing to have a cyber dimension, e.g. attacks on national infrastructure networks, economic institutions, financial systems, etc. To counter these threats, the industry must provide safer ICT products and Big Data analytical methods. In Austria, several projects involving the National Army have been started [14]. In the frame of the Cyber Security Initiative founded in 2011 with the aim to increase the awareness on IT security issues, the Cyber Security Forum was opened this year – a forum where decision makers can meet and share experiences and solutions.

Chapter 5

Challenges and Open Issues

This chapter presents a summary of the important challenges in the Intelligent Data Analytics area. We begin by presenting the drivers and challenges aggregated from the different stakeholders' viewpoints on data and its use. This includes results from interviews with Austrian C-level managers and top-level management staff from industry and the public sector, an online survey, and the workshops with primarily academics and researchers. The chapter subsequently covers in detail not only technological challenges related to Intelligent Data Analytics, but also challenges to be faced in education, personnel, gender and diversity as well as a wider legal or societal framework.

A broad consensus among the experts exists about the importance of data privacy. It plays a key role when operating with data and it is an important driver for increasing the customers' trust in the provisioned services. In most cases adhering to regulations that ensure data privacy results in loss of information (e.g. removing personal details from health records). However, experts also argue that the regulations are to some extent outdated and need to be aligned with today's practices. It was also criticised that an appropriate legal framework for handling and operating on data is still missing. The uncertainty about the legal situation undermines the courage of the industry and also the public sector for publishing data as well as for the take-up of (Big) Data Analytics technologies. As an example consider the smart meter rollout in Austria. Even though there is the legal obligation of energy suppliers that 95% of standard meters are replaced with smart meters by 2019 (cf. [61]), the current regulations cause uncertainty among companies regarding the rollout which is primarily linked to the customers' opt-out option. Additionally, the mostly pessimistic news coverage about data privacy matters negatively impacts the trust in these technological advances and might also increase the number of smart meter opt outs and eventually reduces the data volume available for analytical purposes. As a consequence, a society of distrust rather than a society of trust has been established. Moreover, experts at the workshops expressed concerns because of a rising tendency towards technology aversion in Austria, Germany and Switzerland. Thus, experts suggest to give top political priority to data and IT in general. This strategy would be a political gesture highlighting the importance of these topics. Primarily in Scandinavian countries such as Sweden or Finland, pursuing this strategy has led to top-ranks in the current Networked Readiness Index (NRI), which measures the propensity for countries to exploit the opportunities offered by information and communications technology [139]. Despite Austria's advantageous position (e.g. political stability, secure environment, moderate climate, wide

range of resources available, central geographic location...) it is not among the top ranked business locations (i.e., currently Austria's NRI rank is 21[1]).

Experts highlight that one of the biggest challenges is to determine the value that can be obtained from data. In fact, the actual return on – the still rather high – investment is difficult to assess and remains a major hurdle in technology take up, even though show cases and best practices for a wide range of application domains and industries already exist. This can be attributed to the predominant opinion that such technologies are only applicable in domains or populations generating very high data volumes.

Gaining a 360 degree view on the individual is a primary goal in many domains. This requires standardised access to as many data sources as possible including for example sensor data, structured data gathered through forms, survey, surveillance information, unstructured data from online communication, etc. The experts' opinions about the technological capabilities of available tools are rather diverse. Some state that current tools already offer very sophisticated capabilities for minimizing the time to insight. Others point out that current approaches lack sophisticated data enrichment and in-depth data understanding capabilities. If the limitations in semantic data processing can be overcome, advanced insight generation approaches relying on this metadata will assist in turning data into actionable information. A further technological hurdle is caused by the absence of satisfactory tools for combining data from multiple, different enterprise and non-enterprise data sources [110].

The stakeholders pointed out that the lack of qualified staff that can ask the proper questions, produce the necessary (statistical) models, makes use of the available tools and has the capability to communicate the insights properly imposes a major challenge to the industry. It is important to perform awareness-raising actions at an early stage in order to educate children to become mature and literate in data issues. Computer science education at schools should be compulsory and intensified. Lectures imparting digital skills and media competence should become a foundation of today's school curricula. Graduates need to have the skills to properly classify data, i.e. they need to have gained an understanding of the importance of data. Additionally, innovative concepts and tools for highly efficient learning are required in order to impart maximum knowledge transfer in minimum time. A promising concept is, for instance, the emerging e-learning trend of Massive Open Online Courses (MOOC)[2]. In summary, industry is seeking polymath thinkers rather than experts with a narrow skill set. This species is in short supply and very difficult to find.

In summary, the most important challenges in Intelligent Data Analytics range from Privacy, Security and Data Ownership over algorithmic and technological shortcomings to shortages in the supply of qualified personnel. Above all, determining business cases that will eventually allow obtaining real value from data remains *the* major challenge.

5.1 Data Representation

One of the main challenges observed in Intelligent Data Analytics relates to the data fed into the analytical processes. Data is often unsuitable for the task at hand and must be transformed before it can be analysed — this transformation process is often very complex and time-consuming. Algorithms have to deal with data that is dynamic (like financial market

[1] Networked Readiness Index, see http://www.weforum.org/issues/global-information-technology/the-great-transformation/network-readiness-index

[2] http://en.wikipedia.org/wiki/Massive_open_online_course

data); is multimodal, combining different types of information (text, photos, audio...); is unstructured or semi-structured, potentially written in natural language (where nuance useful to humans complicates machine processing); or is simply extremely large, where the data is not all of equal value and sieving through it to get any insight is still an issue. This *heterogeneity* and *incompleteness* is a major challenge [17]. For many tasks that require reporting carried out by humans, it is usually more efficient to collect unstructured or semi-structured data rather than imposing a completely structured reporting format. It is also important for algorithms to be able to deal with missing data, and to take into account potentially systematically missing values that could lead to a skewing of results. Data cleaning and error correction algorithms are useful for countering these problems, but are usually not infallible. Biased sampled data that may affect secondary data use (with further data processing steps) also represent a challenge that has to be dealt with — if the initial data was collected with a biased sample (e.g. a questionnaire about the use of social media distributed only via Facebook), then lack of knowledge of this bias could result in false conclusions when the data is reused elsewhere. Lack of standards or lack of adherence to standards within certain domains complicates data integration and reuse.

Challenges in data integration and fusion were discussed with the experts in the areas of data quality, context, users, users' privacy, language and social media. Data quality must be assured or maintained. One aspect is that expert knowledge needs to be modeled accurately to assure that data stays correct and trustworthy when integrating with other sources. Methods that have been applied on the data need to be traceable and integrated data results must be reproducible. Incorrect data sets (e.g. text) need to be filtered. Evaluation of data quality should be possible even if the ground truth is incomplete. The integration of contextual aspects (e.g. temporal or geographical) was discussed including the importance of applying sensors for collecting context for data that is integrated. It is not possible to have context-free data and the task for which data was collected is almost always an integral part. Therefore, it is important to include the task as part of the context model and bind it with the data (task-oriented data integration). Contextual information can therefore be used to assign data with a particular domain and there remains a challenge of accurately matching data between domains by matching their contexts. Users require methods of data integration to be understandable with a simple human interface to the integrated data and how it got there. Privacy plays an important role and it is challenging to assure anonymity when combining data from many sources (k-anonymity, cf. [126]). Blurring of data (e.g. in the healthcare sector) or the removal of contextual data (e.g. in the legal sector) is often practiced, but formally not well understood and not very standardised. Social media often provides only very short texts and limited contextual information that challenges effective data integration. New methods need to be developed to adjust to this new requirement. The 'multi-language nature' of data needs to be used much better for data integration. Speech recognition still has areas with unsolved problems (e.g. detecting names) that are paramount for transforming spoken audio and integrating it with textual sources.

Data is seen as an important asset but has not yet been seen as an asset with explicit value. Data should have explicit structural properties that allow determining its quality, that enable data to be integrated, restructured and reduced for multiple (and also secondary) domains and purposes, and that is made accessible over large data platforms for broader audiences. Contextual information is often not explicitly associated with data but instead contained in surrounding infrastructure (such as the application processing the data) or in an expert's head. Semantic structures should be enriched from these implicit sources (e.g. by

controlled vocabularies). Standardised evaluation of data and ontologies are required to test data reliability (e.g. its trustworthiness) and data quality (e.g. causal connections of data points). Metadata markets were proposed as a potential meeting point for requirements and exchange of explicit data descriptions.

5.2 Techniques, Methods and Algorithms

This section presents the challenges for the techniques, methods and algorithms listed in Chapter 3, based mostly on input collected from the workshops.

5.2.1 General

Algorithms play an essential part in every phase of Intelligent Data Analytics. From deciding which pieces of data to store, and in which format, to extracting the right information supporting a business decision, algorithms are involved at every step. As Intelligent Data Analytics is a highly interdisciplinary field which adopts methods and aspects from other research fields, the set of algorithm types available to choose from is large and varied (Chapter 3). The choice of (sets of) algorithms or techniques has to be done depending on the kind of insights sought for the issues of interest, issues to which existing or incoming data might provide answers. At a high level, these algorithms can be split into two categories: data acquisition and data processing or analysis. This split is often not clear, with many algorithmic solutions integrating both.

Among the currently most popular approaches to overcome the problem of scale, we mention parallel algorithms [36] to split the workload among several computing units, extracting representative samples of data to reduce the computing workload [125] and streaming algorithms that process the data as it is collected [62], where the data is unknown before execution. But making an algorithm scalable often causes it to be less flexible. Observing how data is generated and collected we expect that future solution implementations will use adaptable algorithms. Changes in data format and representations as well as the introduction of new hardware platforms are very likely to happen. Therefore, algorithms should be designed having in mind that the context in which they are used may change and they should be able to self-optimize and/or self-verify for correctness.

It is generally a requirement that, even though the amount of data to be processed increases, the algorithms' output is given in a reasonable amount of time. This translates not only into speed requirements for algorithms but also into lower energy usage requirements (green computing), leading to lower costs. What is 'reasonable' depends very much on the kind of task given to the data analysis tools. For example, understanding what a customer's goals were up to a given point in time is not as urgent as deciding what products to display in side ads at the moment a customer browses an online shop, but the former crucially influences the latter [23]. Research into creating better methods to analyse and extract knowledge from large data repositories is rich in improvements and novel ideas (see for example [16, 55]).

It is mentioned in [17] that it is no longer sufficient to rely on processors getting faster according to Moore's law to cope with increasing volumes of data — data volume is now scaling faster than compute resources. Due to power constraints, processor clock speeds are no longer increasing significantly, but processors are being built with increasing numbers of cores. *Parallel data processing techniques* are needed, but techniques developed for inter-node parallelism don't directly apply for intra-node parallelism due to the different architecture.

Parallelism for multiple nodes each with multiple cores adds another layer of complexity. *Cloud computing* introduces extensive sharing of resources, and requires new ways of determining how to run and execute data processing cost-effectively through system-driven optimisation of the program execution. Finally, as hard disk drives are being replaced by solid state drives, and with new storage technologies expected in the near future, data processing algorithms will no longer be faced with a large spread in performance between sequential and random I/O performance, with implications for the design of many algorithms.

5.2.2 Search and Analysis

The workshops discussed search and analysis from four different viewpoints: users, data, algorithms and systems, and contextual issues: *User issues* focus on the understanding of results. For users it is important to get results that carry explanatory power and to provide visualisations that describe and explain the data directly. A low learning curve should support users who wish to understand results from complex data analysis. The search and analysis area is particularly susceptible to the *data representation challenges* described in Section 5.1, as these approaches are often applied directly to raw data. Enterprise search is seen as a special and yet under-recognised area where data sets are much less structured and less integrated than the open web and its search engines. Future applications, with respect to the algorithmic viewpoint of search and analysis, are decision support systems, exploratory search and cross-lingual search. Classic database management systems (DBMS) are viewed by many as important. Despite the technological shift away from strongly structured data formats and SQL technology, these systems will remain important in the near future.

Evaluation, verification and benchmarking of data was considered an important issue. Manually compiled test sets, so called Gold Standards, are largely missing and yet there is a strong need to check facts accurately and discover potential connections within and between data sets reliably. Furthermore, performance and scalability were discussed in the light of developing *algorithms* for large data sets and towards solutions that eventually offer real-time predictions. The modularisation of search engines should be further developed to simplify reuse and repurposing of search engine components into services. Finally, transparency was discussed in the context of search/ranking algorithms as well as their evaluations. This aspect is hardly addressed in the current search engine world – end users have no insight into whether the results that a search engine produces are biased in any way. *Context* is important for enriching data with meaning by adding meta-level information about the potential uses of the data and its adaption to new application areas and new situations. Similarly, algorithms could be expanded with information about their potential uses and application and task areas, leading to semantically adapted search and analysis.

5.2.3 Semantic processing

It is no surprise that a prime focus in the workshop discussions on semantic data analysis were ontologies and the format and structure of data itself, as covered in Section 5.1. The necessity of connecting already existing ontologies and taxonomies to widen their application area was pointed out. Also, semantic structures should be able to link data with appropriate analytical methods. The discussion also highlighted the need for data mining on the terms of high-level semantic information rather than raw data. In this context, it is not clear how effective algorithms are in obtaining semantically enriched data and how to attach semantic information

to formats such as images, videos etc. The creation of ontologies is challenging as it can either occur deductively (top-down modeling) or inductively (bottom-up generative modeling) which generally leads to very different ontologies. Regardless of how ontologies are created, they require constant updates as a natural evolution adapting it to changing requirements. Crowd-sourcing is viewed as one possible instrument to facilitate this modification. Natural language processing is a key component of semantic data analysis and one of its challenges is that systems need to be able to deal with large data sets and the interactive nature of dialog systems as well as specialized and minor languages. From the user perspective, semantic processes need to be combined with workflows and tasks, and ontologies must provide a front-end for users to raise understanding and awareness rather than being hidden within a system. Technological challenges on efficiency are focused on storage and processing of ever growing data sets (e.g. many short texts in lower quality such as tweets or blog comments) with little semantic information. Effective processing requires noise tolerant models that demand less training. Expert knowledge needs to capture a multi-cultural scale with multiple domains being addressed and transformations possible between domains and cultures.

5.2.4 Cognitive Systems and Prediction

Cognitive systems face the challenge of understanding human emotions and human behaviour. Emotions like "attention", "forgetfulness", "forgiveness" impact on applications like traffic control decisions, public transport, or social media company postings, applications that aim at understanding how and why people affected by them react in certain ways. Human daily behaviour, for example, is of importance in Assisted Living Systems that are used by elderly or physically challenged persons in order to quickly react to health damaging events. Another area where behaviour observations are of importance is cognitive systems implementing recommender systems, where there is often a difference between stated and revealed user preferences.

Regarding the machine learning processes involved in cognitive systems and prediction, a set of further challenges were identified by the workshop participants:

- *available training data:* How should good cognitive and prediction/forecasting models be designed when only little, high quality training data is available?
- *domain identification and algorithm efficiency:* When the application domain is correctly identified, suitable machine learning methods and efficient algorithms are relatively easy to choose. Identifying the domain and choosing suitable algorithms in an automated way is considered to be a challenge, especially when Big Data is involved;
- *expert knowledge:* Besides correctly identifying the application domain such that suitable algorithms are used, there is a stringent need of correct and sustainable employment of expert knowledge, both in process design and execution;
- *tangibility:* The knowledge generated by the systems should be made available in ways that are easy to grasp by the systems' end users;
- *re-usability:* Efficient models should, ideally, be easy to transfer to new application domains and should include the expert knowledge specific to the new domain;
- *scalability:* Methods currently do not scale up well, both with respect to data amount and data dimensions. Interactive analysis, for example, is usually not possible on huge amounts of data. Parameter tuning for a large number of parameters is another recurring

issue. It is wished that new models and methods are created with Big Data in mind, and indeed tested on Big Data.

- *streamlining:* Carrying over newly gained knowledge to other application domains should be streamlined and, where applicable, done in real-time;
- *filter bubble:* Determine what should be infused into the learning process/algorithms to avoid the inability to infer knowledge out of the system's usual scope.

A recent trend in modeling cognitive systems is to (re-)place humans in the centre of cognitive systems by means of crowdsourcing. A present challenge related to employing crowdsourcing, for example in product development, is to determine and implement a fair distribution of intellectual property rights.

One of the recurring requirements that workshop participants have voiced is that all operations on the cognitive and prediction processes are traceable and reproducible, especially in complex systems such as cell cancer recognition in oncology departments. This closely relates to and forms a basis for assuming responsibility when system actions or reality distanced predictions are followed by physical and monetary damages. To support accountability of decisions based on cognitive systems or on predictions/forecasts, it is required that processes are transparent and explanatory, and thus increase the trust in such systems. It is a challenge to develop methods for quality assurance, for validation and for verification of processes implemented by the cognitive systems and prediction techniques. This is especially true and needed for the cases where actions are taken based on prediction techniques have effects that are fed back into the cognitive systems that made the prediction.

5.2.5 Interaction with Data

Intelligent Data Analytics requires human-accessible interfaces that allow efficiently sifting through data. Thus it is important to provide a means that enables users to intuitively capture the value of data in order to derive meaning and actionable information. Visualisation is regarded as a key discipline in supporting the user in his/her quest of gaining insights and in making sense and is, thus, an essential part in the user interaction process that facilitates data exploration. This exploration process requires approaches that adequately reduce high-dimensional data to less complex models while making the causality of applied (processing) steps transparent to the user. In many cases, dimensionality reduction negatively impacts the traceability and understanding of why a particular set of information is displayed or discarded. Addressing this challenge, i.e. to provide traceable results, fosters the trust in the technology itself since the communication endpoint facing the user forms the basis for decisions made by the user.

Problem-oriented, personalised visualisation requires approaches that provide different views on the same data depending on the respective context (e.g. geospatial, temporal, provenance, etc.), use (e.g. emergency notifications vs. charts and detailed reports) and most importantly the user (e.g. management, researchers, pilots, surgeons, etc.). Especially streamed, real-time data imposes challenges on several levels in data analytics processes and in particular on visualisation when real-time interaction with huge amounts of data is needed.

It is essential to raise awareness on the limits as well as the appropriateness of a particular (visualisation) approach in a specific context (e.g. speech for hearing-impaired people). Educating users and making them aware of the advantages as well as the limitations of a particular approach is important as well (e.g. GPS hardly works inside buildings).

The ultimate objective, however, is to provide and receive feedback via all senses. Thus, the future of data interaction involves multiple input and output modalities such as visual, audio, haptic, olfactory, etc. Developing solutions that provide the optimal balance between the amount of information fed back to the user and distraction is, therefore, an important but very tough challenge.

5.3 Evaluation

As is clear from Chapter 3, there are many algorithms, methods and techniques available for Intelligent Data Analytics. This means that there are usually multiple ways to solve a problem, and it is usually not immediately clear from the beginning which combination of algorithms and techniques will produce the optimal solution. The often requested "toolbox with guidelines" for solving data analytics problems has to be built on a deep understanding of the techniques and methods and how they perform in different situations. Part of this understanding can be gained through extensive evaluation of the algorithms, techniques and methods.

Currently, whether the optimal solution is found can often depend on the aptitudes and skills of the people solving the problem. Kaggle[3] takes advantage of the dependence on aptitude and skills by crowdsourcing solutions to data analytics problems. Companies post data and associated problems as competitions on the Kaggle platform, and anybody can submit solutions to the problems. Incentives include prize money, fame (people winning multiple competitions are promoted on the Kaggle main page), and occasionally employment offers. However, while such platforms may be a useful source of well-functioning solutions, they do not provide deep understanding of how and why the solutions work and why certain solutions are better than others.

Solutions published in the scientific literature are usually evaluated by comparing a proposed solution to one or more other solutions on one or more datasets selected by the authors. These datasets can be created by evaluation campaigns or originate elsewhere. Evaluation campaigns such as the Text Retrieval Conference (TREC)[4] for information retrieval and the PASCAL Challenges[5] for machine learning play a central role in creating common datasets and tasks on which solutions can be compared. Nevertheless, problematic aspects of this type of evaluation include the use of unsuitable or proprietary datasets, or comparison to poor baselines, leading to "improvements that don't add up" [20] or an "illusion of progress" [72]. In the computational sciences in general, little focus has been directed towards the reproducibility of experimental results, raising questions about their reliability [58]. There is currently work underway to counter this situation, ranging from presenting the case for open computer programs [79] to considerations about the legal licensing and copyright frameworks for computational research [124]. Research infrastructures to allow reproducible computational research [58] by allowing sharing of data, algorithms, techniques and methods are a promising approach to improving the reproducibility of computational results and also to extend the range of problems on which Intelligent Data Analytics techniques and methods can be evaluated.

It is important that Intelligent Data Analytics solutions have a positive impact on end users by allowing them to complete their work more effectively and efficiently. To achieve this,

[3] http://kaggle.com
[4] http://trec.nist.gov
[5] http://pascallin2.ecs.soton.ac.uk/Challenges/

extensive insights into how end users make use of Intelligent Data Analytics solutions and how these solutions impact on their everyday work are needed. These insights could be obtained by observation of end users performing their tasks, but this approach is time-intensive, and the unconstrained situation could lead to less scientific insight. Other approaches such as more constrained tasks performed in a "laboratory setting" under control of the researchers, or even performed using crowdsourcing, are a possibility.

5.4 Privacy and Security

Privacy is the selective ability of individuals or groups to withhold themselves or information about themselves entirely or in part[6]. Often, individual privacy is related to anonymity: information that does not allow a person to be identified. *Security*, on the other hand, "is the practice of defending information from unauthorised access, use, disclosure, disruption, modification, perusal, inspection, recording or destruction[7]" and generally covers more technological aspects such as encryption in storage, transmission and access control (e.g. authentication). Security can be used to protect privacy by, for example, encrypting sensitive information. However, security could also breach privacy, for example when a company stores copies of personal data in multiple geographical locations for increased security [34]. A recent paper from the Cloud Security Alliance presents the top ten Big Data security and privacy challenges [4]. These challenges range from secure computations in distributed programming frameworks, over secure data storage and transactions logs to scalable and composable privacy-preserving data analytics.

Privacy concerns increase with high data volumes and advanced analytics applications. Data protection in Austria is part of the constitution, existing since 1978, and was last revised in 2000. The Austrian data protection authority (in German *Datenschutzbehörde*) is a governmental authority charged with data protection. The European Data Protection Supervisor is the independent supervisory authority devoted to protecting personal data and privacy and promoting good practice in EU institutions by monitoring how the EU administration processes personal data and advising on policies and legislation. These local authorities are connected with additional data privacy instruments on the European level. One of the most powerful and comprehensive legal efforts toward privacy regulation today is the EU Data Protection Directive. It regulates the processing of personal data within the European Union. The Directive on Privacy and Electronic Communications (2002/58/EC), also called E-Privacy Directive, covers new digital technologies and electronic communications services and technologies such as data retention, unsolicited emails and cookies. In addition, the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No. 108) "is the first binding international instrument which protects individuals against the collection and processing of their personal data.[8]" It regulates transfer, processing and flow of personal data, and outlaws the processing of "sensitive" personal data in the absence of legal regulation also with respect to transborder exchange (e.g. information about a persons ethnical background, political interests, religion, etc.). This convention allows individuals to request disclosure and correction, which is however restricted by so-called overriding interest (e.g. interests of the state or defence). As the latest advancement, the

[6] Paraphrased from http://en.wikipedia.org/wiki/Privacy
[7] http://www.en.wikipedia.org/wiki/Information_security
[8] http://conventions.coe.int/Treaty/en/Summaries/Html/108.htm

European Commission unveiled a draft of the European General Data Protection Regulation in January 2012 that will supersede the EU Data Protection Directive. This regulation extends the scope of the EU data protection law to all foreign businesses that process data of Europeans. Violations are punished severely. This propagates European data privacy beyond Europe with a potential to further standardise privacy efforts internationally.

Violation of data privacy (e.g. if personal data is transmitted without consent, is false, or was not deleted as requested), as far as the Austrian constitution is concerned, are addressed by the Austrian data protection authority (who assists with consultancy) and the civil court who applies EU regulations. This however does not quite meet the current technological state with high data volumes, high-frequency and multi-source recording, and advanced data analytics that may exploit all these properties. In such a context, previously law-abiding anonymised data sets can be used to re-identify people and put the entire legal effort in jeopardy. One of the most notorious cases in the past is AOL Research that published twenty million search queries for 650,000 users of the AOL's search engine, summarizing three months of search activity in 2006 [103]. User names and IP numbers were suppressed in an effort to anonymise the data set. However, AOL decided to maintain the connections between queries by assigning unique user identifiers to enable research on the data set. Soon thereafter, it was discovered that people could be re-identified from the anonymous data based on what people searched for and on the personal information included in the search queries (e.g. names and places). Perhaps the most disturbing fact was that no advanced data analysis was needed to discover very private and potentially harmful information about individuals[9]. This risk naturally increases with bigger data sets and better tools.

Although initially addressed by European and international law, there is need for more technological regulation by the law. We do not yet know how to share private data in a way that ensures sufficient data utility in the shared data while limiting disclosure and protecting anonymity. The existing paradigm of differential privacy is a very important step in the right direction, but it unfortunately reduces information content too far in order to be useful in most practical cases. In addition, real data is not static but increases and changes over time; none of the prevailing techniques result in any useful content being released in this scenario. Yet another very important direction is to rethink security for information sharing in Big Data use cases. Many online services today require us to share private information (think of Facebook applications), but beyond record-level access control we do not understand the implications of sharing data, how the shared data can be linked, and how to give users fine-grained control over this sharing [17].

Our online survey revealed that more than 70% of participants regard privacy and security as an important short-term to mid-term challenge. This indication of both importance and urgency was further strengthened in the three workshop events where privacy and security was one of the central topics. Many other themes discussed at the workshops quickly diverted into privacy and security discussions. This suggests that despite participants' shared natural fascination about the possibilities of data analytics, they are very aware and concerned about its abuse. The law was generally viewed as being too slow to adapt to the fast-changing technological developments of an increasingly globally operating world. Since the law generally operates on a national level, there are many grey areas in a world-wide data network. Tensions exist between an open style that allows for innovation and creativity and their restriction to ensure privacy and the rights of the individual. While innovation is important, ethical

[9]One of the re-identified users was a woman who extensively searched for ways to kill her husband.

boundaries are often broken and the law is limited to the task of fixing loopholes in the social system. Generally, technology and law should collaborate better. Also, society has to take more active interest in technological developments and create influence in legal issues to speed up their adjustment.

The loss of trust is one of the key elements when talking about privacy and security. Unfortunately, this loss of trust is almost never as valuable as the the potential gain for those who violate that trust. High level security solutions should be advanced and not rely on existing or future legal support. It is more important to prevent abuse technologically than by law. Security checks should be the technological measure to demonstrate how safe private data is. The Snowden case[10] created some awareness about what it means to post or transmit personal information online. However, there is still a large educational gap where people are not aware of how to handle private data and treat it with the same care as other valuable possessions. One problem is that insecure user activities are generally very accessible. Secure counterparts usually require much more effort to use. Since the common user is not aware or capable of applying the extra technology that would make a transaction secure, the unsafe default is commonly used (e.g. by not encrypting email correspondence since it would require additional software and the understanding and handling of public/private key pairs). What makes it even worse is that data, once leaked, has no expiration date and can be stored anywhere forever. It should be possible to better control data with respect to the maximum length of storage and a right for personal data being forgotten by design. The *Right to Forget* is about users getting to control their own data, or data posted about themselves by others. This is not only an issue in the context of data stored by sensors such as smart meters, but it is a rather complex problem with backups and the ultimate challenge on the internet.

5.5 Data Ownership

The problems faced when talking about *data ownership* are illustrated in a 2003 Information Management column by D. Loshin [88]. It starts from the definition of ownership and lists different paradigms in which potentially different entities may have rightful claims to ownership: Creator, Enterprise, Funder, Decoder, Packager, Subject (e.g. the person in the photograph), and Purchaser. The different paradigms are partially a result of the distinction between two types of data: those related to individual physical persons and those not related. The spectrum is defined by the relatedness parameter.

For the general public, data ownership is tightly related to personal data protection, while for the professional public, it has a stronger link to Cyber-Security. As such, the two aspects, data protection and cyber-security can be seen as two sides of the same coin, and this has been observed by the EU Justice Commissioner in a recent speech before the NATO Parliamentary Assembly [117]. Overall, the regulatory issues are now being re-considered across the world, as evidenced by the two new proposed directives of the EC mentioned above, the relatively controversial Cyber Intelligence Sharing and Protection Act (CISPA) in the US, and a recent study covering 63 jurisdictions [12].

The matter of data ownership is difficult because data can be changed in unforeseeable ways, at which point it may be considered new data and rightfully assigned new ownership. However, the precise point at which processed data becomes new data is unclear. One definition, related to data ownership, is based on the possibility or impossibility to re-generate the original data

[10] http://en.wikipedia.org/wiki/Edward_Snowden

from the "new" data. However, the notorious AOL search log data release incident [19] (also mentioned in Section 5.4) showed that original data may be reconstructed given sufficient external data, outside the control of the owner. The issue is complicated beyond the means of this report. A short report of current problems from a legal perspective was recently published by Graham and Lewington [68].

Finally, it is worth noting that ownership of data, as in the case of some physical objects, is not only a set of rights, but also a set of responsibilities. Owning data, selling, licensing, making it available for processing, to generate new data implies a statement of quality or integrity. The consumer estimates these properties by an implicit or explicit credibility assessment of the input data and, in the face of large or complex data, automatic tools may need to be developed to assist the consumer in making this assessment. This is the aim of data governance, which "embodies a convergence of data quality, data management, data policies, business process management, and risk management surrounding the handling of data in an organisation. Through data governance, organisations are looking to exercise positive control over the processes and methods used by their data stewards and data custodians to handle data[11]". This relates closely to issues of proof and trust in data quality and their associated analytical methods.

One central discussion topic during the workshops was Open Data and the difficulty to decide when data is new and not simply a narrowly-related form of derived data. This creates a complicated interplay with copyright, data protection rights, and rights to use data including licensing. Another issue was the tension between open innovation versus data ownership which limits innovation, if intellectual property rights can be implemented and how they relate to product liability. Finally, more than 80% of online survey participants indicated that the challenge of data ownership needs to be addressed in the short to mid-term.

5.6 Data Economy and Open Data

The direct impact of Open Data on the EU27 economy was estimated at €32 Billion in 2010, with an estimated annual growth rate of 7% [131]. In order to encourage this development, the European Commission has set as a priority "to coordinate the development and implementation of a strategy for the European data value chain leading to a set of actions that will help to nurture a coherent European data ecosystem and that will contribute to increased efficiency in a range of data-intensive sectors as well as to the emergence of a high number of innovative data-related products and services[12]". A data value chain is defined as "a value chain where most of the participants have data resources as an important factor of production and produce as their output or consume as additional inputs other data resources or data technologies (e.g. database management systems, visualisation toolkits, data mining or machine learning frameworks, etc.) or services (e.g. recommendation systems, navigation systems, etc.)[13]". The provision of a data ecosystem should allow Small and Medium sized Enterprises (SME) to easily create innovative solutions based on data, and hence create revenue streams.

In search of a good analogy, data has often been referred to as the "new oil[14]". Already in 2006, Michael Palmer wrote "Data is just like crude. It's valuable, but if unrefined it cannot

[11] http://en.wikipedia.org/wiki/Data_governance
[12] http://ec.europa.eu/dgs/connect/en/content/data-value-chain-european-strategy
[13] http://ec.europa.eu/digital-agenda/events/cf/ict2013/document.cfm?doc_id=26738
[14] http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/

really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analysed for it to have value[15]." This analogy has been taken further by Jer Thorpe[16] who warns of "data spills" (when personal data is inadvertently released), "dangerous data drilling practices" and that a great deal of profit is currently being made through the use of human-generated information such as browsing habits and movement information ("where oil is composed of the compressed bodies of long-dead micro-organisms, this personal data is made from the compressed fragments of our personal lives").

Open Data (the provision of data, available for free use and re-use with an open license in machine readable open formats[17]) forms one of the foundations for Data Value Chains. According to a recent CapGemini report [131], Austria is classified as a *beginner* in terms of positioning and pace of adoption of Open Data initiatives, as it has poor Open Data availability and poor data portal usability. In contrast, our workshop participants generally considered that the Open Data domain in Austria is quite advanced. It was, however, expressed that there is catching up to do, especially since data that conforms to the Open Data principles [48] are hard to come by. The United Kingdom is a country with strong government support for open data that is in the Trend Setter class in [131]. A recent UK initiative is the Open Data Institute[18], which functions as a *Data Incubator*. Data Incubators offer support to start-ups and SMEs in order to remove administrative or technical obstacles in getting access to data, but also offer financial support to allow these start-ups and SMEs to complete their data-driven project plans.

The European Commission (EC) fosters Open Data activities very intensely via several activities as part of the EU Digital Agenda[19]. In the Horizon2020 Framework Programme there are concrete funding mechanisms for Open Data[20], where Open Data is also a vertical topic in in the programme area of Societal Challenges[21].

Furthermore, the EC has set up several Open Data activities, including but not limited to: a) the support of member states in publishing increasing amounts of Open Data, b) the request towards funded R&D projects that created data needs to be published as Open Data for further re-use and c) the revision of the Public Sector Information Directive (PSI Directive) that has to be implemented into national law of EU28 member states until July 2015[22]. Last but not least the EC pushes the topic of Linked Open Data[23] (LOD) in all above mentioned activities to bring open data to its full potential. Linked Open Data provides an efficient and sustainable set of methods and techniques to enable an ideal basis for a Data Value Chain[24]. Furthermore, Open Data and Linked Open Data principles are increasingly used inside of organisations to enable powerful data management.

However, Open Data on its own is not sufficient to feed a data ecosystem. A recent McKinsey report [93] points out the "clear potential to unlock significant value by applying

[15] http://ana.blogs.com/maestros/2006/11/data_is_the_new.html

[16] http://blogs.hbr.org/2012/11/data-humans-and-the-new-oil/

[17] http://en.wikipedia.org/wiki/Open_data

[18] http://theodi.org/

[19] http://ec.europa.eu/digital-agenda/en

[20] ICT 15 Open Data and Big Data Innovation and take-up as well as INSO-1 ICT-enabled open government

[21] http://ec.europa.eu/programmes/horizon2020/h2020-sections

[22] http://ec.europa.eu/digital-agenda/en/news/what-changes-does-revised-psi-directive-bring, http://blog.okfn.org/2013/04/19/the-new-psi-directive-as-good-as-it-seems/

[23] http://en.wikipedia.org/wiki/Linked_data, http://linkeddata.org/

[24] http://5stardata.info/

advanced analytics to a combination of open and proprietary data." Access to proprietary data is currently in general difficult to impossible, and action is needed to facilitate and encourage access to proprietary data. Furthermore, a data ecosystem would facilitate the reuse of data in multiple domains (e.g. the reuse of weather data in the agricultural, insurance and energy domains) as well as the reuse of intelligent data analytics solutions across multiple domains (e.g. process modelling methods in both the manufacturing and logistics domains).

5.7 Data Curation and Preservation

Data constitutes an enormous investment and value. This is growing as data is being reused in meta-studies or repurposed and integrated with other sources across domains. This requires measures to be taken to ensure that the investment made into data is made sustainable by ensuring long-term data availability via appropriate data curation. This is also essential to support verification of decisions made based on data, or to allow re-evaluation of older data with newer models. Beyond the mere challenge of redundantly and securely storing vast amounts of data in distributed settings, which poses significant engineering challenges, there are numerous unsolved issues that require significant research to tackle.

For companies required to adhere to particular quality standards as, for example, in Life sciences, data preservation is of great importance. In order to comply with reporting obligations covering data that dates back 15 or more years, companies perform backups but also keep and run their legacy hardware and make print outs of their data [113].

With the value of scientific data having been recognised[25], most research institutions are currently developing *data management policies*, while most research proposals need to provide a *data management plan*. Currently, many data management policies are not connected to operational procedures, and most plans are simply text stating intentions with no means of enforcement, monitoring and verification. Methods need to be devised to consistently document the policies and steps in a machine-readable and machine-actionable manner, or it will be impossible to scale up data curation operations to meet the requirements posed by the massive amounts of data in an enormous variety of styles and formats [2]. *Process management plans* go further than just the data, and allow the data context to be captured. This should ensure that essential components of this context, such as specific processing steps, can be kept operational over longer periods of time [107]. This requires completely new machine-actionable models of process management plans [100].

Data citation allows persistent links to be established between data sets and result presentations. Several approaches for assigning persistent identifiers (PIDs) to support data citation have been proposed [9, 1]. Current approaches do not support citation of arbitrary subsets in a machine-processable way and provide only limited support for dynamic data. Finally, for data to remain useable and useful, we need to be able to read and interpret it in an authentic manner over time. This raises specific challenges on two levels: the *logical (format)* and *semantic (interpretation)*. As formats evolve and older formats loose their support, tools need to be developed to transfer them into newer data repositories. At the semantic level, data changes its meaning as terminology evolves. Ensuring correct and authentic interpretation

[25] See Recommendation 4 on *Preservation and re-use of scientific information* in the European Commission Recommendation of 17.7.2012 on access to and preservation of scientific information (http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf).

across data sets collected over longer periods of time will require novel means of (automatic) data annotation. This includes all aspects of data collection and processing (e.g. characteristics of physical sensors or of data collection techniques), to interpretation aspects encoded in transformations, filtering and appraisal decisions taken. Furthermore, terminology needs to be developed to be able to capture data semantics in a machine-processable manner [3].

5.8 Austrian Shared Computing Infrastructure

The processing and sustainable storage of large amounts of data in appropriate infrastructures has been studied by many computer science research groups for many years. Especially in data-centric science, the possibility to share resources for computing and storage is essential for sharing data and algorithms, but also to ensure access to computing facilities of sufficiently high performance to deal with huge amounts of data. Different approaches can be grouped under the three broad topics: High Performance Computing (HPC), grid computing and Cloud Computing, for which the relevant competences have been available for many years in Austria and Europe.

Since 1989 various Austrian activities and initiatives have aimed to gather the requirements of researchers from all disciplines so as to coordinate and implement corresponding (new) infrastructure. These include the Austrian Centre for Parallel Computing (ACPC), the AustrianGrid I and the AustrianGrid II projects, the Vienna Scientific Cluster[26] (VSC) and the Austrian Centre for Scientific Computing[27] (ACSC). The VSC is the largest high-performance computing infrastructure available in Austria, while the ACSC provides several infrastructures with different software and computer architectures (shared memory, distributed memory, GPGPU, Cloud) to their members. A uniform access to all infrastructure is not available. Furthermore, the development of a joint strategy for all cooperation platforms and the targeted integration and communication of requirements with other existing initiatives and platforms (academic as well as industrial in nature) in Austria (Austrian Academy of Sciences, ACOnet[28], EuroCloud[29], etc.) and also internationally takes place only to a limited extent. In Austria, project and discipline-specific development of computing resources also takes place (Austrian Academy of Sciences, Central Institute for Meteorology and Geodynamics, etc.) — these are not connected to or reconciled with the current two central Austrian platforms: ACSC and VSC.

The European Strategy Forum on Research Infrastructures[30] (ESFRI) recommends a hierarchical structure of a European HPC infrastructure, which PRACE[31] is currently attempting to implement. At the highest level there are a few European data centres that provide top computing power. These are provided as part of PRACE currently available. Below these are national data centres, already implemented in many European countries, and invariably not located at a single university, but as independent organisations that bundle the HPC skills of a country and have capacity for the largest research projects of all national research institutions (universities, Universities of Applied Sciences, Academies of Sciences, non-university research centres). Austria is one of the few European countries that operates no national academic data

[26] http://vsc.ac.at
[27] http://acsc.uibk.ac.at
[28] http://www.aco.net
[29] http://www.eurocloud.at
[30] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri
[31] http://www.prace-ri.eu

centre, but builds computing infrastructure in a competitive environment between universities and research institutions.

This strategy leads to an extremely fragmented computing landscape, which impedes initiatives to share data and algorithms so as to bring together researchers with backgrounds in specific subjects (such as astrophysics, genetics, economics) and extensive experimental data, representatives of industry with real-world problems linked to huge collections of data, and researchers in the area of Intelligent Data Analytics.

5.9 Qualified Personnel

According to the McKinsey report [91], there are three types of talent required to take full advantage of Intelligent Data Analytics. *Deep analytical talents* are people with technical skills capable of integrating and analyzing data to derive insights. *Data-savvy managers and analysts* who have the skills to be effective consumers of data insights, i.e., capable of posing the right questions for analysis, interpreting and challenging the results, and making appropriate decisions. *Supporting technology personnel* who develop, implement, and maintain the hardware and software tools such as databases and the analytics programs needed.

Although this McKinsey report looked at the requirements from the view of the industry, a recent paper in Nature [94] pointed out that for the growing area of eScience, a new breed of researcher equally familiar with science and advanced computing is required. Such a researcher would also be expected to have deep analytical talent. These people with deep analytical talent have been given the name of *data scientists*. According to Wikipedia, a data scientist has the task of extracting meaning from data, and, in order to be capable of doing this, should master techniques and theories from many fields, including mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualisation, uncertainty modeling, data warehousing, and high performance computing [138].

Our workshops confirmed the high-level and multi-layered expectations on the requirement profile of a data scientist, especially as far as interdisciplinary technological and mathematical/statistical know-how is concerned. However, a shortage, as estimated for the USA by McKinsey [91], that predicted an expected gap of 140,000–190,000 people with deep analytical talent in the USA in 2018, was not in the scope of our workshop discussions. Surprisingly, senior university researchers did not point to an expected gap in their working area, nor did they refer to any scenarios on how to prevent Austria from a possible shortage. This is why the aspect of gender or migration, speaking of the idea how to attract more people and/or import experts from other demographic groups was neglected throughout the World Cafes.

On the other hand, our survey as well as expert interviews with representatives from Austrian industry and public organisations showed that shortages of highly qualified personnel are expected and are regarded as an important challenge. Companies even referred to a current lack of good people highlighting the necessary interplay between analytical and technical competence with an empathetic personality, i.e. being eager to talk to customers of different application fields. From a business point of view, data scientists are urgently required, especially as far as the capability to build the right models for data analysis and – in the following – streamlining operations, improving production, informing business decisions, etc. is concerned.

In summary, we believe that it will take time to increase the rate of production of data scientists, and universities should consider creating the necessary interdisciplinary degree

programs to achieve this. Internationally spoken, efforts are already being made to increase the attractiveness of this career option with headlines such as "Data Scientist: The Sexiest Job of the 21st Century" (cf. [76]). We herewith apply suggestions coming from the World Café discussions where experts agreed on the necessity to improve current university curricula. First, the claimed interdisciplinarity and the capability for knowledge transformation requires know-how about different application areas and thus, joint curricula between technological and mathematical/statistical faculties as well as social sciences. In order to apply technological know-how to different areas, holistic approaches towards problems are required. From the university side, this may include early formation in school labs but also the possibility to bring own ideas and projects to the university and thus enrich formal curricula through creative and individual aspects. These suggestions are believed not only to improve formal education and training but would additionally enhance the development of a joint terminology throughout the different research fields.

5.10 Gender and Diversity

In Austria, more women graduate from universities than men. Subsequently 42% junior scientists holding a doctoral degree are female [123]. Even though the gender gap has diminished in education, and although the number of female researchers in Europe increases more than two times stronger than the group of male researches, women are still not equally represented as men in higher positions: Only 20% of the top research positions are held by women [51] and women remain marginalised in a number of research areas, especially in engineering and technology. Female full professors in these research fields for example do not exceed 7.9% in the whole European Union [51].

When taking a closer look at the disciplines involved in the field of Intelligent Data Analytics, only a small number of female leaders are represented in industry and research: an Austrian Gender Monitoring Report (cf. [40]) reveals that female project leaders make up 11% in ICT, which is significantly lower than in other sectors (an average percentage of 25% is provided in the report).

Although Gender Mainstreaming efforts and supporting efforts to promote women have been extremely intensified during the last decades, structural and social barriers still hinder women from advancing their career as far as men. Studies mostly refer to a masculine understanding of the scientific and/or engineering profession (cf. [53]), informal selection and promotion processes (cf. [32]), but also psychological barriers and traditional role-ascriptions and role-expectations of women [99]. Further, informal exclusion makes women feel unwelcome [28, 63]. Thus, access to research is gendered and structural barriers for women still exist.

During the study, we put a focus on integrating and purposefully inviting women to participate in our research and reflected on gender issues with the gender expert Dr. Brigitte Ratzer[32].

However, throughout the study, especially during the identification of experts, high gender segregation in the field of data analytics became apparent with respect to senior and/or management positions. Despite our own reflection on the topic of gender and its relevance throughout the whole study, only a few women actually took part in the process. While 5

[32]Dr. Brigitte Ratzer is the head of the Center for Promotion of Women and Gender Studies at Vienna University of Technology. She has more than 40 publications in the field of Gender and Technical Research (http://www.tuwien.ac.at/en/services/gender_studies/home/team/brigitte_ratzer/)

(out of 12) members of the advisory board were female, only 4 women participated in the online survey. In total, 11 (of of 61) workshop participants were women. In fact, the Vienna workshop attracted 22.2% female participants, Graz 12.5% and Salzburg saw 16.7% female participants.

A surprising observation was that gender aspects in and of data analytics were neither brought up during the workshops nor by the interviewed experts. Not any respondent of the online survey raised this topic. Neither did anyone refer to gender issues during the discussion about personnel, nor did — men or women — mention any specific experiences in this context.

Gender as a relevant category or social structure was thus not recognised in the environment of data analytics throughout this study. This, however, does not mean that it is less important. On the contrary, it shows that emphasis has to be put on gender issues in Intelligent Data Analytics in order to prevent especially this new, innovative area from homo-social reproduction [32]. Gender assumptions that remain invisible to the scientific community nevertheless influence scientific priorities, research questions and choices of methods which may — in the long run – lead to a bias into a scientific field and its assumptions (cf. [121]).

There seems to exist a background assumption that gender plays no significant role in Intelligent Data Analytics. One can argue that the implicit picture of a data scientist – as the picture of scientists — follows traditional assumptions such as being a male, white mid-aged men dedicating all his life to work [15, 114]. Such assumptions do not only affect role ascriptions of men and women in the field, but equally addresses the question of who is defined and accepted as an expert in the field and who is going to be qualified and promoted.

Invisibility of women in research often refers to their status: they less often hold management positions and conduct research in scientific niches as structural barriers such as recruiting and award procedures and/or male networks hinder women from striving for a scientific career the same way as men do. Being confronted with a male-dominated area where gender seems to be no or an allegedly neutral category, the European Commission (cf. [52]) suggests research questions that address the current state of knowledge on gender norms, identities or relations in the respective research area. What happens if gender is not being addressed or being misunderstood or misrepresented in the field of Intelligent Data Analytics? First of all, it is about data scientists and data researchers, their current research, their practices etc. Knowing this, further questions can be addressed: what are the implications of the findings? What do we miss, which aspects have not been considered yet? Which marginalisation processes need to be investigated and addressed?

Additionally, women — as well as members of minority groups — are much more at risk to be assigned to less prestigious and administrative tasks, which puts them in a position of continuous competitive disadvantage [137, 32]. The important questions remaining are "Which concepts led to this gender-denying attitude?" and "How can we close the gender gap in the disciplines linked to Intelligent Data Analytics?"

5.11 Societal and Economic Challenges

Optimally, society should be prepared in advance for the consequences of the increasing ubiquity of data analysis, rather than passively reacting to the changes. At present, the topic of data analytics is mostly visible in the public media in Austria in a negative way, usually related to privacy fears (e.g. doctors sell patient data to pharmaceutical companies[33], NSA

[33] http://help.orf.at/stories/1723343/

spying[34]). Such breaches are obviously a problem, and it is necessary that the use and abuse of data be more carefully regulated by law. Nevertheless, the topic is currently only fully understood by a rather small "elite" group of people.It is necessary to educate a much wider group of people to successfully live in and take advantage of the knowledge society.

The long term aim of such an education is to make Intelligent Data Analytics a commodity,i.e. a tool for everybody in the knowledge society, allowing all members of society to make more informed decisions. The first step is to make people more sensitive to the data that they generate and its value and hence to give them control over it. This can be initially done through a better awareness of which data is private and should not be shared, which data about them is owned by other players (e.g. online shopping portals, social networks), the effects of long term storage of their data and their rights and possibilities to delete data. The possibilities that their data provides to others, such as linking separate data sets, should also be made clear. With this better awareness of data and its value, it should be possible, in the medium to long term, for people to make informed decisions about what happens with their data. A "trade in data" can be started. Incentives can be created for data to be made available to certain trusted and transparent organisations for certain tasks — some people currently choose to make their genome sequences available to companies in return for information on their future disease probabilities and the knowledge that their genomes are being used in medical research to improve the estimation of these probabilities. Alternatively, valuable personal data could even be sold or leased by its owners if they decide to do so.

[34] http://derstandard.at/1381373815989/NSA-naher-IT-Dienstleister-fuer-Gesundheitsministerium-taetig

Chapter 6

Austrian Intelligent Data Analytics Competence Landscape

This Section describes the Austrian Intelligent Data Analytics Competence landscape. The Austrian Data Analytics community is divided into three groups: *researchers* in academia or research institutes; *service providers* (often small and medium enterprises) with the expertise to implement solutions to Data Analytics challenges for clients; and *end users* — companies with data and challenges to be solved by Data Analytics. After a summary of each of these groups, the competence landscape is presented. In summary, Austrian strengths are in the areas of statistics, algorithmic efficiency, machine learning, computer vision and Semantic Web.

6.1 Research Landscape

For the universities, research groups working in the area of Intelligent Data Analytics or a related area were identified through a manual analysis of the university websites. For each potentially relevant research group, a list of all publications of the leading researchers (professors and associate professors) was extracted from the DBLP Computer Science Bibliography[1], under the assumption that these leading researchers are usually co-authors on the majority of papers published by a research group. The titles of all publications were extracted and tokenised, and a count of the word occurrence over all publication titles allowed the main research foci to be determined. This was complemented by a short perusal of the research group web page. For research institutes and universities of applied science, more emphasis was placed on obtaining the research foci from the web pages, due to a usually smaller number of publications. With this approach, we aimed to extract the major areas of competence of each organisation, while leaving out the minor areas of competence (e.g. areas in which only a handful of papers have been published)[2].

[1] http://www.informatik.uni-trier.de/~ley/db/

[2] In such a procedure, it is of course possible to omit some institutes or competences, and we make no guarantees for correctness or completeness. We thank all people who submitted corrections to the Position Paper Competence Landscape, and have tried to take these into account. We apologise for any remaining omissions. We hope to eventually use the information gathered here as the basis for an online competence landscape, in which further corrections can be made.

Table 6.1: *Colour legend for Tables 6.2 and 6.4.*

Colour	Meaning
	Carinthia
	Lower Austria
	Styria
	Salzburg
	Tyrol
	Upper Austria
	Vienna
	Company with headquarter outside Austria

Table 6.2 summarises the data obtained. All research groups are clustered by their host organisation, and their research foci are classified into the groups defined in Chapter 3, leading to a matrix showing the Intelligent Data Analytics research competences available in the Austrian universities, universities of applied sciences and research centres. The colours indicate the province, according to the legend in Table 6.1.

While it is clear that all groups evaluate their algorithms, the evaluation column indicates groups that have organised a larger-scale evaluation activity, such as an evaluation campaign or challenge.

6.2 Service Providers

The service providers in the area of Intelligent Data Analytics or a related area were identified through desk research and the manual analysis of company websites. The focus was on identifying and reviewing Austrian companies. However, some large international players active in the field of Intelligent Data Analytics were reviewed as well. During this review, we have a) identified the application domains and industries the companies operate in, and b) tried to elicit the technological foundations of their offerings. Table 6.4 summarises the results of the review of the service providers. The service providers are categorised according to their activities in the selected application domains. While *Manufacturing and Logistics*, the *Public Sector and Government*, *Finance and Insurance* as well as *Transportation and Travel* are addressed by a rather large number of providers, *Market Research*, *Defence and Intelligence* as well as *Law* are less prominently addressed by the reviewed providers (see Figure 6.1).

6.3 End Users

This study focusses on creating an Intelligent Data Analytics technology roadmap, so we have not created a landscape of the end users of Intelligent Data Analytics solutions, apart from the broad contours of the application landscape described in Section 4. A further study focussing on the end users of Data Analytics solutions and Big Data is currently underway[3].

[3] http://bigdataaustria.wordpress.com

Table 6.2: *Research foci. Colour meanings are in Table 6.1. Organisation full names are in Table 6.3.*

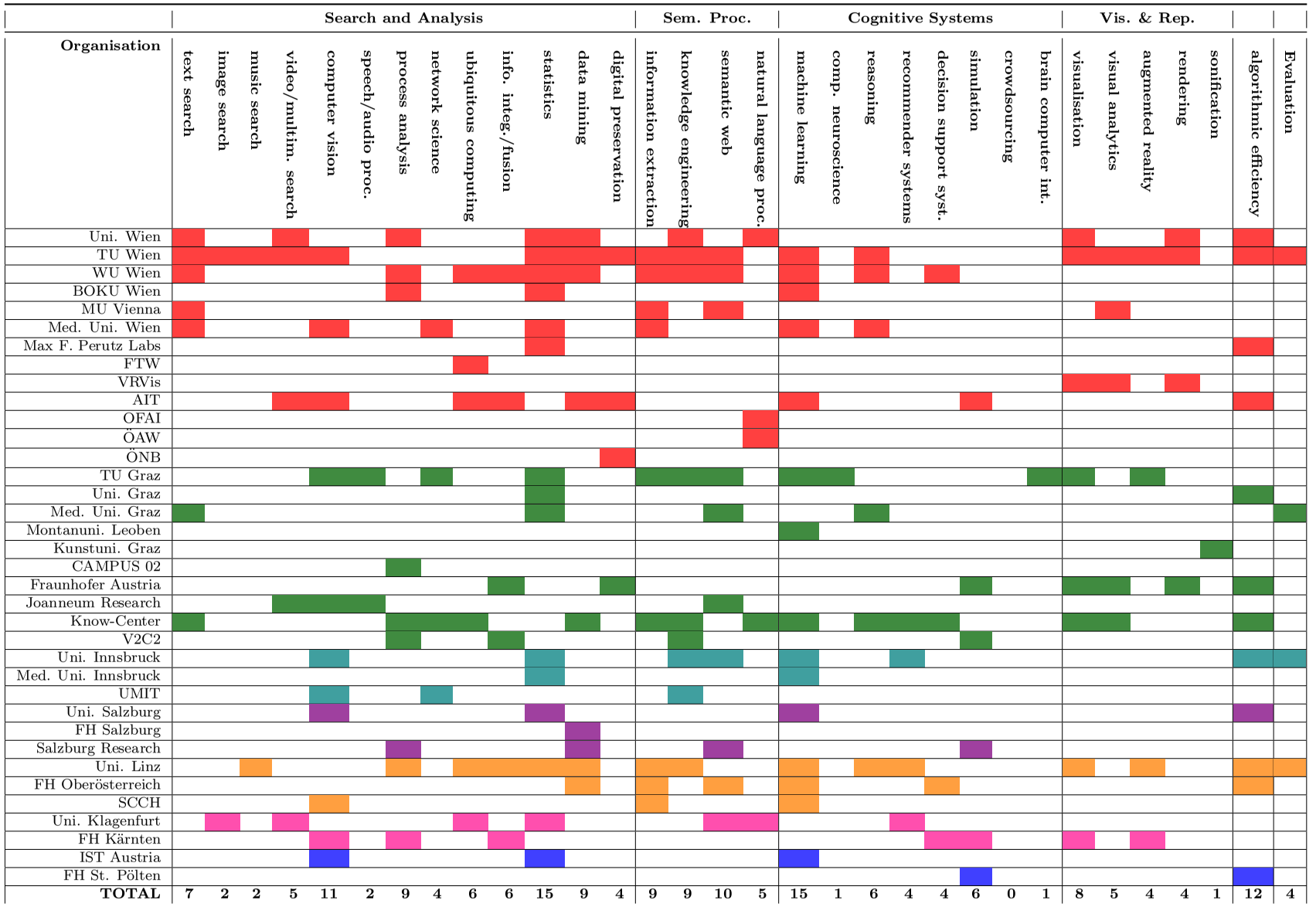
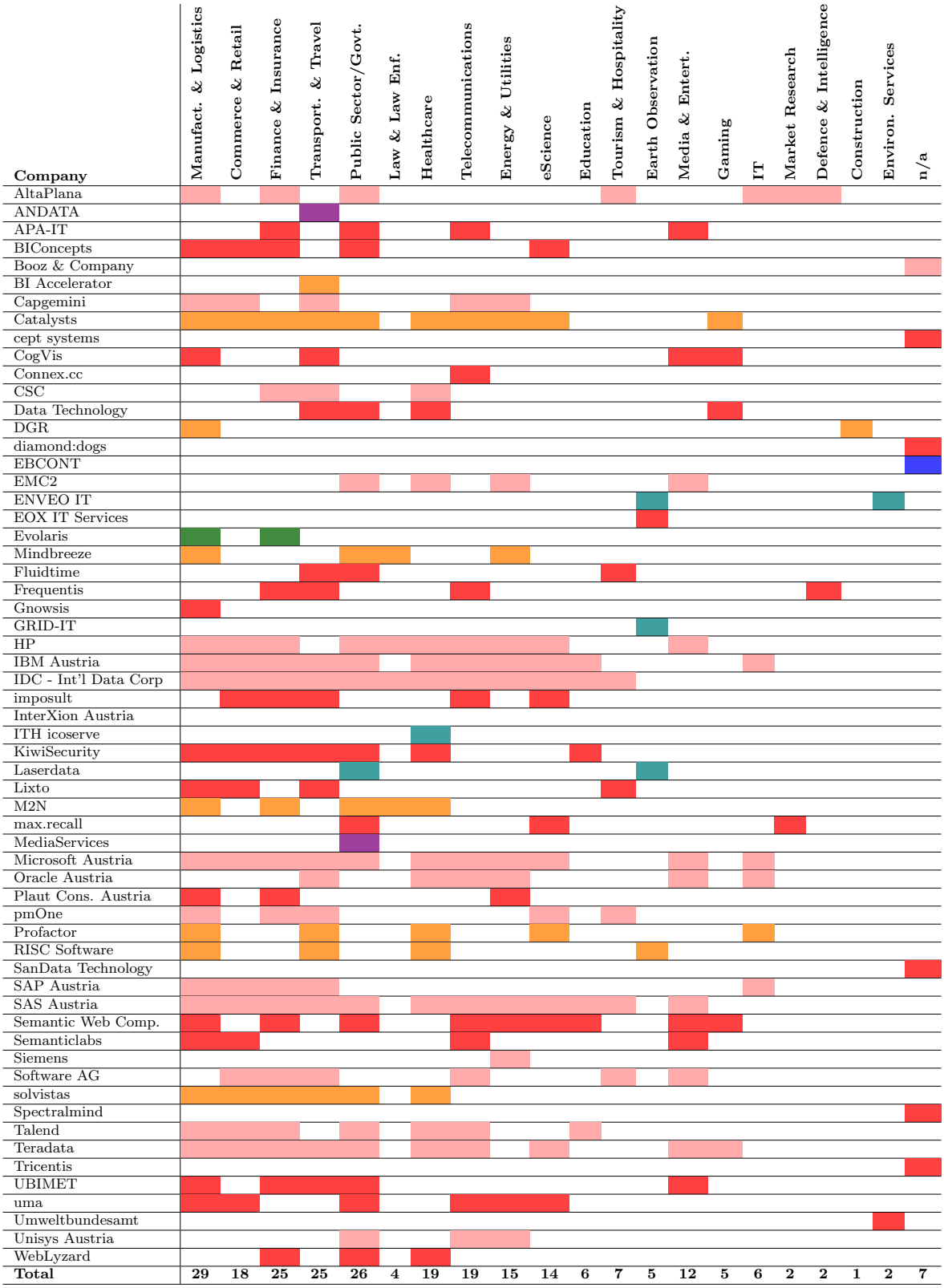


Table 6.3: *Expansion of the Research Institute Abbreviations used in Table 6.2.*

Abbreviation	Full Name
Uni. Wien	University of Vienna
TU Wien	Vienna University of Technology
WU Wien	Vienna University of Economics and Business
BOKU Wien	University of Natural Resources and Life Sciences, Vienna
MU Vienna	MODUL University Vienna
Med. Uni. Wien	Medical University of Vienna
Max F. Perutz Labs	Max F. Perutz Laboratories
FTW	Telecommunications Research Center Vienna
VRVis	Zentrum für Virtual Reality und Visualisierung
AIT	Austrian Institute of Technology
OFAI	Austrian Research Institute for Artificial Intelligence
ÖAW	Austrian Academy of Sciences
ÖNB	Austrian National Library
TU Graz	Graz University of Technology
Uni. Graz	University of Graz
Med. Uni. Graz.	Medical University of Graz
Montanuni. Leoben	University of Leoben
Kunstuni. Graz	University of Music and Performing Arts Graz
CAMPUS 02	CAMPUS 02 University of Applied Sciences
Fraunhofer Austria	Fraunhofer Austria
Joanneum Research	Joanneum Research
Know-Center	Know-Center
V2C2	Virtual Vehicle Research Center
Uni. Innsbruck	University of Innsbruck
Med. Uni. Innsbruck	Innsbruck Medical University
UMIT	UMIT - The Health & Life Sciences University
Uni. Salzburg	University of Salzburg
FH Salzburg	Salzburg University of Applied Sciences
Salzburg Research	Salzburg Research
Uni. Linz	Johannes Kepler University Linz
FH Oberösterreich	University of Applied Sciences Upper Austria
SCCH	Software Competence Center Hagenberg
Uni. Klagenfurt	Alpen-Adria-Universität Klagenfurt
FH Kärnten	Carinthia University of Applied Sciences
IST Austria	Institute of Science and Technology Austria
FH St. Pölten	St. Pölten University of Applied Sciences

Table 6.4: *Application domain by company. Colour meanings are described in Table 6.1.*



Company	Manufact. & Logistics	Commerce & Retail	Finance & Insurance	Transport. & Travel	Public Sector/Govt.	Law & Law Enf.	Healthcare	Telecommunications	Energy & Utilities	eScience	Education	Tourism & Hospitality	Earth Observation	Media & Entert.	Gaming	IT	Market Research	Defence & Intelligence	Construction	Environ. Services	n/a
AltaPlana																					
ANDATA																					
APA-IT																					
BIConcepts																					
Booz & Company																					
BI Accelerator																					
Capgemini																					
Catalysts																					
cept systems																					
CogVis																					
Connex.cc																					
CSC																					
Data Technology																					
DGR																					
diamond:dogs																					
EBCONT																					
EMC2																					
ENVEO IT																					
EOX IT Services																					
Evolaris																					
Mindbreeze																					
Fluidtime																					
Frequentis																					
Gnowsis																					
GRID-IT																					
HP																					
IBM Austria																					
IDC - Int'l Data Corp																					
imposult																					
InterXion Austria																					
ITH icoserve																					
KiwiSecurity																					
Laserdata																					
Lixto																					
M2N																					
max.recall																					
MediaServices																					
Microsoft Austria																					
Oracle Austria																					
Plaut Cons. Austria																					
pmOne																					
Profactor																					
RISC Software																					
SanData Technology																					
SAP Austria																					
SAS Austria																					
Semantic Web Comp.																					
Semanticlabs																					
Siemens																					
Software AG																					
solvistas																					
Spectralmind																					
Talend																					
Teradata																					
Tricentis																					
UBIMET																					
uma																					
Umweltbundesamt																					
Unisys Austria																					
WebLyzard																					
Total	**29**	**18**	**25**	**25**	**26**	**4**	**19**	**19**	**15**	**14**	**6**	**7**	**5**	**12**	**5**	**6**	**2**	**2**	**1**	**2**	**7**

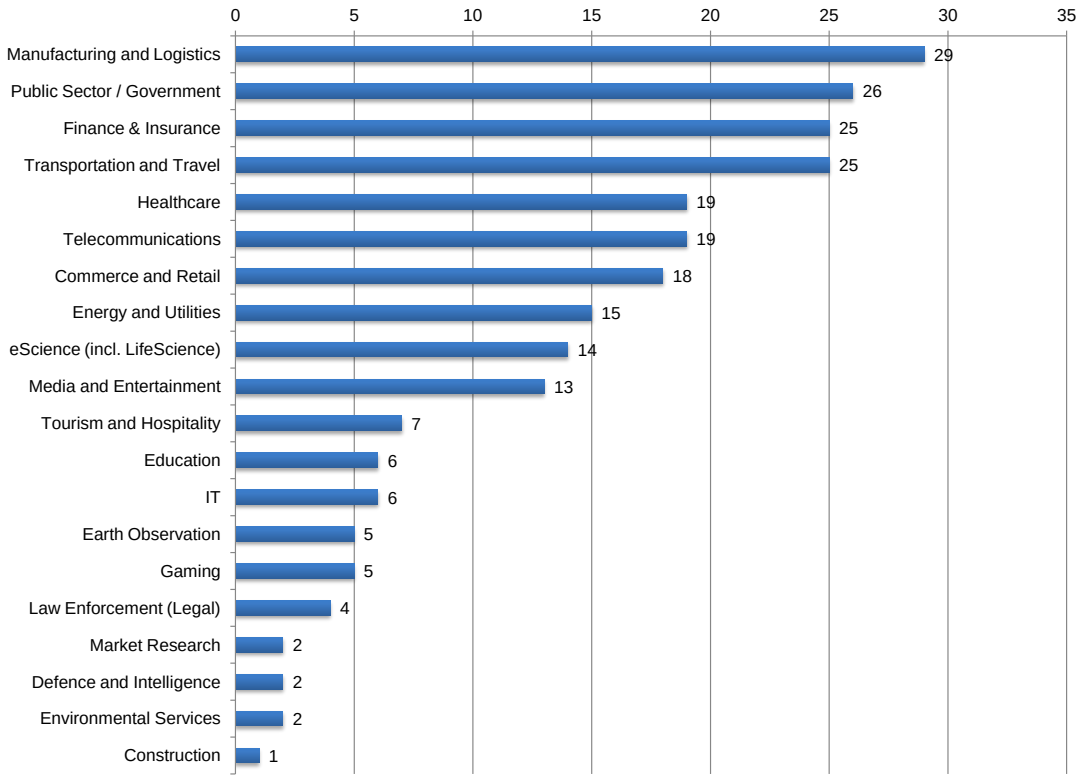


Figure 6.1: *Distribution of service providers across application domains.*

Chapter 7

Intelligent Data Analytics Roadmap

In this chapter we present the technology roadmap for Intelligent Data Analytics (Daten durchdringen – Intelligente Systeme) for Austria, more specifically for the FFG funding programme *ICT of the Future* (IKT der Zukunft). It builds upon an extensive literature review and interactions with stakeholders through an online survey, discussions in workshops and structured expert interviews. This technology roadmap brings together the perspectives of technology and industry stakeholders in this area, identifies the requirements for new ICT in this area, and prepares for the expected developments, requirements and guidelines in the ICT field. The technology roadmap covers three areas – *Technology* , *Coordination* and *Human Resources* – which influence, rely on and cross-fertilize each other. For example, the development of lead technologies in the field of Intelligent Data Analytics requires highly qualified personnel, which is short in supply. Training qualified personnel requires access to realistic data, which can be achieved through improved coordination of the relevant stakeholders in Austria.

Figure 7.1 provides an overview of the roadmap. It spans across three periods, namely short term (up to 2015), mid term (up to 2020) and long term (up to 2025). On the left, the nine roadmap objectives – grouped into the three areas – are given. The first four objectives address technological topics focusing on:

1. advancing the current data integration and data fusion capabilities,
2. increasing algorithmic efficiency,
3. turning raw data into actionable information, and
4. automating knowledge workers' processes.

These technological objectives require dedicated R&D funding, which will, in the mid to long term future, result in novel, Austrian-made lead technologies in the area of Intelligent Data Analytics. The next three objectives focus on coordination measures aiming at supporting the stakeholders' capabilities to innovate and extend their competitive position. Achieving these objectives will improve Austria's visibility, integration, and attractivity in the international ICT research and development context. These coordination-oriented objectives require investment from the public sector and industry in order to build an Austrian Data-Services Ecosystem. The Ecosystem will make data accessible and interoperable in order to generate greater economic value. Further objectives involve the elaboration of a legal framework for dealing with data,

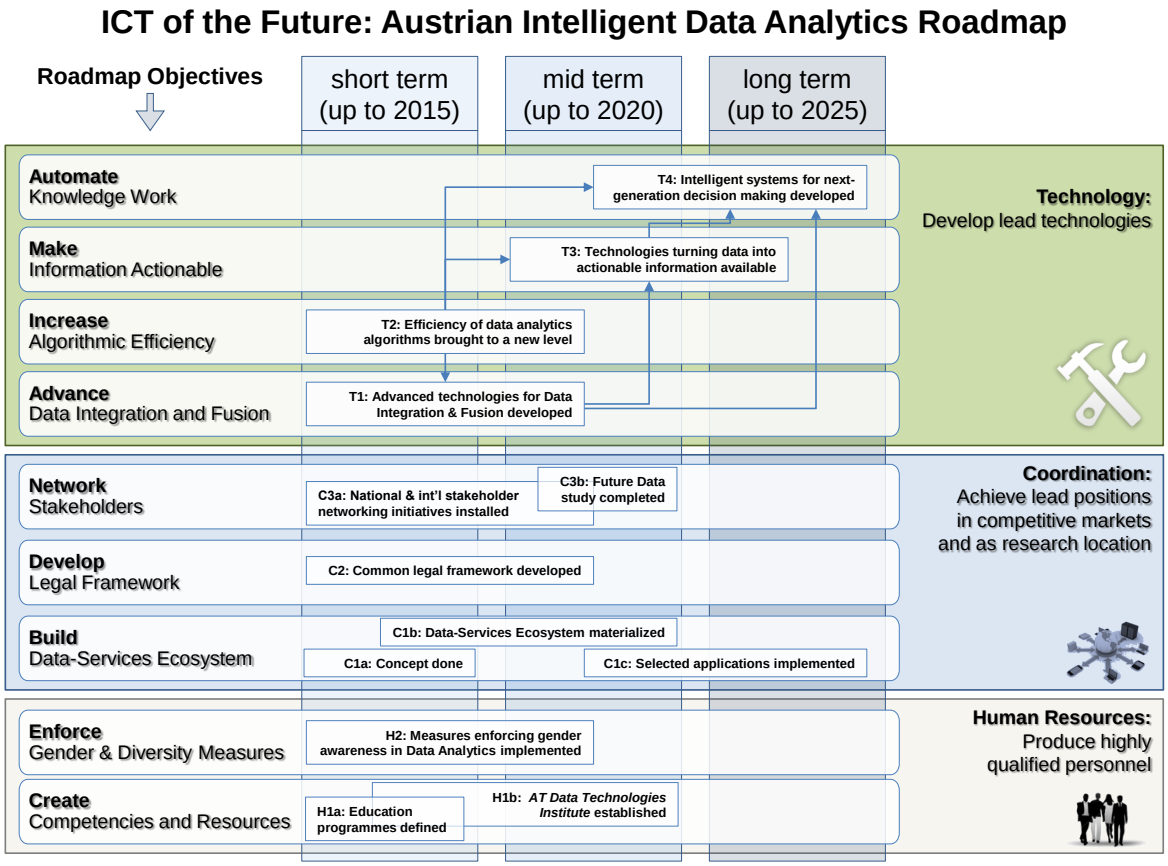


Figure 7.1: *Intelligent Data Analytics Roadmap.*

and the launch of various initiatives — including a dedicated "Austrian Data Technologies Institute" — which will strengthen the networking of and know-how exchange between Austrian and international stakeholders in the field. The remaining two objectives cover the area of Human Resources and call for addressing the urgent need for highly qualified personnel in data technologies. They advocate public investment in novel education programmes that assist in creating polymath thinkers capable of coping with the requirements emerging from (Big) Data Analytics. Improving the gender and diversity awareness in the field of Intelligent Data Analytics is the second of these two objectives.

For each objective, one or more actions are given (see Figure 7.1). Each action is enclosed by means of a rectangle and numbered from [T,C,H]1 to [T,C,H]4 with *T* referring to *Technology*, *C* referring to *Coordination* and *H* referring to *Human Resources*. The rectangles mark the preferable starting period and the estimated duration within which the respective objective shall be achieved. Note that the identification of objectives, the elaborated sequence, priorities and dependencies of actions are based on the results of the online survey, the workshop input and the expert interviews.

The technology roadmap specifically addresses the overarching goals of the ICT of the Future programme which are summarised in the following:

1. **Develop lead technologies:** Increase both the quantity and the quality of ICT-research and development that can achieve and sustain technological leadership; Enable the exploration of new ICT research topics and application fields;

2. **Achieve lead positions in competitive markets:** Strengthen the capability of firms to innovate, support firms in establishing and extending their competitive position;

3. **Establish and extend a lead position as a location for research:** Secure and improve Austria's visibility, interlinkedness and attractivity in the international context in the area of ICT research and development;

4. **Produce highly qualified personnel:** Train and attract lead researchers; improve the availability of a sufficient number of trained researchers as the backbone of excellent ICT-research and development.

The remainder of this chapter describes a catalogue of short-, medium- and long-term objectives and details actions to achieve these objectives in order to strengthen Austria's innovation capacities in the Intelligent Data Analytics area. Furthermore, we present potential lighthouse projects as a route to address some of these objectives.

7.1 Technology

Four technology areas in which research and development should be concentrated are proposed.

7.1.1 Advance Data Integration and Data Fusion Technologies

When humans consume information, heterogeneity is comfortably tolerated and errors are largely compensated. Machine analysis algorithms expect homogeneous data, and hardly understand nuance and fail with erroneous data. Thus, increasing or at least maintaining data quality is key; careful integration of (often) multimodal data is essential. Besides data quality, data integration and fusion is driven by a wide range of requirements emerging from aspects such as the user, context, privacy, languages, modality, etc. It is important that expert knowledge is accurately modeled in order to assure that data remains correct and trustworthy when integrating with other sources. Methods that have been applied on data need to be traceable; integrated results have to be reproducible. Potentially erroneous data sets (e.g. text documents) need to be preprocessed and filtered with reliable data processing technologies.

Justifying the quality of results of an integration or fusion process is paramount. This requires novel and reliable evaluation approaches that can be used to assess data quality even if the ground truth is incomplete. Advanced data integration and data fusion technologies need to account for contextual aspects (e.g. individual, temporal or geographical) in order to ensure high-quality record linkage. This implies that also the procedures for data collection have to be re-thought and the inclusion of monitors or sensors for recording the context while the data was generated is key. Therefore, it is important to include this task as part of the context model and attach it to data (task-oriented data integration). Novel approaches considering such contextual information shall therefore not only match the data itself but also consider policies or domain information attached to data. However, short texts such as in social media channels often provide only very limited contextual information, which in turn challenges effective data integration methods.

The increasing amount of data and data generating devices will produce inconceivable data volumes. However, a substantial amount of the generated data will simply be copies of data. Experts see major challenges in identifying and removing duplicate data when merging this multitude of data sources. The more noise taints the data the harder it gets to derive real insights and actionable information from data. The trend towards obtaining actionable information requires novel strategies for integrating or fusing data from different modalities such as text, images, audio, video, etc., in order to provide holistic and synchronised data on which future predictive analytics technologies shall operate. Furthermore, the multi-language nature of data needs to be considered. Eventually, this will lead to higher data quality and more powerful technologies for data integration and fusion.

Privacy plays an important role and it is rather challenging to assure anonymity when combining data from many sources (k-anonymity). Blurring of data (e.g. in the medical sector) or the removal of contextual data (e.g. in healthcare) is often practiced, but formally not well understood and not very standardised and requires more sophisticated and reliable methods for pseudonymisation.

The roadmap objective *Advance Data Integration and Data Fusion Technologies* calls for immediate research and development of novel approaches that can cope with these requirements and will eventually produce lead technologies that advance the field and improve data quality. Since heterogeneous data is an intrinsic issue of all data-related applications, technological advancements in this field will impact all application domains. For this technological objective researchers and research-oriented companies are the intended recipients of funds provided by the public sector (bmvit) and/or industry. Setting immediate actions will ensure that novel technologies for Data Integration and Data Fusion will become available in the medium term. From an engineering and research perspective addressing this objective requires cross-disciplinary expertise ranging from *Search and Analysis* and *Semantic Processing*, over *Data Representation* and *Algorithmic Efficiency* to *Evaluation*. To push the research and development of technologies that advance Data Integration and Data Fusion in the medium term, the following aspects and actions shall be considered:

- Fund research and development of quality-preserving, quality-enhancing approaches for fusing and/or integrating data originating from many sources with different modalities (e.g. audio, video, text, sensors...) and/or different media types and/or different languages. Research de-duplication approaches.

- Launch research projects specifically focusing on context-aware (e.g. individual, temporal or geographical) record linkage for fusing and/or integrating data.

- Funding of research and development of approaches for fusion and/or integration of data that can cope with very large amounts of data.

- Fund projects focusing on future data representation and management technologies, e.g. in-memory databases, NoSQL etc.

- Fund research and development of approaches that assure anonymity when crossing data from many sources including approaches for anonymisation, pseudonymisation, and the like.

- Engage with open (government) data providers in order to provide researchers and developers with access to real data linked to real problems (optimally through a Data-Services Ecosystem, cf. Section 7.2.1).

This objective responds to the following challenge: Data Representation (Section 5.1).

7.1.2 Increase the Efficiency of Data Analytics Algorithms

It is no longer sufficient to rely on processors getting faster according to Moore's law to cope with increasing volumes of data — data volume is now scaling faster than compute resources [17]. Due to power constraints, processor clock speeds are no longer increasing significantly, but processors are being built with increasing numbers of cores. *Parallel data processing techniques* are needed, but techniques developed for inter-node parallelism don't directly apply for intra-node parallelism due to the different architecture. Parallelism for multiple nodes each with multiple cores adds another layer of complexity.

Efficient algorithms decrease not only the processing requirements — investing in efficient algorithms saves costs and makes computing greener. When applying adequate measures, efficiency of data analytics algorithms can be brought to a new level (see Figure 7.1, *Increase Algorithmic Efficiency*) in the medium term. The roadmap objective *Increase the Efficiency of Data Analytics Algorithms* requires immediate research and development initiatives enabling the development of green lead technologies that provide innovative solutions to algorithmic problems and increase algorithmic efficiency. Efficient algorithms have the potential to boost all data-related applications. Thus, technological advancements in this field will impact all application domains. For this technological objective researchers and research-oriented companies are the intended recipients of funds provided by the public sector (bmvit) and/or industry. From an engineering and research perspective addressing this objective will contribute and advance all areas of the programme ICT of the Future ranging from *Search and Analysis* over *Semantic Processing*, to *Data Representation*. Additionally, *Evaluation* is an important constituent of this field since it allows to benchmark novel approaches, to estimate the suitability of a solution for a specific problem and aids in selecting the best solution for a task. To advance the research and development of green and highly efficient algorithms in the medium term, the following actions shall be considered:

- Funding of research and development of highly efficient and power-saving algorithms in the areas of Search and Analysis, Semantic Processing, Cognitive Systems, and Visualisation and Interaction.
- Launch research projects that approach algorithmic challenges in Intelligent Data Analytics from new angles, specifically with approaches such as parallel programming, high-performance computing, quantum computing, multi-core architectures and the like.
- Fund projects focusing on algorithms for real-time processing.
- Funding of projects researching new models for distributed processing of large data sets.
- Fund the research and development of approaches for evaluation and benchmarking of algorithms and install a standardised and open benchmarking environment (optimally through the Data-Services Ecosystem, cf. Section 7.2.1).

- Fund the development of guidelines for matching optimal but efficient algorithms to problems and make them available to research and industry (make it accessible via a public information exchange platform, cf. Section 7.2.3, Intelligent Data Analytics Web Platform).
- Establish incentives for efficient coding, e.g. a Green Computing Prize.
- Consolidate available infrastructures and computing resources, install novel computing architectures and provide access to research and industry (optimally through the Data-Services Ecosystem, cf. Section 7.2.1).

This objective responds to the following challenge: Techniques, Methods and Algorithms, in particular scale challenge described in Section 5.2.1.

The above actions should take advantage of synergies with the ICT of the Future horizontal topic on "Using resources responsibly and in a sustainable manner."

7.1.3 Make Information Actionable

Data overwhelms us. Thus, eliminating the noise and extracting the "valuable matter" from data is paramount. This "valuable matter" can be used to make specific (business) decisions and transforms pure data into valuable, actionable information and boosts companies' competitive advantage. Such information has to be accurate, timely, comprehensive and predictive. It needs to be analytical to stimulate exploratory thinking and provide context to ensure strategic alignment and direction. Information becomes actionable when it is understood in context. Understanding how information items relate to others is critical and in order to identify such relationships frameworks for integrating concepts are required. Consequently, making information actionable requires to develop taxonomies defining a common language for concepts, business terms, and information objects, etc. [46].

The roadmap objective *Make Information Actionable* calls for immediate research and development of novel approaches that can cope with these requirements and will eventually produce novel predictive technologies. The availability of efficient algorithms and expertise in this area (cf. Section 7.1.2) as well as the advances in governance, integration and fusion of data (cf. Section 7.1.1) facilitate the development of versatile approaches for transforming pure data into valuable, actionable information in the mid to long term. Extracting the "valuable matter" from data in order to make grounded and informed decisions is paramount to all data-related applications. Technological advancements in this field will thus impact all application domains. For this technological objective researchers at universities, research centers and research-oriented companies are the intended recipients of funds provided by the public sector (bmvit) and/or industry. From an engineering and research perspective addressing this objective requires cross-disciplinary expertise ranging from *Search and Analysis* and *Semantic Processing*, over *Data Representation* and *Visualisation* to *Evaluation* and will build upon the research done for boosting algorithmic efficiency and achievements in integrating and fusion data. To this end, the following actions shall be taken into account:

- Fund research and development of powerful predictive analytics technologies that perform real-time reasoning based on and analysis of current and historical facts in order to make predictions about future, or otherwise unknown, events [140]. Fund and launch research that combines innovative approaches from several areas including statistics, modeling, visualisation, machine learning, data mining, etc.

- Fund research of technologies for personalised, context aware human-data and/or machine-data interaction. Develop innovative interaction paradigms for data exploration incl. virtual environments, virtual reality, sonification, etc. in order facilitate the development of new data experiences.
- Funding of the development of standards for languages, language resources and core components of Intelligent Data Analytics technologies to facilitate interoperability and reusability with the ultimate objective to arrive at "standardised search and analytics" interfaces and modules.
- Fund research of multimodal and/or multilingual language resources such as ontologies, taxonomies or thesauri. Fund and launch research and development projects to automatically build and populate language resources.
- Fund projects for large-scale semantic enrichment and interlinking of multi-modal, multi-lingual data (incl. crowd sourcing).
- Fund projects to make information actionable through cross-language and cross-media data analytics incl. machine translation.

This objective responds to the following challenge: Techniques, Methods and Algorithms (Section 5.2). It also allows Austria to build on the areas of research strength identified in the workshops.

7.1.4 Automate Knowledge Work

According to McKinsey [92], advances in artificial intelligence, machine learning, natural language processing and natural user interfaces (e.g., voice recognition) are making it possible to automate many knowledge worker tasks, i.e. intelligent systems can perform knowledge work tasks involving unstructured commands and subtle judgments. This opens up possibilities for sweeping change in how knowledge work is organised and performed. Sophisticated analytics tools can be used to augment the talents of highly skilled employees, and as more knowledge worker tasks can be done by a machine, it is also possible that some types of jobs could become fully automated. However, due to the complexity of many tasks in knowledge work, it is unlikely that advances in technology will make knowledge work positions completely redundant. It is more likely to create a demand for workers with new skills who can perform new kinds of tasks. Through application of knowledge work automation technologies, significant benefits in terms of increased efficiency and hence reduced costs can be obtained in areas including healthcare, drug discovery, E-discovery (in the legal domain), recruiting and venture capital investing [90].

The roadmap objective *Automate Knowledge Work* will take advantage of the progress already made in achieving objectives 1–3. In the medium- to long-term funding research and development following the outlined direction will produce novel systems that will assist in solving complex knowledge work tasks in an automated fashion. Supporting knowledge work is relevant to all data-related applications. Technological advancements in this field will impact all application domains but is of particular importance for knowledge-driven domains such as healthcare, eScience, etc. For this technological objective researchers at universities, research centers and research-oriented companies are the intended recipients of funds provided by the public sector (bmvit) and/or industry. From an engineering and research perspective

addressing this objective requires cross-disciplinary expertise ranging from *Search and Analysis* and *Semantic Processing*, over *Data Representation* and *Visualisation* to *Evaluation* and will build upon the research done for boosting algorithmic efficiency and achievements in integrating and fusion data. To this end, the following actions shall be considered:

- Fund research into automation of knowledge work and the technologies needed for this. The research should take advantage of advancements already made in achieving Objectives 1–3.
- Fund research into the abstraction of knowledge work tasks. The results of this research should simplify the development of optimal technological support for knowledge work tasks through a general classification by the types of knowledge work tasks. Roughly seen, this should lead to the "toolbox and guidelines" for Intelligent Data Analytics algorithms often requested in the workshops.
- Fund research to ensure that the requirements of the end user are taken fully into account in the design of knowledge work support systems. This should be done in synergy with the ICT of the Future horizontal topic on human-centered computing.

This objective responds to the following challenge: Techniques, Methods and Algorithms (Section 5.2). It also allows Austria to build on the areas of research strength identified in the workshops.

7.2 Coordination

Austria has a strong research community in all areas necessary for conquering data. Nevertheless, there is currently fragmentation within this community and within the infrastructure used by this community, which should be overcome by improved coordination at the community and infrastructure levels.

7.2.1 Build a Data-Services Ecosystem for Austria

Currently, a large number of information sources remain mostly unused. The roadmap objective *Build a Data-Services Ecosystem for Austria* will make this data accessible and interoperable and assist in generating greater economic value. While it will encourage further *open data* to be made available, it will go beyond open data by also allowing *closed data* to be shared in a controlled way, potentially even generating revenues for the owners of the closed data through use of their data. Trends towards data markets have the potential to produce completely new ecosystems that enable dealing (in the broadest sense of the word) of/with data — data will become a commodity that is exchanged, auctioned, sold and bought. Data and service providers and a wider data service market is currently emerging in Europe and globally. Directing public funds into this direction in the short to medium-term will ensure that Austria participates in this global trend. In fact, this requires the involvement and active participation of all stakeholders including industry, governmental and non-governmental organisations, etc. Advancements in this field will impact all application domains but is of particular importance for knowledge-driven domains such as healthcare, eScience, digital humanities, etc.

The future data challenges shall be addressed within a robust governance framework to protect privacy. There is thus an opportunity for the industry to become a more strategically

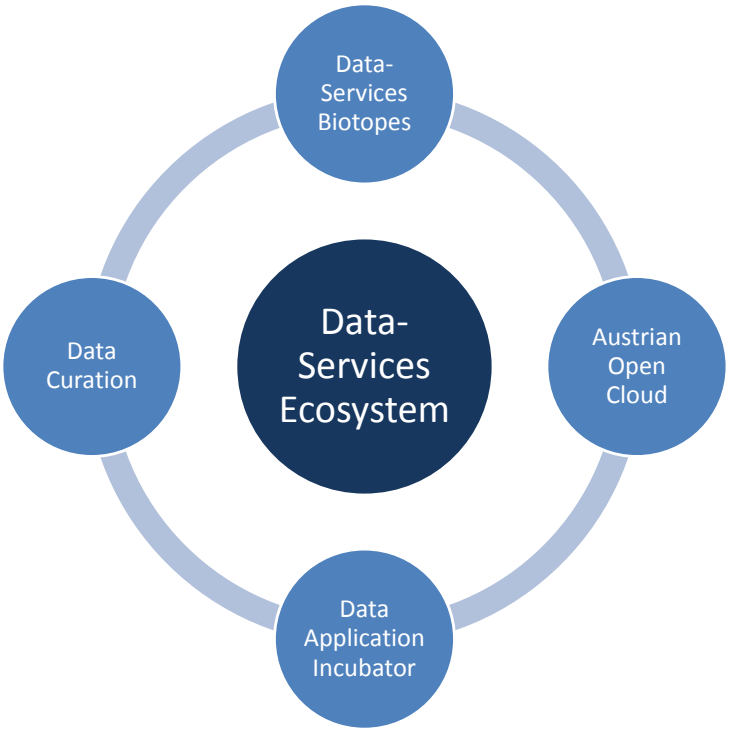


Figure 7.2: *Main components of the Data-Services Ecosystem.*

engaged customer of data services and to help build a stronger collaborative culture – one that involves stakeholders aligning around the common aim of delivering better value for everybody. Stakeholders are not necessarily experts in the area of data management implying that this is a primary task of the Data-Services Ecosystem. Providing services for data provisioning, hosting, access, curation, citation, billing and licensing is thus essential. Figure 7.2 shows the main components of an Austrian Data-Services Ecosystem. The *Data-Services Biotopes* allow the controlled sharing of data and processing services on the data within specific application domains, with interactions between the Biotopes explicitly encouraged. The *Austrian Open Cloud* provides the substrate on which the Ecosystem runs. The *Data Application Incubator* provides funding and assistance for start-ups creating data-centred applications, while *Data Curation* services are offered to ensure that data remains available and usable over long time periods. Each of these components is described in more detail in Section 8.3, which describes the Lighthouse Project to implement this Data-Services Ecosystem. To create the Austrian Data-Services Ecosystem, the following actions shall be considered:

- Fund a study project aiming at the concept development for an Austrian Data-Services Ecosystem. Ensure that all stakeholders are involved from the very beginning; examine the current situation in Austria, Germany and the EU (cf. Figure 7.1, *C1a: Concept done*).

- Launch measures (e.g., roadshows, workshops) to educate and to encourage data owners requiring solutions to make their data and corresponding problem descriptions available to Intelligent Data Analytics scientists (cf. Figure 7.1, *C1a: Concept done*).

- Subsequently fund the implementation of the Austrian *Data-Services Ecosystem* lighthouse project described in Section 8.3 (cf. Figure 7.1, *C1b: Data-Services Ecosystem materialized*).
- Furthermore, fund application-specific lighthouse projects (see Section 8.4) that shall be attached to the Austrian Data-Services Ecosystem (cf. Figure 7.1, *C1c: Selected applications implemented*).

This objective responds to the following challenges: Data Economy and Open Data (Section 5.6); Austrian Shared Computing Infrastructure (Section 5.8); Data Curation and Preservation (Section 5.7); Evaluation (Section 5.3) and some of the Societal and Economic Challenges (Section 5.11). This Ecosystem should also simplify and thereby encourage the take-up and further use of open data in Austria, and facilitate the creation of innovative start-ups using this data. It hence responds to the poor open data uptake weakness identified in the workshops. The application-specific lighthouse projects combined with the Data Application Incubator should provide more continuity from basic through applied research to start-up support, thereby overcoming a further weakness identified in the workshops.

7.2.2 Develop Legal Framework and Technological Framework Controls

Much data that could be useful in creating innovative applications is not (easily) reusable, and a legal framework allowing more straightforward specification of conditions under which data can be shared is required. In fact, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations are less forceful. The absence of a well-defined legal framework for data and in particular the privacy issues related to data are huge concerns [17].

Providing a solid legal framework and the necessary regulations gives stakeholders the certainty to operate within the legal boundaries and stimulate innovation. To instantiate this framework, a technological system that ensures that the access to and privacy of the data satisfies the legal requirements is needed. This system should also allow access to subsets of data if the data provider wishes to provide different levels of access to different end users to stay within the legal framework. An idea that circulated at the workshops was to make "Österreich zur Schweiz der Daten[1]". This would give Austria a unique position in this very sensitive context.

A basis for a common legal framework can be developed within the medium term (see Figure 7.1, *Develop Legal Framework*) if the following interdisciplinary aspects and recommendations are considered. Ensuring a solid legal framework for data will on the one hand assist in achieving lead positions in competitive markets and, on the other hand, raise Austria's attractiveness as a location for research and investment. Furthermore, the availability of a solid legal framework will positively impact the take-up of technologies in all application domains.

To improve the legal situation with respect to data and to develop a draft of rules and regulations for data use in the medium term, the following actions shall be considered. These actions are linked to the roadmap item *C2: Common legal framework developed.*

[1] Austria to become the Switzerland of Data

- Fund investigative projects/studies answering questions, e.g. what does the government know about its citizens; what are the potentials of turning general prohibition for data collection into a right to use data with care; what are the directions and impulses within the EC, etc. Identify the top legal issues, involved regulations, etc. in Austria. Detailed investigations of topics such as Privacy, Security, Data ownership, Compliance issues, Service Levels, Reliability, Indemnification and Limitations of Liabilities and how they are handled in Austria, in other countries and at EU level.

- Fund interdisciplinary projects in order to elaborate and draft a legal framework for data issues in Austria.

- Fund the creation of the technological framework controls (optimally within the Data-Services Ecosystem, cf. Section 7.2.1) that adhere to and ensure the rules and regulations. There is a large potential for synergies with the ICT of the Future topic on "Justifying trust: safe and secure systems", in particular in the areas of security and information security and privacy.

This objective responds to the following challenges: Privacy and Security (Section 5.4); Data Ownership (Section 5.5); and some of the Societal and Economic Challenges (Section 5.11). It also responds to a weakness identified in the workshops, namely that industry and research tend to be reluctant about disclosing data. By creating a clear legal and technical framework, data exchange and the release of open data should be encouraged.

7.2.3 Network Stakeholders

The linkage and knowledge exchange between Austrian stakeholders in the field of Intelligent Data Analytics is poorly developed. Proactive initiatives to network stakeholders and to grow awareness across stakeholders will assist future engagement between differing organisations and increase the trust in data technologies. Austrian companies are lacking information about legal matters, and competences of Austrian research organisations and universities and, on the other hand, academia is missing information about available data, service providers, hardware etc. In summary, it is fairly intransparent which hardware and human resources, data, etc. are available for or work in the field of Intelligent Data Analytics in Austria. Additionally, it is also frequently the case that (research) results simply end up in drawers.

Bringing together experts from different disciplines, e.g. from computer science, mathematics, legal, economic, ethics, public authorities, etc. will improve the information exchange between Austrian stakeholders. Thus, experts call for institutionalised collaboration and networking platforms. Supported through events and conferences, such platforms (or clusters) shall become the key driver for successful innovation in the area of Intelligent Data Analytics in Austria. An open and searchable directory covering the assets of Austrian stakeholders from research and industry will increase transparency, assist in establishing future collaborations and support knowledge exchange.

The following actions aim at addressing the lack of linkage and know-how exchange between stakeholders in the short to medium term. These will assist in establishing a lively and strong Austrian community in data technologies. In the end, linking national and international stakeholders and increasing the awareness and understanding of data technologies is critical to stimulate public and private investment as well as the understanding of the potential that these technologies have. Except for the last action, all others are associated with the roadmap

item *C3a: National & int'l stakeholder networking initiatives installed.* The last action is linked to the roadmap item *C3b: Future Data study completed.*

- Bundle and institutionalise marketing, collaboration and networking activities of stakeholders in the field of Intelligent Data Analytics via the "Austrian Data Technologies Institute (ADTI)" (cf. Section 7.3.1).

- Implement a stakeholders competences directory and establish an *Intelligent Data Analytics Web Platform*[2] for information exchange of stakeholders in the area of data technologies. Provide a searchable directory of the competences of data providers, service providers, research and academia. Offer a registration facility for additional stakeholders. Provide guidance and advice on legal topics, data governance, frameworks, transparency and openness. Give pointers for users to national and international (re)sources of data, metadata and data services. Promote the Intelligent Data Analytics Web Platform via the "Austrian Data Technologies Institute (ADTI)".

- Provide funds to ensure regular dialogue between stakeholders to maintain awareness of innovations. Thus, coordinate initiatives for promoting awareness, collaboration and knowledge exchange among all stakeholders. Install a regular conference or workshops involving industry, government, academia, research funders, and the public.

- Interact with the public to ensure balanced treatment on the outcomes and impact of Intelligent Data Analytics. In a first step, reach a broader audience by explaining the results of this study by means of explanation videos. Finance the development of infographics in order to convey study results quickly and clearly. Make the outcomes publicly available via the Intelligent Data Analytics Web Platform (see above).

- Fund workgroups focusing on the development of concepts for supporting (e.g. Intellectual Property, productisation) of company start-ups in the area of data technologies. Establish incubators and boot camps (as part of the "Austrian Data Technologies Institute (ADTI)"). Strengthen the collaboration with early stage investors (e.g. AWS Gruenderfonds[3], Crowd Funding[4]) such that the transition from research to data technology products becomes seamless.

- Join forces on an international level. Launch a collaboration (e.g., workgroup) between Germany and Austria (in a first step between the German BMWi and bmvit) and compare the results of related studies. Derive points of departure for joint future initiatives (in particular in the light of H2020). Evaluate the involvement in the Open Government Partnership[5] (OGP) – a multilateral initiative that aims to secure concrete commitments from governments to promote transparency, empower citizens, fight corruption, and harness new technologies to strengthen governance. Consider hosting events such as the European Data Forum[6] in Austria.

[2] e.g. http://www.conqueringdata.at/

[3] http://www.gruenderfonds.at/

[4] For example platforms such as 1000x1000 https://1000x1000.at/ or more research oriented platforms like http://www.inject-power.at

[5] http://www.opengovpartnership.org

[6] http://data-forum.eu/

- Fund a follow-up study to this roadmap study in the time approaching 2020 in order to elicit the progress made by then and to align time frames and recommendations with the actual social, economic and technological advances.

This objective responds to the overall challenge of ensuring that the many players in the Austrian Intelligent Data Analytics Competence Landscape interact with each other as effectively as possible in achieving all objectives. It also responds to a weakness identified in the workshops, namely that the Austrian stakeholders working in areas related to Conquering Data are poorly networked. The interaction with the public responds to a weakness identified in the workshops, namely that the Austrian public tends to be unaware of the benefits of data-based innovation and negatively biased to such innovation by fear of data misuse.

7.3 Human Resources

Austria is facing a shortage of highly qualified people necessary to successfully implement the required measures. It is therefore recommended to create these human resources and competences through educational measures at all levels, from schools through universities and universities of applied sciences to companies.

7.3.1 Create Competences and Resources

Qualified personnel being capable of coping with the challenges of Intelligent Data Analytics is scarce and thus very precious; and it will become even more precious in the short to medium term — in Austria and all over the world. The lack of qualified staff that can ask the proper questions, produce the necessary (statistical) models, make use of the available tools and has the capability to communicate the insights properly imposes a major challenge to the industry.

The new breed of skilled personnel equally familiar with science and advanced computing is being voiced to be the most required and valuable asset in Data Analytics. Such personnel would also be expected to have deep analytical talent, master techniques and theories from many fields, including mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualisation, uncertainty modeling, data warehousing, and high performance computing. It is about producing polymath thinkers rather than experts with a narrow skill set. These high-level and multi-layered expectations on a professional group encounter a gap of highly qualified personnel noticed currently but also predicted for the future.

It is important to perform awareness-raising actions at an early stage in order to educate children to become mature and literate in data issues. Computer science education at schools has to be compulsory and intensified. Lectures imparting digital skills and media competence have to become a foundation of today's schools' curricula. Graduates need to have gained an understanding of the importance of data and the implications of its collection and use.

On the other hand, innovative concepts and tools for highly efficient learning are required in order to impart maximum knowledge transfer in minimum time. A promising concept is, for instance, the emerging e-learning trend of Massive Open Online Courses (MOOC).

The following actions aim at addressing these personnel shortages by investing public funds into activities improving formal education and training with holistic approaches. In the end, these contribute to the production of highly qualified personnel in the data technology area. Apart from the last action, all others are associated with the roadmap item *H1a:*

Education programmes defined. The last action is linked to *H1b: AT Data Technologies Institute established.*

- Launch studies, workshops, working groups or task forces in order to define the future higher-education programmes. Foster engagement of the industry, higher education institutes and research councils to adjust or expand relevant MSc and PhD programmes, accredited, undergraduate courses and industry placements. By doing so, specific requirements of the industry become visible for university and – vice versa – decision makers in industry learn about academic formation processes, which enhances their understanding of the profoundness of studies in informatics. Early co-operation between universities and companies allows for attracting high potentials by means of internships, master and PhD theses. This, in turn, shortens the recruiting process of companies and facilitates their access to qualified personnel. Integrate gender measures in formation and training (cf. Section 7.3.2).

- Run high quality trainings; install (virtual) classrooms (MOOCs) for introductions to Intelligent Data Analytics, uses, and case studies of collaborative projects of stakeholders. Provide professional education to decision makers in companies in Austria on the possibilities and potentials of Intelligent Data Analytics, and to technical and research staff in companies on using the various Intelligent Data Analytics tools and methods. A tool for this could be the FFG Innovation Lectures Programme or courses at universities of applied sciences.

- Establish standards and standardised certifications, e.g., through the ECDL Foundation – the certifying authority of the leading international computer skills certification programme – ECDL / ICDL.

- Install a professorship for Intelligent Data Analytics at an Austrian university. As experts learn from experts the idea of having an emphasis on data analytics research is twofold: First, a dedicated professor will enter new fields of research and, in the long run, further research fields will be developed and explored, mostly in connection with the international scientific community. Second, this person will serve as a role model, attracting junior researchers for a career in this field. Participation in ICT Labs at master level[7] would also be possible with a dedicated professorship.

- Fund studies, workshops, working groups or task forces in order to improve data skills and data know-how in high-school education (statistics, capability to classify data) and encourage scholars to enter the relevant university programmes. This could also be done on behalf of a professorship but also with a stronger interaction between universities and high schools. Best practice examples for school labs are to be found in the EC initiative Living Schools Lab[8]. Interested young people and students from all kinds of high schools should be further connected with higher educational programmes. This smooths the inscription to ICT, respectively Data Analytics and also tests preferences and interests before entering university.

- Create and establish the "Austrian Data Technologies Institute (ADTI)" in order to bundle and institutionalise applied research, education, marketing, collaboration and

[7] http://www.masterschool.eitictlabs.eu/
[8] http://lsl.eun.org/about

networking for Intelligent Data Analytics (e.g., as a "K-Zentrum"). Collaborate with existing initiatives and organisations such as the Österreichische Computer Gesellschaft[9] (OCG) or digital networked data[10]. Foster collaboration between academia, research funders and technology firms. Develop skilled personnel by running high-quality trainings in the field of data technologies at the "Austrian Data Technologies Institute (ADTI)". Collaborate with industry, universities of applied sciences and universities. Conduct applied research projects in this area. As a guideline for implementation, we refer to the *Smart Data Innovation Lab*[11] launched 2014 in Germany. Furthermore, the ADTI may act as the organisational and coordinating body of the Data-Services Ecosystem.

This objective responds to the following challenge: Qualified Personnel (Section 5.9).

7.3.2 Enforce Gender and Diversity Measures

To address shortages in personnel we have learned that we need analytical leaders, content experts and business analysts at the same time. Moreover, they should be communicative, have an ability to connect on an emotional level and be able to adapt to diverse organisational and industrial cultures. As mentioned before, this holds true for men and women. Although short-term gaps in human resources may be coped with through brain circulation activities, the question remains how to attract more women to informatics and especially the profession of data scientist.

From an economic point of view, innovation is only possible through the qualification and support of excellent researchers, regardless of them being male or female, black or white, old or young. This comes along with the idea for creative processes that are believed to be of higher use for an economy if diverse ideas and moral concepts are integrated in their development. For example, Rastetter argues that groups with people from diverse backgrounds, nationalities, ages and gender are much more innovative and creative than homogenous teams [116]. Furthermore, from an equal rights position, women — and members of other minority groups – have the same rights as men to enter a career and to successfully develop it [142].

Gendered projects strengthen the awareness of project leaders, enhance discussions between gender experts and decision makers so that the relevance of gender in projects is realized. In the long run, gender expertise will help research and industry identify new markets and develop new technologies as new and innovative ideas are able to meet the needs of complex and diverse user groups [52].

It is thus necessary to highlight gender as relevant as it is in future projects and explicitly raise women to top positions in order to enhance pre-gendering of data analytics through diversity. Considering gender issues and the specific ideas and needs of a minority group in data analytics will, in turn, also be of use for people with diverse social and geographical background, age, family status, education and values [142].

The following actions aim at addressing shortcomings in the gender and diversity situation in Austria. In the end, enforcing these measures will produce a diverse range of highly qualified personnel in the data technology area. On the other hand, a stronger consideration of diversity (in gender, but also in backgrounds, competences and interests) will be more able to meet research and industry's requirements on data scientists and the diverse application fields of

[9] http://www.ocg.at/
[10] http://networkeddata.at/
[11] http://www.sdil.de/de/

this professional group. These actions are linked to the roadmap item *H2: Measures enforcing gender awareness in Data Analytics implemented.*

- Fund and launch workgroups that evaluate new directions for teaching and research. This involves, for example, holistic research designs combining real use cases with theoretical backgrounds. A stronger project and problem-based learning environment at universities will lead to a better professional understanding of a data scientist. This also holds true for a stronger variation of teaching, including different materials, approaches and interactions in every class to address different learning styles. To explicitly address gender issues in a technical male-dominated area, international universities include compulsory lectures on gender and diversity in the standard curriculum[12]. Introductory phase and introductory courses at universities should explicitly address the mentioned needs by offering classes not only in the sector of programming, data mining etc., which enforce mathematical-technical competences, but should additionally enforce communication and language skills. The Carnegie Mellon University may serve as an international reference case as they enhanced the number of female students by changing the requirements for studying informatics in this way: programming skills were no longer required and communication skills were asked for. Women enter universities with specific prior knowledge. In Austria, girls more often attend a humanistic secondary education while boys attend technical colleges [123]. By offering, for example, programming classes for those with lower skills would not only support female students, but would also serve as a measure to generally make the university curricula more permeable, allowing people from different backgrounds and with different competence profiles to associate themselves from the beginning on.

- Increase the visibility of outstanding research performances of women. An excellent example to achieve this, is the Laura Bassi Initiative [33]. Eight research centres in different research areas which were explicitly to be led by excellent female researchers were started throughout Austrian federal states in the last decade. The initiative promotes women who do excellent research instead of promoting women because of being a women. In their teams, men and women receive positions according to their own scientific reputation and a formal selection process.

- Quotas: The concept of female quotas is, however, often criticised from both men and women [142, 45, 97]. Women do personally not appreciate approaches that are based on their alleged deficits, as they prefer to make their life and their career on their own. Nevertheless, national (cf. [142]) and international studies (cf. [87]) show a boosting effect for women if they once received a management position in a project. This even holds true if the promotion of the female was driven by external reasons such as a financial bonus. This means that quotes are still a reliable and effective measure to enhance female opportunities in male-dominated areas and should be pursued in the future.

- Prevent marginalisation of women by supporting (research) projects involving mixed teams.

This objective responds to the following challenge: Gender and Diversity (Section 5.10).

[12]For example see University of Bremen. Description of gender awareness initiatives can be found at http://www.gender-curricula.com/gender-curricula/gender-curricula-detailansicht/?uid=9

Chapter 8

Lighthouse Projects

As a route to implementing some of the objectives in Chapter 7, lighthouse projects are proposed. At the workshops, a table discussed propositions for lighthouse project topics, and the input received is first summarised. Subsequently, an outline of the proposed lighthouse projects is given, and some potential lighthouse project outlines are presented.

8.1 Input from the Workshops

Apart from identifying the adequate volume of research funding[1] in Austria, the primary question is whether to focus on initiatives having a *broad impact* or a *deep impact*. Initiatives expected to deliver a broad impact are centered on infrastructures or platforms while the funding of highly specialised projects tends to generate fairly deep impact (excellence clusters). In the end, it is essential to raise the awareness of the general public about the importance and value of data as well as on handling of data. Above all, the experts argued that one of the most important aspects to address is to provide a solid and reliable legal framework, in which future activities (both commercial and research) can trust and on which they can reliably build upon.

When taking into account the Austrian expertise in the field of Intelligent Data Analytics experts proposed to focus on a few, specialised application and research domains. These were in particular quantum physics, life sciences and healthcare, digital humanities, and next-generation production paradigms (Industry 4.0).

On the other hand, experts suggested investing in Austrian infrastructure or platform initiatives, which provide data sovereignty, guarantee legal certainty and have the potential to raise the awareness about data and data handling in Austria and beyond. Initiatives falling into this category may materialise as data marketplaces or as platforms that increase transparency and enable interdisciplinary knowledge exchange.

In particular, the stakeholders proposed an eScience platform that provides management and access to publications, research and results, the underlying data, as well as the parameters, processes, configurations, settings, software, i.e. the context that was used to generate the results. In the first place, this approach will allow for better reproducibility of research results

[1] The volume of public funding was generally regarded as too little. Moreover, the practice of obtaining public funding was criticised in so far that i) funding has to be deliberately split into non-overlapping projects (projectified) and ii) the majority of projects does not receive funding; thus, a large amount of time of Austrian researchers is wasted.

and experiments. This applies not only to academia but also to commercial sectors such as the pharma industry in order to guarantee the reproducibility of clinical trials. Additionally, this eScience platform will help to bridge the gap between actual (business) needs and (researched) technologies addressing these needs. Furthermore, it will foster the communication and interaction between different disciplines (e.g. information science, biology, legal, mathematics, etc). Above all, this platform supports the goals of public interest economics and has the potential to strengthen Austria's innovation output.

Another infrastructure idea is driven by the concept of data markets, i.e. an infrastructure that provides the legal, financial and technological framework for securely and reliably dealing with data and associated services that operate on data. Such data markets shall provide the necessary means for contracting, licensing and payment. Moreover, this infrastructure may host data challenges (cf. Kaggle). It will bring together problem owners with solution providers and can become the Austrian/German/Swiss platform for crowd-sourced predictive analytics and facilitate the identification and mediation of experts.

8.2 Outline of Proposed Lighthouse Projects

As a result of the discussions at the workshops, the necessity for an innovation-facilitating structure such as an eScience infrastructure or data market in Austria became apparent. We therefore recommend a broad impact lighthouse to create a *Data-Services Ecosystem* in Austria, seen in the centre of Figure 8.1 and described in Section 8.3. The *Data Application Incubator* component of this Ecosystem will facilitate access to data and appropriate algorithms and methods for start-ups and SMEs, resulting in the creation of multiple *Data-Centred Applications*.

However, a key role of the Data-Services Ecosystem is to facilitate research, innovation and technology transfer in the Intelligent Data Analytics domain. It can also be seen as an enabler for application-specific lighthouse projects, as each lighthouse project would make use of the Data-Services Ecosystem to share data, algorithms, methods and solutions. The role of the application-specific lighthouses would be to channel the work on achieving the recommendations into application areas that are of specific interest and importance in Austria. The Data-Services Ecosystem implemented by the Broad Impact Lighthouse would encourage and ensure cross-fertilisation of work between the application areas, avoiding the development of application-specific solutions in "silos."

Four potential application-specific lighthouse projects are shown in Figure 8.1, illustrating application domains in which such lighthouse projects would create impact in Austria. It makes more sense to focus such lighthouse projects by application area rather than by recommendation, as it is easier to interest the relevant industrial stakeholders in such focussed projects in their area of interest. Multiple parallel lighthouse projects would increase the amount of cross-fertilisation of Intelligent Data Analytics algorithms, methods and approaches between application areas through the Data-Services Ecosystem. The choice of application area is based on the survey responses described at the beginning of Chapter 4. The application domain seen as important by the highest number of respondents is *healthcare*. Also within the top four application domains are *manufacturing* and *energy*, which are already identified as important application areas in the ICT of the Future programme. The eScience domain is also seen as important, and could be taken as being covered by the Data-Services Ecosystem. However, we present an example of a lighthouse focussing on Digital Humanities, an area in

which there is particular scientific promise in Austria due to the large amount of relevant data available, and which has a potential commercial application in tourism.

As an example of how such application-specific lighthouse projects would be structured and their relation to the Data-Services Ecosystem, we have chosen, in order to avoid too much repetition, to present the general outline of such a lighthouse in Section 8.4. This is followed by the outlines of the instantiation of such lighthouse projects for two domains of different scale and complexity. The *healthcare and life sciences* domain (Section 8.5) is an example requiring a large-scale lighthouse involving a broad range of stakeholders, while the *Digital Humanities* lighthouse (Section 8.6) is at a smaller scale, with more focus on SMEs and start-ups.

The general relation between the recommendations and the lighthouse projects is shown in Figure 8.2. The Broad Impact lighthouse would focus effort on implementing the recommendations on building a Data-Services Ecosystem and developing a corresponding legal framework. All application-specific lighthouses deal principally with developing lead technologies through advancing data integration and fusion, increasing algorithmic efficiency, making information actionable and automating knowledge work within a specific application area. As an example in the figure, the courses of two potential application-specific lighthouses are shown, one on manufacturing and one on healthcare and life sciences. In practice, many such lighthouses can run in parallel.

The Data-Services Ecosystem Lighthouse could best be implemented as a planning project followed by a large-scale implementation project. The application-specific lighthouses could be implemented in at least three ways: (i) use existing project structures such as K-Projects or Christian Doppler Laboratories; (ii) create new large-scale project structures within the ICT of the Future programme; or (iii) reserve parts of the funding in the normal ICT of the Future calls for research and development within the application-specific lighthouse topics

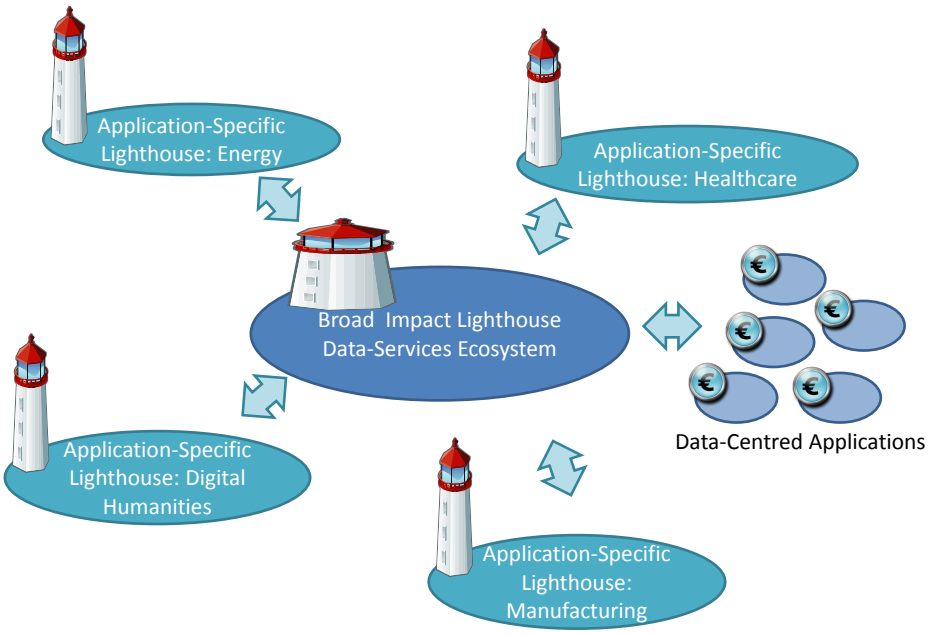


Figure 8.1: *Overview of the proposed lighthouse projects.*

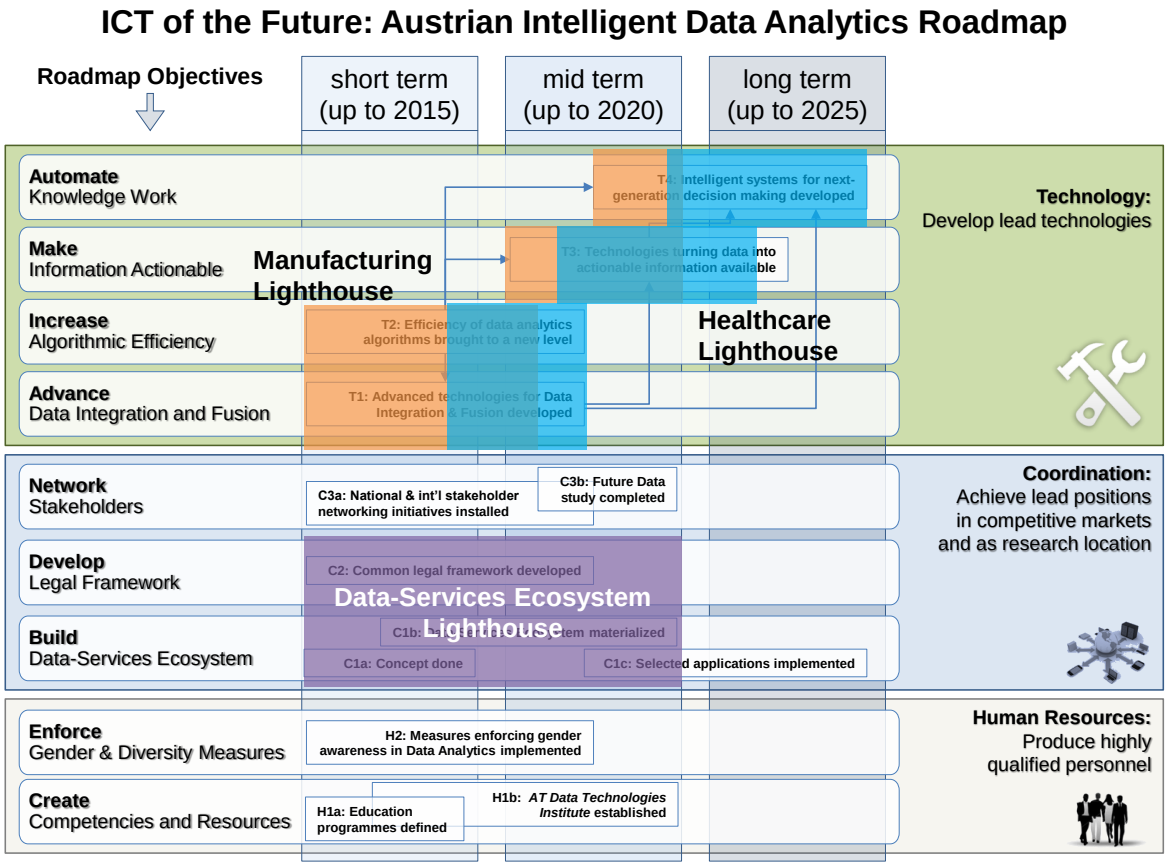


Figure 8.2: *The relation of the lighthouse projects to the roadmap. The Data-Services Ecosystem lighthouse is shown in purple. Two potential application-specific lighthouses are also shown in blue and orange.*

(as is currently done in the ICT of the Future programme for the topics "ICT for intelligent energy networks and systems" and "ICT-supported manufacturing systems").

While it would be possible to carry out each application-specific lighthouse project without having the Data-Services Ecosystem in place, this would be inefficient, as sub-parts of the Data-Services Ecosystem would have to be recreated within each application-specific lighthouse project to make them possible (e.g. an experimental platform to allow evaluation of multiple analysis approaches on large datasets would have to be created in each project). Also, the cross-fertilisation of Intelligent Data Analytics algorithms, methods and approaches between application areas would be difficult without the Data-Services Ecosystem.

8.3 Broad Impact Lighthouse: Data-Services Ecosystem

A Data-Services Ecosystem Lighthouse Project will achieve the following:

- Build an **Austrian Data-Services Ecosystem** consisting of:
  - **Data-Services Biotopes**, through which the sharing of data and services is facilitated

  - **Austrian Open Cloud**, on which the Ecosystem runs
  - **Data Application Incubator**, encouraging the creation of innovative data-centred start-ups
  - **Data Curation**, ensuring that research and other data remains available and usable over long timespans

- Create a framework for resolving **legal**, **privacy** and **security** issues related to data.

The **Data-Services Ecosystem** is an innovation-facilitating structure at national level. Part of the value provided would be similar to that provided by existing national open data organisations, such as the Open Data Institute[2] in the UK as well as the European level Data Incubator project currently open for proposals[3]. These assist start-ups and SMEs in gaining access to open data as well as proprietary data, and also provide funding for start-ups and SMEs developing data-centred applications. However, the Data-Services Ecosystem would go beyond this, as it will also focus on bringing together algorithms, tools, and methods for intelligent data analytics, and on providing data curation and preservation for researchers in Austria, ensuring that the results of publicly-funded research projects continue to be available. This will result in multiple synergies:

- The industry makes closed data and associated problems available in a controlled way for Intelligent Data Analytics researchers to work on, and to receive solutions to their problems;
- Intelligent Data Analytics researchers have access to multiple sources of open and closed data on which to evaluate tools, also encouraging the reuse of Intelligent Data Analytics solutions across multiple domains;
- SMEs and start-ups have access to data and Intelligent Data Analytics tools for creating innovative applications, with all data access issues and revenue flows regulated within the Ecosystem;
- Researchers have an eScience infrastructure on which to share and preserve their data and associated code, hence encouraging reproducibility of scientific results and sustainability of publicly-funded research.

This Data-Services Ecosystem will encourage the development of *sustainable collaborations* between domain experts and data owners (from industry and academia) and researchers in the Intelligent Data Analytics area; the development of an *ecosystem of analytical tools and research practices* that is sustainable, reusable, extensible, learnable and straightforward to translate across application areas; and the connection of the ecosystem to the *education of Data Scientists* to ensure early experience with real-life challenges. As a comparison, an initiative to maximise the impact of data science on academic research in the USA is the $37.8 million partnership between New York University, the University of California, Berkeley and the University of Washington with support from the Gordon and Betty Moore Foundation and Alfred P. Sloan Foundation[4]. This has the aim to: (1) increase interactions

[2] http://theodi.org

[3] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/87-ict-15-2014.html

[4] http://www.sloan.org/fileadmin/media/files/press_releases/datascience.pdf

and collaborations between data scientists and researchers in subjects collecting large amounts of data; (2) establish sustainable career paths and alternative metrics for data scientists; and (3) work towards a sustainable ecosystem of analytical tools and research practices.

As it is not effective to populate the Ecosystem with data and algorithms in all application domains in an uncoordinated way, the Ecosystem is made up of **Data-Services Biotopes**. These Biotopes can be populated with data and algorithms through focussed actions, such as the *application-specific lighthouses* described later in this chapter. This would result in functioning biotopes in specific domains, which can also interact to ensure reuse of tools across multiple application domains. A prototype for a Data-Services Biotope is CloudEO[5], which works in the earth observation domain. It allows earth observation data and processing tools for this data to be uploaded, and makes these available for use in application development by companies, with part of the revenue stream flowing to the algorithm and data providers once the application generates revenue. Furthermore, it serves as a data incubator (through cooperation with the European Space Agency Business Incubation Centres), but does not make research and eScience activities possible.

Setting up this Ecosystem does not necessarily require building and maintaining a data centre or other infrastructure. As an alternative, a cloud service provider could be used. This has the advantage that the costs are proportional to the storage and computing time used, and hence will start off low at the beginning and increase as the Ecosystem is accepted and adopted. Another alternative is to work on unifying much of the currently rather fragmented Austrian research computing infrastructure under a single national umbrella organisation, such as the **Austrian Open Cloud**, that could provide the Data-Services Ecosystem using existing Austrian infrastructure. An interesting model for providing such a cloud at national level is the Massachusetts Open Cloud[6] (MOC) [11], currently being set up in the State of Massachusetts, USA. The MOC enables multiple entities to provide computing resources and services, and its design allows a range of heterogeneous system architectures to be included in the cloud. Each entity providing a computing resource or service will be responsible for its operation and for determining the rate that users are charged. The MOC is responsible for the shared services of the cloud, and for exposing and collecting charges (along with a small overhead to pay for MOC operations). The MOC is funded by the state, private industry and five universities: Boston University, Harvard, MIT, Northeastern University and the University of Massachusetts[7]. Start-ups can also provide hardware or software services through the MOC and derive revenue from this.

An outline of a promising Data-Services Ecosystem for Austria is shown in Figure 8.3. The main stakeholders are in the corners of the diagram, and are described below:

Data-Centred Sciences and Humanities: Researchers with backgrounds in specific subjects (such as quantum physics, astrophysics, genetics, economics, sociology) and extensive experimental data, as well as researchers working in areas of the humanities dealing with large amounts of data (digital humanities).

Industry and Public Organisations: Industrial and public organisations that have large amounts of data (both open and proprietary) and challenges and problems associated with the data that they require to be solved by Intelligent Data Analytics approaches.

[5] http://www.cloudeo-ag.com
[6] http://www.bu.edu/cci/moc/
[7] http://gcn.com/Articles/2013/12/16/Massachusetts-open-cloud.aspx

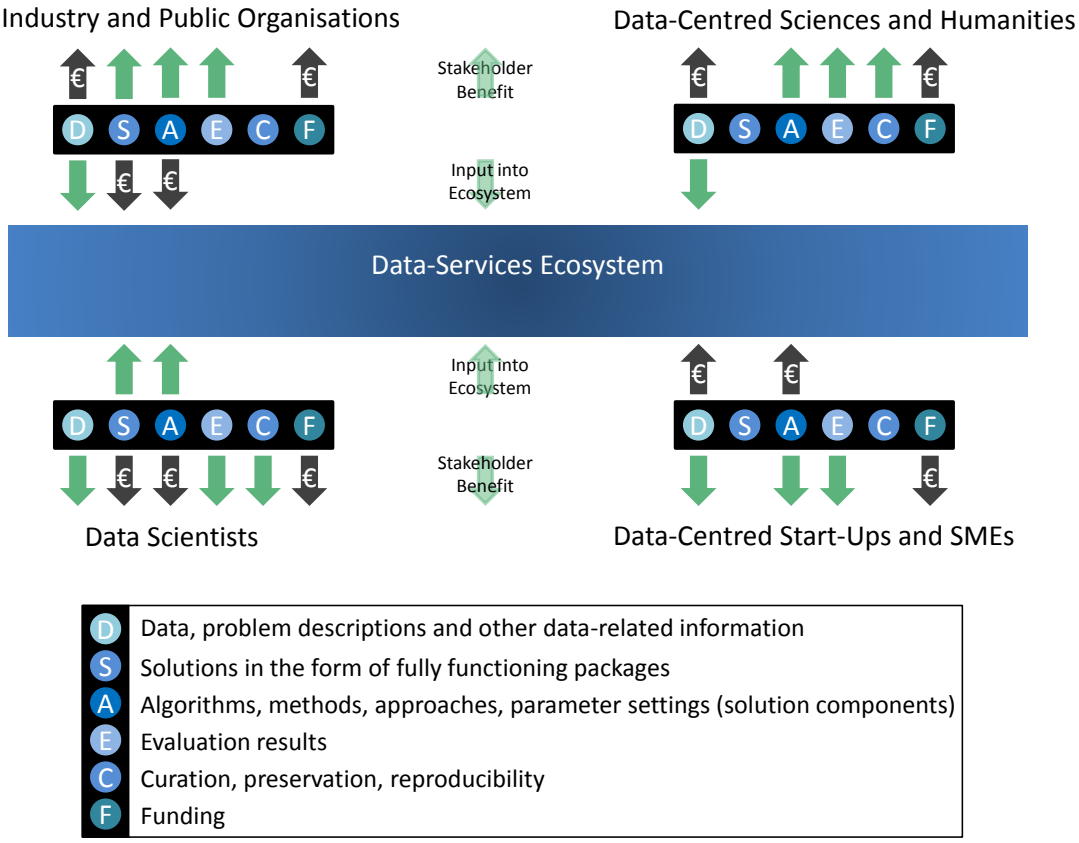


Figure 8.3: *Overview of the Data-Services Ecosystem and its stakeholders. For each stakeholder group, the inputs into the Ecosystem and the benefits obtained from the Ecosystem are shown by arrows. Arrows containing a Euro sign indicate a financial input or benefit, while other types of input or benefit are shown by a green arrow. The types of inputs and benefits are summarised in the table below the figure.*

Data Scientists: Researchers in the area of Intelligent Data Analytics and its related areas (computer science, applied mathematics, statistics). This also includes researchers from specific subjects such as physics and genetics that have specialised in data analysis.

Data-Centred Start-Ups and SMEs: Innovative companies that have the capability of placing data and intelligent data analytics algorithms in innovative applications that can generate revenue.

For each of the stakeholders shown in Figure 8.3, the inputs into the Data-Services Ecosystem and the benefits obtained from the Ecosystem are shown, where the types of inputs and benefits are listed below the diagram. We now describe some of the cycles within this Ecosystem.

First, this ecosystem will serve to bring together the *Data-Centred Sciences and Humanities* stakeholders with *Data Scientists*. This cooperation will have the specific aim of recognizing what it takes to move each of the sciences forward, through using novel analytics techniques on the scientific data to produce new insights in the corresponding scientific fields, and pushing the development of new analytics techniques and the extensive evaluation of existing

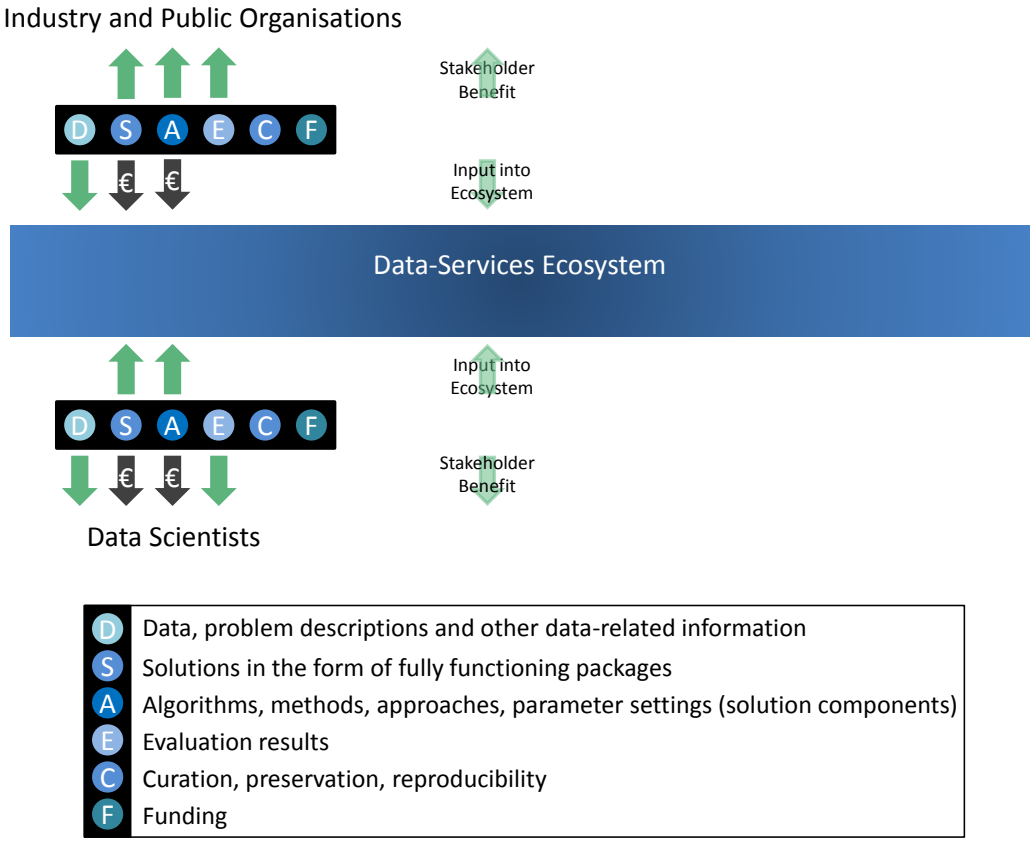


Figure 8.4: *The scenario of industry paying Data Scientists for solutions to their problems on their data.*

techniques on extensive data from multiple fields. On the Intelligent Data Analytics side, this extensive experimental and evaluation capability will lead to the creation of reusable, extensible analytical tools, evidence-based usage guidelines for the tools and research practices that can be applied across research areas. The curation and preservation Ecosystem services will ensure that data and experiments (especially those supported by public funding) remain available, and will also increase reproducibility of experimental results. Public funding for such projects will also be available through the Ecosystem, but the funding will be well targeted due to the clear overview of what already exists and what is still needed in the Ecosystem.

Second, *Industry and Public Organisations* will be able to provide problems and associated data (both open and proprietary) for *Data Scientists* to solve. The industry pays the Data Scientists providing the best solution for the solution and associated algorithms, as illustrated in Figure 8.4. This could be, for simpler problems, through a competition in which the winning solution gets prize money. For more complex problems, the solution could be created within a project in which the company pays one or more research groups to develop the solution, potentially also with additional public funding through the Ecosystem. This concept of users of the Ecosystem funding the Ecosystem is a key to the sustainability of the Ecosystem, and the most effective ways of doing this remain to be investigated. Further sustainability options such as payments for Ecosystem use from funded research projects should also be investigated.

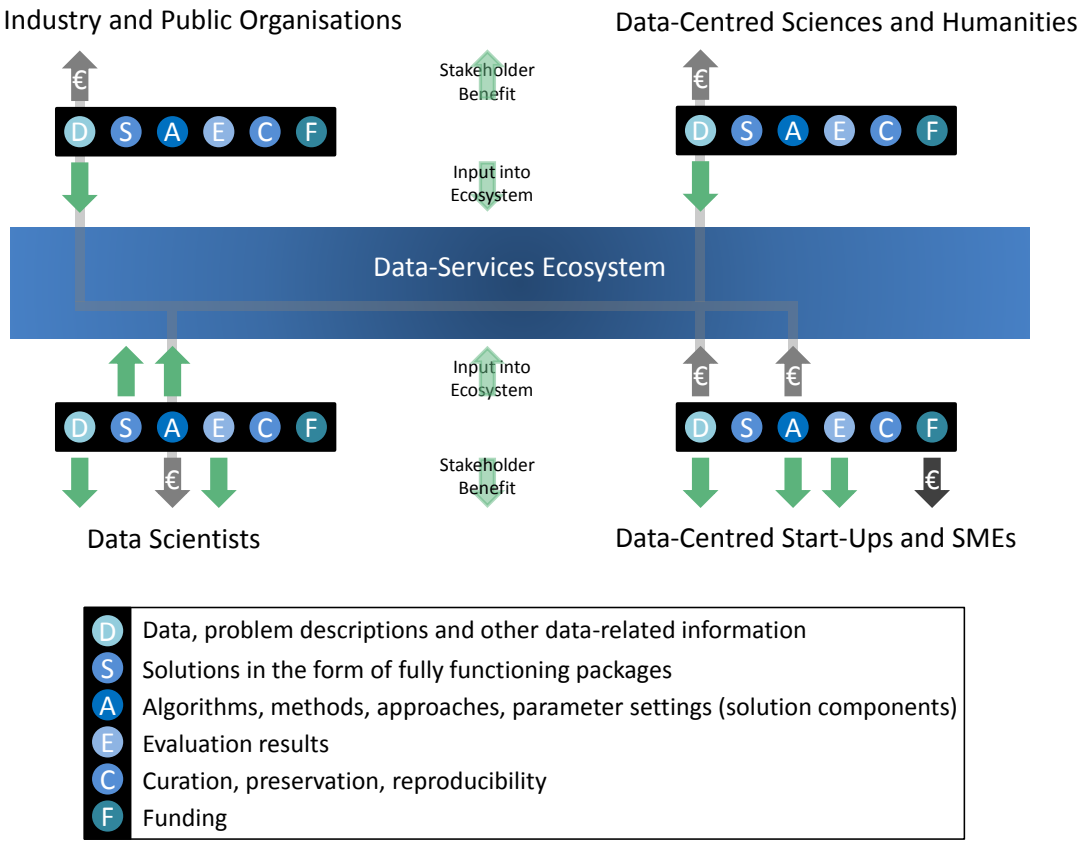


Figure 8.5: *The scenario of a start-up that makes use of data from Science and Industry and algorithms from a research organisation to create an application. The lighter grey arrows connected by lines demonstrate the flow of remuneration for the data and algorithms from the application income.*

A **Data Application Incubator** is a very promising option for sustainability. Through the data incubator, *Data-Centred Start-Ups and SMEs* are able to release data-centred applications using data in the Ecosystem as commercial services. Through the analytical tools available in the Ecosystem, these data entrepreneurs will have access to tools and associated evaluation information to guide the choice of tools. Once the applications are released, part of the income from these applications then flows to the Ecosystem and, depending on the usage agreements, to providers of adopted analytical tools and data providers. The latter should be an encouragement to data providers to make data available, while a potential income through reuse of provided algorithms is an encouragement for Data Scientists. The financial component of this Ecosystem cycle is shown by the lighter gray arrows connected by lines in Figure 8.5. As an example, imagine a start-up that creates a planning application for farmers that integrates weather data from the Central Institution for Meteorology and Geodynamics (ZAMG) with satellite data from the European Space Agency (ESA). Data analysis algorithms from the University of Vienna are combined with visual analytics algorithms from the VRVis research centre inside a user interface developed by the start-up. A part of the revenue generated by the application flows to the data providers and algorithm providers, as well as to the

ecosystem for hosting the service. The start-up also receives seed funding from the Ecosystem. As an example of the impact that a Data Application Incubator can have, the UK Open Data Institute, in it's first year of existence, supported 12 start-ups, where these start-ups generated over €800,000 in commercial contracts and €800,000 in investment.

As pointed out in Chapter 5, **Data Curation** is of the utmost importance to ensure that data continues to be usable over long time spans. This has been recognised for scientific data, and the number of countries with institutions handling research data is increasing, ranging from those offering data curation services (e.g. Data Archiving and Networked Services[8] (DANS) in The Netherlands) to those assisting others in managing their research data (e.g. Australian National Data Service[9] (ANDS) in Australia). Internationally, the Research Data Alliance[10] (RDA) aims to accelerate and facilitate research data sharing and exchange. An example of a well used eScience service is *myExperiment*,[11] which allows users to find, use and share scientific workflows and other research objects — it currently has over 7,500 members and 2,500 workflows. However, commercial data must often also be preserved, and a service to be offered by the Data-Services Ecosystem could well be consulting on data curation and preservation for vital commercial data.

Initially, this lighthouse project should focus on data that can be shared without legal constraints. However, from the beginning, the legal framework for data sharing and ownership should be carefully crafted. This legal framework should serve to encourage increased sharing of data by industry and government, with clear rules on what can and cannot be done with the data, and a straightforward way to assemble an understandable data usage agreement when uploading data. Once the data has been uploaded under certain conditions, then the Ecosystem should ensure that users of the data adhere to these conditions through the implementation of security measures. As more experience is gained, the legal framework and associated security measures should be extended to handle more sensitive data (such as medical and pharmaceutical data).

This lighthouse project directly addresses the challenges on a shared computing infrastructure, data curation and preservation, data economy, data ownership and open data, and privacy and security. It provides the Ecosystem for evaluation of techniques on multiple data sets, hence leading to improvements in data representation, data fusion, data integration and Intelligent Data Analytics techniques and algorithms. The Ecosystem facilitates cooperation between multiple application domains and Intelligent Data Analytics researchers, as well as assists in the education of qualified personnel.

8.4 General Design of an Application-Specific Lighthouse

This section presents a general template for an Application-Specific Lighthouse and its relation to the Data-Services Ecosystem. Such a template can be used as a starting point for the design of lighthouses working in specific application areas (e.g. manufacturing, energy, healthcare, digital humanities, ...). Essentially, an application-specific lighthouse serves to populate a *Data-Services Biotope* within the *Data-Services Ecosystem* in a coordinated way. Two examples

[8] http://www.dans.knaw.nl/en
[9] http://www.ands.org.au
[10] https://rd-alliance.org/
[11] http://www.myexperiment.org/

of the instantiation of the design of Application-Specific lighthouses are given in the following two sub-sections.

A general application-specific lighthouse will achieve the following goals:

- **Develop lead technologies** for data integration and fusion, algorithmic efficiency, actionable information and decision support (Objectives 1–4) in the chosen application area;
- Implement the necessary interfaces to the Data-Services Ecosystem to **allow sharing of data with Intelligent Data Analytics researchers**. Ensure that data owners requiring solutions see the advantages of the Data-Services Ecosystem;
- Clear up any Application-Specific **legal, ethical or privacy issues** within the Data-Services Ecosystem;
- **Develop Intelligent Data Analytics expertise** in the chosen application area by giving students the opportunity to work on real data within the Data-Services Ecosystem;
- Encourage the **creation of start-ups providing data-centred applications** on the Data-Services Ecosystem.

8.5 Application-Specific Lighthouse: Life Sciences and Healthcare

This is an example of an Application-Specific Lighthouse project. The theme of life sciences and healthcare was chosen due to the strong position that Austria currently occupies in this area, the top position of this application area in the survey results, as well as the immense challenges still open in this area for both medical professionals and life science researchers. Medical professionals are suffering from a dramatic information overload, and this will only get worse in the future. This reflects the scale, complexity, and rapid rate of change of medical sciences, the ever burgeoning array of disease prevention and treatment options, plus the immense challenge of intelligent real-time utilisation of omics data (genomics, proteomics, metabolomics, ...). Life science researchers, on the other hand, need information from the clinical domain in order to be able to carry out their research more effectively and focus on challenges with potentially high impact in human medicine.

The life sciences and healthcare lighthouse will achieve the following goals:

- **Develop lead technologies** for data integration and fusion, algorithmic efficiency, actionable information and decision support (Objectives 1–4) in the healthcare and life sciences area;
- Implement the necessary extensions to the Data-Services Infrastructure in the area of life sciences and healthcare to **facilitate the reuse of life science data together with secondary use of healthcare data**, and hence encourage the movement of knowledge from the lab bench to the bedside;
- Create a framework for resolving **legal**, **ethics**, **privacy** and **security** issues related to secondary use of patient data, including ways for citizens to be more involved in the decisions regarding the reuse of their data and the outcomes of these decisions;

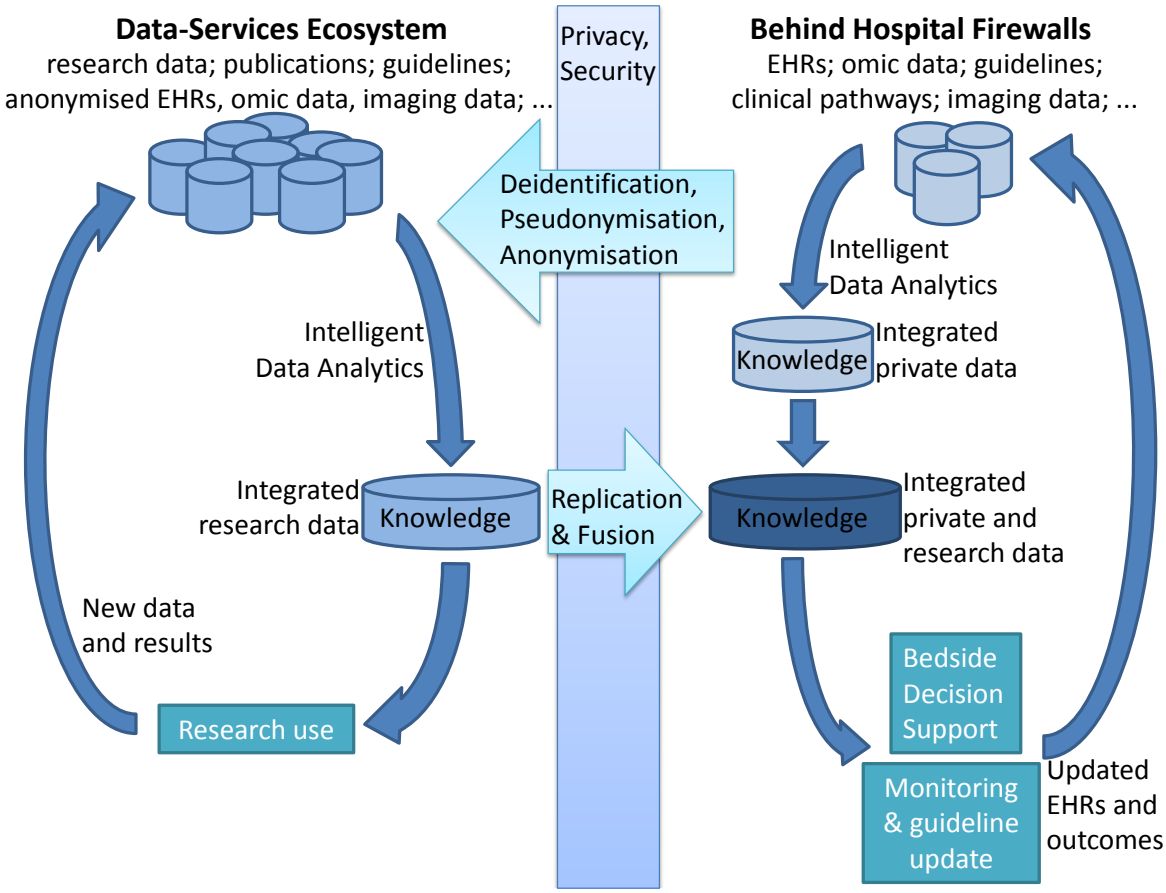


Figure 8.6: *The double cycle of health and life science information inside and outside a hospital.*

- Develop the profession of **Health Data Scientist** and the corresponding curriculum;
- **Inform the public** of research results that have an influence on their life and health.

A potential structure in the clinical domain is shown in Figure 8.6. The cycle on the left is the Data-Services Ecosystem described in Section 8.3, allowing research data, processing workflows and results to be shared among scientists. The cycle on the right takes places behind a hospital firewall (or within a social security provider), and involves the analysis of patient data to gain new insights. There are two links between the cycles, going through the hospital firewall: the lower link provides the analysis cycle within the hospital with new knowledge obtained through research in the "outside world" (Data-Services Ecosystem) that could be used to influence the hospital procedures; the upper link provides carefully controlled access to de-identified, pseudonymised or anonymised patient data, including electronic health records (EHRs), omics data and imaging data, to the external research infrastructure, allowing this data to be used to generate new knowledge through analyses outside the hospital. This infrastructure is of even more use if multiple hospitals and related organisations provide data to the Data-Services Ecosystem.

There have already been demonstrations of the advantages to be gained through mining huge numbers of anonymised medical records, such as the discovery of the dangers of the

Vioxx[12] drug by mining the Kaiser Permanente patient records [67]. The creation of the *legal and ethics framework* is an important building block in this lighthouse project, and needs to be carefully done. In the creation of this framework, it may be of use to examine the potential negative outcomes of not making patient information available for research use, caused by potentially life-saving knowledge within the patient records not being found. For example, Feufel et al. "believe that it is inefficient, and in some cases unethical, to store [population-based data] without installing mechanisms to allow access and publication in appropriate and useful ways" [56]. A recent article at The Telegraph points out the advantages of carefully sharing anonymised UK National Health Service (NHS) patient records.[13] Ideally, this framework should be created with extensive consultation with all affected citizens, where the advantages and disadvantages as well as the already gained new knowledge and potential breakthroughs in new knowledge are carefully discussed. The security controls for accessing the data and the procedures for granting access to the data will also have to be carefully designed. The task of sustainability of the data sharing activities is more challenging — pharmaceutical companies are very likely prepared to pay for access to this data, so this should be carefully considered during the creation of the legal and ethics framework. Such payment for the data could also be a welcome additional source of income for current financially stricken hospitals and social insurance organisations. Due to the sensitive nature of the data sharing, the public should be regularly informed of research results applicable to their life and health, to ensure continuing support for the use of the data in this way.

The *health data scientist* is a new profession that is currently missing in the health domain, given the huge amount of data generated in this domain, and the rather small amount of use that is currently made of this resource. A health data scientist would be a person with extensive data analytics competence, but also sufficient training in the health domain to ease communication with people working exclusively in this domain.

This lighthouse project directly addresses the challenges on a shared computing infrastructure, data curation and preservation, data ownership and open data, and privacy and security. It provides the infrastructure for evaluation of techniques on multiple data sets, hence leading to improvements in data representation, data fusion, data integration and intelligent data analytics techniques and algorithms in the health domain. Due to the sensitive nature of medical data, the data economy challenge can only be met after extensive work on the challenges related to data privacy and security.

8.6 Application-Specific Lighthouse: Digital Humanities

Austria has a huge amount of historical data, some of which is digitised or being digitised, which is spread over multiple organisations. From a research point of view within the humanities, the use of digital humanities techniques on integrated and fused data has the potential to lead to interesting research results. Research results of this type are also often easy to make understandable for the wide public. On the commercial side, this historical data has the potential to be used in tourism applications. However, the current difficulty of getting access to the widely distributed data makes it unattractive to develop such applications. This lighthouse

[12] http://en.wikipedia.org/wiki/Rofecoxib

[13] http://blogs.telegraph.co.uk/news/marthagill/100253859/the-nhs-wants-to-share-our-data-and-its-a-good-idea/

should also encourage the creation of such start-ups through simplifying the modalities for accessing and using the data.

The digital humanities lighthouse will achieve the following goals:

- **Develop lead technologies** for data integration and fusion and actionable information in the digital humanities area;
- Implement the necessary interfaces to the Data-Services Ecosystem to **allow sharing of data with Intelligent Data Analytics researchers and digital humanities researchers**. Ensure that data owners see the advantages of using the Data-Services Ecosystem to share data;
- Clear up any **legal, ethical or privacy issues** specific to the digital humanities area within the framework created in the Broad Impact Lighthouse;
- **Develop expertise** in the digital humanities area for both digital humanities researchers and intelligent data analytics researchers by giving students the opportunity to work on real data within the Data-Services Ecosystem;
- Encourage the **creation of start-ups**, particularly in the tourism area, providing data-centred applications on the Data-Services Ecosystem.

As an example, imagine an Austrian start-up that creates a "Sounds of Austria" service and associated app that allows one to listen to historical recordings based on location. For this app, data from the Austrian Academy of Sciences Phonogram Archive and Austrian Mediathek are used along with advanced approaches to sound visualisation developed at the University of Linz, all embedded in a sophisticated user interface created by the start-up and provided on the Data-Services Ecosystem. Proceeds from the sales of the app are divided among the data providers, algorithm providers, infrastructure providers and the start-up.

Chapter 9

Conclusion

This report presents a technology roadmap for the Austrian ICT of the Future research theme *Conquering Data: Intelligent Systems*. The technology roadmap covers three areas, Coordination, Technology and Human Resources and provides a catalogue of nine objectives for the short-, medium- and long-term focus of this funding programme.

Coordination: Austria has a strong research community in all areas necessary for conquering data. Nevertheless, there is currently *fragmentation* within this community and within the infrastructure used by this community, which should be overcome by improved coordination at the community and infrastructure levels.

At the community level, improved networking is needed to bring together researchers from all scientific and technological disciplines needed for effectively conquering data, but also to bring together the researchers with data owners requiring solutions to their data challenges.

The Austrian computing infrastructure is currently also fragmented, with computing infrastructure being built up in an uncoordinated way by institutions and groups of institutions. Better coordination of this infrastructure is needed to ensure the most effective use of it by the widest range of people. We furthermore propose to go beyond this and build an Austrian *Data-Services Ecosystem*. This Ecosystem functions through contributions of its users, and provides services in return. The Ecosystem is based on four ingredients:

Computing: An *Austrian Open Cloud* uniting existing Austrian computing infrastructure under a single cloud, with a lightweight cloud operator responsible for shared cloud services and exposing and collecting usage charges. This allows single-point access to the full diversity of Austrian computing infrastructure in a transparent way as an Ecosystem service.

Data: Data can be straightforwardly made available through the Ecosystem for scientific use or even for commercial use, with the possibility to charge through the Ecosystem services for commercial use. The majority of legal issues to do with data sharing are cleared up, and the Ecosystem provides services such as templates for data sharing agreements and security services for sensitive data. Assistance in obtaining data that is not yet shared and in curation of data is also provided.

Algorithms: Users of the Ecosystem can provide algorithms for use by others. The usage parameters are easily definable, and commercial use can be remunerated through the

ecosystem. This encourages comparison of the algorithms on available data sets, useful for advancing scientific knowledge but also for solving practical problems.

Incubator: Start-ups and SMEs take advantage of the Ecosystem services to create innovative data-centred applications, which generate revenue for themselves but also for the Ecosystem and those providing inputs. The provision of seed funding is a further Ecosystem service for start-ups.

A lighthouse project is proposed to take on the task of building this Data-Services Ecosystem.

Technology: The roadmap proposes four technology areas in which research and development should be concentrated. *Data Integration and Fusion* is the challenging task of bringing together heterogeneous data types from multiple sources, both within and across organisations. This is required to move beyond the current situation in which data tends to be stored in separate silos, and hence to take advantage of the full potential of the data. Due to the increasing amount of data to be processed, it is necessary to *increase the efficiency of many algorithms* so that the outcomes of analyses are available within the required time frame. Building on these areas, *information will be made actionable*, meaning that the "valuable matter" required for decisions will be extracted from the data. Due to the work on data fusion and integration, data can be analysed together to get deeper insights, while the use of interactive analyses and visual analytics goes beyond what is currently possible due to the increased algorithmic efficiency. The actionable information will then feed into a revolution in the *automation of knowledge work*, where knowledge work will not only become more efficient due to increased automation, but new types of tasks will be created.

In order to focus developments in these technology areas on application domains of importance to Austria, the use of application-specific lighthouse projects is proposed. Such projects will channel the development work towards solving challenges in a specific domain of application, while the Data-Services Ecosystem will allow cross-fertilisation of technologies between application domains. This cross-fertilisation is important as it allows problems specific to certain application domains to be abstracted to "problem types" across multiple application domains, and hence the application of solutions also in domains for which they were not initially developed. Application domains of particular interest to Austria and hence targets for application-specific lighthouses are manufacturing, energy, healthcare and digital humanities.

Human Resources: Austria is facing a shortage of highly qualified people necessary to successfully implement the required measures. It is therefore recommended to *create these human resources and competences* through educational measures at all levels, from schools through universities and universities of applied sciences to companies. A significant increase in qualified talent could be gained through implementing *gender and diversity measures* to increase the number of females choosing to study in the area of Intelligent Data Analytics.

In summary, Intelligent Data Analytics has the potential to greatly benefit the Austrian society and economy. It is essential for a successful innovation economy to provide an ecosystem in which data-centred innovation and technology transfer can take place. There are still many challenges to overcome from both a technological and societal point of view before Austria is ready to take full advantage of this opportunity.

Bibliography

[1] FORCE11 Data Citation Synthesis Group. http://www.force11.org/node/4432, last visited: September 2013. (Cited on page 48.)

[2] RDA Practical Policies Working Group. https://www.rd-alliance.org/working-groups/practical-policy-wg.html, last visited: September 2013. (Cited on page 48.)

[3] RDA Working Group on Terminology. https://rd-alliance.org/working-groups/data-foundation-and-terminology-wg.html, last visited: September 2013. (Cited on page 49.)

[4] Top Ten Big Data Security and Privacy Challenges. Cloud Security Alliance. http://www.isaca.org/Groups/Professional-English/big-data/GroupDocuments/Big_Data_Top_Ten_v1.pdf, last visited: September 2013. (Cited on page 43.)

[5] ELAK - ein europaweites E-Government Vorzeigeprojekt. APA-OTS Presse sendung OTS0100, 2004. 2004-07-28. (Cited on page 29.)

[6] Aster Data Cluster: High-Performance Analytics for Gaming. Teradata Corp., 2010. http://www.asterdata.com/resources/assets/sb_Aster_Data_4.0_Gaming_Industry.pdf, last visited: August 2013. (Cited on page 33.)

[7] Findings regarding the market events of may 6, 2010. U.S. Securities and Exchange Commission and the Commodity Futures Trading Commission, 2010. http://www.sec.gov/news/studies/2010/marketevents-report.pdf, last visited: December 2013. (Cited on page 30.)

[8] Riding the Wave: How Europe can gain from the rising tide of scientific data. European Commission, 2010. (Cited on pages 22 and 24.)

[9] CODATA Task Group on Digital Data Citation: Best Practices: Research & Analysis Results, 2012. http://www.codata.org/taskgroups/TGdatacitation/docs/CODATA_DDCTG_BestPracticesBib_FINAL_17June2012.pdf, last visited: September 2013. (Cited on page 48.)

[10] Reform der Verwaltungsgerichtsbarkeit geht voran. APA-OTS Presse sendung OTS0182, 2012. 2012-11-05. (Cited on page 31.)

[11] The Massachusetts Open Cloud (MOC), 2012. http://www.bu.edu/cci/files/2012/11/MOC.pdf, last visited: December 2013. (Cited on page 82.)

[12] Data protection laws of the world. DLA Piper, 2013. http://bit.ly/16GAKE1, last visited: August 2013). (Cited on page 45.)

[13] eGovernment Solutions in Austria, 2013. http://www.egov-suite.com/en/references/austria.html, last visited: November 2013. (Cited on page 29.)

[14] Österreichisches Bundesheer investiert 10 Millionen Euro in Technologie und Forschung. APA-OTS Presse sendung OTS0195, 2013. 2013-08-22. (Cited on page 33.)

[15] J. Acker. Hierarchies, jobs, bodies: A theory of gendered organizations. *Gender and Society*, 4(2):139–158, 1990. (Cited on page 52.)

[16] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica. BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data. In *Proceedings of EuroSys*, pages 29–42. ACM, 2013. (Cited on page 38.)

[17] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom. Challenges and Opportunities with Big Data, 2012. http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf, last visited: August 2013. (Cited on pages 37, 38, 44, 65, and 70.)

[18] R. Aldred. From community participation to organizational therapy? World Café and Appreciative Inquiry as research methods. *Community Development Journal*, 46(1):57–71, 2011. (Cited on page 9.)

[19] N. Anderson. The ethics of using AOL search data. Ars Technica, 2006. http://arstechnica.com/uncategorized/2006/08/7578/, last visited: August 2013. (Cited on page 46.)

[20] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 601–610. ACM, 2009. (Cited on page 42.)

[21] R. Baker and K. Yacef. The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining. *Journal of Educational Data Mining*, 1:3–17, 2009. (Cited on page 26.)

[22] A. Bankhamer. Schatzsuche 2.0, 2013. http://www.monitor.at/index.cfm/storyid/15012_Big_Data_-_der_Datengoldrausch_im_21._Jahrhundert-Schatzsuche_2.0, last visited: December 2013. (Cited on page 27.)

[23] M. Barlow. *Real-Time Big Data Analytics: Emerging Architecture*. O'Reilly, 2013. (Cited on page 38.)

[24] B. Batinic. Internetbasierte Befragungsverfahren. *Österreichische Zeitschrift für Soziologie*, 28(4):6–18, 2003. (Cited on page 9.)

[25] C. Bauckhage and K. Kersting. Data Mining and Pattern Recognition in Agriculture. *KI - Künstliche Intelligenz*, pages 1–12, 2013. (Cited on page 32.)

[26] S. R. Bird. Unsettling universities' incongruous, gendered bureaucratic structures. *Gender, Work & Organization*, 18(2):202–230, 2011. (Cited on page 10.)

[27] M. H. Birnbaum. Human research and data collection via the internet. *Annual review of Psychology*, 55:803–832, 2004. (Cited on page 9.)

[28] J. C. Blickenstaff. Women and science careers: Leaky pipeline or gender filter? *Gender and Education*, 17(4):369–386, 2005. (Cited on pages 10 and 51.)

[29] A. Bogner and W. Menz. Das theoriegenerierende Experteninterview. Erkenntnisinteresse, Wissensformen, Interaktion. In A. Bogner, B. Littig, and W. Menz, editors, *Das Experteninterview: Theorie, Methode, Anwendung*, pages 33–70. VS Verlag für Sozialwissenschaften, 2005. (Cited on page 10.)

[30] D. Boyd and K. Crawford. Critical Questions for Big Data. *Information, Communication & Society*, 15(5):662–679, 2012. (Cited on page 2.)

[31] E. Brat, S. Heydorn, M. Stover, and M. Ziegler. Big Data: The Next Big Thing For Insurers? The Boston Consulting Group Perspectives, March 2013. (Cited on page 27.)

[32] M. V. D. Brink and Y. Benschop. Gender practices in the construction of academic excellence: Sheep with five legs. *Organization*, pages 1–18, 2011. (Cited on pages 10, 51, and 52.)

[33] Bundesministerium für Wissenschaft und Forschung - BMWFJ. Laura Bassi Centres of Expertise. At the interface of science and industry, 2013. https://www.ffg.at/sites/default/files/downloads/laurabassi_brosch.pdf, last visited: December 2013. (Cited on page 76.)

[34] J. Burn-Murdoch. Data security and privacy: can we have both? The Guardian, July 2013. http://www.theguardian.com/news/datablog/2013/jul/31/data-security-privacy-can-we-have-both, last visited: September 2013. (Cited on page 43.)

[35] D. Cheney. Text mining newspapers and news content: new trends and research methodologies. In *Proc. IFLA World Library and Information Congress*, 2013. (Cited on page 24.)

[36] C.-T. Chu, S. K. Kim, Y.-A. Lin, Y. Yu, G. Bradski, A. Ng, and K. Olukotun. Map-Reduce for Machine Learning on Multicore. In *Proceedings of NIPS*, pages 281–288. MIT Press, 2006. (Cited on page 38.)

[37] P. S. Churchland, C. Koch, and T. J. Sejnowski. What is computational neuroscience? In E. L. Schwartz, editor, *Computational Neuroscience*, pages 46–55. MIT Press, 1993. (Cited on page 17.)

[38] Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council. *Frontiers in Massive Data Analysis*. The National Academies Press, 2013. (Cited on page 13.)

[39] B. Cotton. Mastering information for competitive advantage: Smarter computing in the travel and transportation industry, 2012. http://public.dhe.ibm.com/common/ssi/ecm/en/xbl03022usen/XBL03022USEN.PDF, last visited: August 2013. (Cited on page 26.)

[40] E. Czernohorszky, V. Schadauer, and S. Schinkinger. Gender monitoring bericht 2010, 2011. http://www.zit.co.at/fileadmin/user_upload/ZIT/Dienstleistungen/GenderMonitoringBericht2010.pdf, last visited: December 2013. (Cited on page 51.)

[41] Y. Dandawate, editor. *Big Data: Challenges and Opportunities*, volume 11 of *Infosys Labs Briefings*. Infosys Labs, 2013. http://www.infosys.com/infosys-labs/publications/Documents/bigdata-challenges-opportunities.pdf, last visited: August 2013. (Cited on page 1.)

[42] T. Davenport. At the big data crossroads: turning towards a smarter travel experience, 2013. http://www.amadeus.com/web/binaries/blobs/60/112/Amadeus_Big_Data.pdf, last visited: December 2013. (Cited on page 27.)

[43] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, Jan. 2008. (Cited on page 18.)

[44] N. Döring. *Sozialpsychologie des Internet. Die Bedeutung des Internet für Kommunikationsprozesse, Identitäten, soziale Beziehungen und Gruppen*. Hogrefe, Göttingen, 2nd edition, 2003. (Cited on page 9.)

[45] H. Dryburgh. Work hard, play hard: Women and professionalization in engineering—adapting to the culture. *Gender and Society*, 13(5):664–682, 1999. (Cited on page 76.)

[46] S. Earley. Transforming information into knowledge, part ii, 2012. https://www.earley.com/knowledge/articles/transforming-information-knowledge-part-ii, last visited: December 2013. (Cited on page 66.)

[47] E. H. Ecklund, A. E. Lincoln, and C. Tansey. Gender segregation in elite academic science. *Gender & Society*, 26(5):693–717, 2012. (Cited on page 10.)

[48] G. Eibl, J. Höchtl, B. Lutz, P. Parycek, S. Pawel, and H. Pirker. Rahmenbedingungen für Open Government Data Plattformen. Technical report, 2012. (Cited on page 47.)

[49] S. Ellis. Big Data and Analytics Focus in the Travel and Transportation Industry, 2012. http://h20195.www2.hp.com/V2/GetPDF.aspx%2F4AA4-3942ENW.pdf, last visited: August 2013. (Cited on page 26.)

[50] O. Etzioni, R. Tuchinda, C. A. Knoblock, and A. Yates. To buy or not to buy: mining airfare data to minimize ticket purchase price. In L. Getoor, T. E. Senator, P. Domingos, and C. Faloutsos, editors, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pages 119–128. ACM, 2003. (Cited on page 26.)

[51] European Commission. She Figures 2012. Gender in Research and Innovation, 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/she-figures-2012_en.pdf, last visited: December 2013. (Cited on page 51.)

[52] European Commission. Gendered innovations: How gender analysis contributes to research. report of the expert group 'innovation through gender', 2013. http://ec.europa.eu/information_society/newsroom/cf/horizon2020/document.cfm?doc_id=3853, last visited: December 2013. (Cited on pages 52 and 75.)

[53] W. Faulkner. The gender(s) of "real" engineers: Journey around the technical/social dualism. In P. Lucht and T. Paulitz, editors, *Recodierungen des Wissens. Stand und Perspektiven der Geschlechterforschung in Naturwissenschaften und Technik*. Campus, Frankfurt, 2008. (Cited on pages 10 and 51.)

[54] E. A. Feigenbaum and P. McCorduck. *The fifth generation*. Addison-Wesley, 1983. (Cited on page 16.)

[55] D. Feldman, C. Sung, and D. Rus. The single pixel GPS: learning big data signals from tiny coresets. In *Proceedings of SIGSPATIAL/GIS*, pages 23–32. ACM, 2012. (Cited on page 38.)

[56] M. A. Feufel, G. Antes, J. Steurer, G. Gigerenzer, J. A. M. Gray, M. Mäkelä, A. G. Mulley, Jr., D. E. Nelson, J. Schulkin, H. Schünemann, J. E. Wennberg, and C. Wild. What is Needed for Better Health Care: Better Systems, Better Patients or Both? In G. Gigerenzer and J. A. M. Gray, editors, *Better Doctors, Better Patients, Better Decisions: Envisioning Health Care 2020*, pages 117–134. MIT Press, 2011. (Cited on pages 23 and 89.)

[57] C. Fouché. An invitation to dialogue 'the world café' in social work research. *Qualitative Social Work*, 10(1):28–48, 2011. (Cited on page 9.)

[58] J. Freire and C. T. Silva. Making computations and publications reproducible with VisTrails. *Computing in Science & Engineering*, 14(4):18 –25, Aug. 2012. (Cited on page 42.)

[59] U. Froschauer and M. Lueger. *Das qualitative Interview zur Analyse sozialer Systeme*. WUV Studienbücher Sozialwissenschaften, Vienna, 2nd edition, 1998. (Cited on page 11.)

[60] S. Fuchs, J. V. Stebut, and J. Allmendinger. Gender, Science and Scientific Organisations in Germany. *Minerva*, 39:175–201, 2001. (Cited on page 10.)

[61] Futurezone. Smart-Meter-Pflicht ab Ende 2019, 2012. http://futurezone.at/science/smart-meter-pflicht-ab-ende-2019/24.579.186, last visited: December 2013. (Cited on page 35.)

[62] J. Gama. *Knowledge Discovery From Data Streams*. Chapman & Hall/CRC, 2010. (Cited on page 38.)

[63] A. F. Gilbert. Disciplinary cultures in mechanical engineering and materials science: Gendered/gendering practices? *Equal Opportunities International*, 28(1):24–35, 2009. (Cited on pages 10 and 51.)

[64] B. G. Glaser and A. L. Strauss. *The Discovery of Grounded Theory. Strategies for Qualitative Research*. Aldine, New York, 1987. (Cited on pages 10 and 11.)

[65] J. Gläser and G. Laudel. *Experteninterviews und qualitative Inhaltsanalyse*. VS Verlag für Sozialwissenschaften, Wiesbaden, 3rd edition, 2009. (Cited on page 10.)

[66] S. Gorard and T. Cook. Where does good evidence come from? *International Journal of Research and Method in Education*, 30(3):307–323, 2007. (Cited on page 7.)

[67] D. J. Graham, D. Campen, R. Hui, M. Spence, C. Cheetham, G. Levy, S. Shoor, and W. A. Ray. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *The Lancet*, 365(9458):475–481, 2005. (Cited on page 89.)

[68] R. Graham and A. Lewington. The Big Data Explosion: A New Frontier in Digital Law, 2013. http://www.scl.org/site.aspx?i=ed31114, last visited: August 2013. (Cited on page 46.)

[69] J. C. Greene and V. J. Caracelli. Advances in mixed-method evaluation. the challenges and benefits of integrating diverse paradigms. *New directions for evaluation*, 74, 1997. (Cited on page 7.)

[70] J. C. Greene, V. J. Caracelli, and W. F. Graham. Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11:255–274, 1989. (Cited on page 7.)

[71] U. Gretzel. Technology and tourism: Building competitive digital capability, 2013. http://www.tourism.australia.com/documents/Technology_and_Tourism.pdf, last visited: August 2013. (Cited on page 29.)

[72] D. J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006. (Cited on page 42.)

[73] E. Hand. Word play. *Nature*, 474:436–440, 2011. (Cited on page 24.)

[74] L. Hardesty. Ecova customers cut electric consumption intensity 8.8%, shows study, 2013. http://www.energymanagertoday.com/ecova-customers-cut-electric-consumption-intensity-8-8-shows-study-093633/, last visited: December 2013. (Cited on page 23.)

[75] J. A. Harding, M. Shahbaz, and A. Kusiak. Data Mining in Manufacturing: A Review. *J. of Manufacturing Science and Engineering*, 128, 2006. (Cited on page 25.)

[76] Harvard Business Review. Data scientist: The sexiest job of the 21st century, 2012. http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/, last visited: December 2013. (Cited on page 51.)

[77] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009. (Cited on pages 23 and 24.)

[78] B. Inc. The big data opportunity, 2012. http://tdwi.org/~/media/9836C350D9F641669563B19DD563B6E1.pdf, last visited: August 2013. (Cited on page 33.)

[79] D. C. Ince, L. Hatton, and J. Graham-Cumming. The case for open computer programs. *Nature*, 482(7386):485–488, Feb. 2012. (Cited on page 42.)

[80] S. International. Now arriving: Big data in the hospitality, travel and tourism sector, 2013. https://scpelc.egnyte.com/h-s/20130510/c61863288b5f4c68, last visited: August 2013. (Cited on page 30.)

[81] F. Jahanian. From data to knowledge to discovery, 2013. http://admin.icordi.eu/Repository/document/Presentations/RDALaunch_Presentations/FromDataToKnowledgeToDiscovery_FarnamJahanian.pdf, last visited: August 2013. (Cited on page 13.)

[82] R. King. From cars to catamarans, how big data plays in sports. http://www.zdnet.com/from-cars-to-catamarans-how-big-data-plays-in-sports_p2-7000019911/,lastvisited:September2013, 2013. (Cited on page 33.)

[83] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. P. B. Simperl, editors, *The Semantic Web: Research and Applications, 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31-June 4, 2009, Proceedings*, volume 5554 of *Lecture Notes in Computer Science*, pages 723–737. Springer, 2009. (Cited on page 32.)

[84] W. Koerbitz, I. Önder, and A. Hubmann-Haidvogel. Identifying Tourist Dispersion in Austria by Digital Footprints. In L. Cantoni and Z. P. Xiang, editors, *Information and Communication Technologies in Tourism 2013*, pages 495–506. Springer Berlin Heidelberg, 2013. (Cited on page 30.)

[85] G. Kramer, editor. *Auditory Display: Sonification, Audification, and Auditory Interfaces*, volume XVIII of *Sante Fe Institute Studies in the Sciences of Complexity*. Addison-Wesley, 1994. (Cited on page 18.)

[86] H. Krcmar, P. Wolf, M. Wolf, L. Rau, and V. Till-Stavrakakis. Government Monitor 2013: Nutzung und Akzeptanz von elektronischen Bürgerdiensten im internationalen Vergleich. Technical report, 2013. (Cited on page 29.)

[87] M. Lindberg, I. Danilda, and B.-M. Torstensson. Women resource centres - a creative knowledge environment of quadruple helix. *Journal of the Knowledge Economy*, 3(1):36–52, 2012. (Cited on page 76.)

[88] D. Loshin. Who owns data? Information Management, 2003. http://www.information-management.com/issues/20030301/6389-1.html, last visited: August 2013. (Cited on page 45.)

[89] Louisiana AgCenter. Seven big data lessons for farming, 2013. http://www.agprofessional.com/news/Seven-big-data-lessons-for-farming-223632331.html, last visited: December 2013. (Cited on page 32.)

[90] V. Mahidhar and D. Schatsky. The future of knowledge work. *Signals for Strategists*, 2013. (Cited on page 67.)

[91] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute, 2011. (Cited on pages 13, 16, 21, 22, 24, 25, 28, 30, 32, and 50.)

[92] J. Manyika, M. Chui, J. Bughin, R. Dobbs, P. Bisson, and A. Marrs. *Disruptive technologies: Advances that will transform life, business, and the global economy.* McKinsey, 2013. (Cited on page 67.)

[93] J. Manyika, M. Chui, P. Groves, D. Farrell, S. V. Kuiken, and E. A. Doshi. *Open data: Unlocking innovation and performance with liquid information.* McKinsey Global Institute, 2013. (Cited on pages 21 and 47.)

[94] C. A. Mattman. A vision for data science. *Nature*, 493:474–475, 2013. (Cited on pages 24 and 50.)

[95] P. Mayring. *Einführung in die qualitative Sozialforschung. Eine Anleitung zu qualitativem Denken.* Beltz, Weinheim/Basel, 2002. (Cited on page 11.)

[96] P. Mayring. *Qualitative Inhaltsanalyse. Grundlagen und Techniken.* Beltz, Weinheim, 9th edition, 2007. (Cited on page 11.)

[97] J. McIlwee and J. Robinson. *Women in engineering: Gender, power and workplace culture.* State University of New York Press, 1992. (Cited on page 76.)

[98] M. Mehmet and D. Wijesekera. Data analytics to detect evolving money laundering 71-78. In K. B. Laskey, I. Emmons, and P. C. G. Costa, editors, *Proceedings of the Eighth Conference on Semantic Technologies for Intelligence, Defense, and Security, STIDS 2013*, volume 1097, pages 71–78. CEUR Workshop Proceedings, 2013. (Cited on page 31.)

[99] S. Metz-Goeckel. Diskrete Diskriminierungen und persoeliches Glueck im Leben von Wissenschaftler/innen. *Erkenntnis und Methode*, 2009. (Cited on page 51.)

[100] T. Miksa and A. Rauber. Increasing preservability of research by process management plans. In *Proc. 1st International Workshop on Digital Preservation of Research Methods and Artefacts (DPRMA)*, 2013. (Cited on page 48.)

[101] H. Müller. ELAK, the e-filing system of the Austrian Federal Ministries, 2008. http://www.epractice.eu/en/cases/elak, last visited: November 2013. (Cited on page 29.)

[102] A. Norton. Predictive Policing - The Future of Law Enforcement in the Trinidad and Tobago Police Service. *Int. J. of Computer Applications*, 62:32–36, 2013. (Cited on page 31.)

[103] P. Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, 57:1701, 2009. http://ssrn.com/abstract=1450006, last visited: August 2013. (Cited on page 44.)

[104] A. Olesker. White Paper: Big Data Solutions For Law Enforcement, 2012. IDC White paper. (Cited on page 31.)

[105] T. O'Reilly, M. Loukides, J. Steele, and C. Hill. *How Data Science is Transforming Health Care*. O'Reilly Media, 2012. (Cited on page 23.)

[106] D. Osswald and G. Girard. Improving Business Outcomes with Big Data and Analytics in Communications, Media, and Entertainment, 2012. IDC White paper. (Cited on page 32.)

[107] K. Page, R. Palma, P. Holubowicz, G. Klyne, S. Soiland-Reyes, D. Cruickshank, R. G. Cabero, E. G. Cuesta, D. D. Roure, J. Zhao, and J. M. Gómez-Pérez. From workflows to research objects: an architecture for preserving the semantics of science. In *Proc. Linked Science Workshop*, 2012. (Cited on page 48.)

[108] M. Q. Patton. *Qualitative evaluation and research methods*. Sage, Newbury Park, 3rd edition, 1990. (Cited on page 10.)

[109] G. L. Paul and J. R. Baron. Information inflation: Can the legal system adapt? *Richmond Journal of Law & Technology*, XIII(3), 2007. (Cited on page 31.)

[110] Personal communication. DI Harald Leitenmüller, 2013. Chief Technology Officer at Microsoft Austria. (Cited on page 36.)

[111] Personal communication. Dr. Franz Haider, 2013. Chief Information Officer at the Austrian Minstery for Transport, Innovation and Technology. (Cited on page 29.)

[112] Personal communication. Dr. Franz Kainberger, 2013. Abteilungsleiter Stellvertreter Neuroradiologie und Muskuloskeletale Radiologie, AKH Wien. (Cited on pages 22 and 23.)

[113] Personal communication. Dr. Manuela Prokesch, 2013. Stv. Direktorin der pre-klinischen Abteilung, QPS Austria. (Cited on page 48.)

[114] A. Prasad and P. Prasad. Otherness at large: Identity and difference in the new globalized organizational landscape. *Gender, Identity and the Culture of Organizations*, 2002. (Cited on page 52.)

[115] J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, editors. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*. Springer, 2005. (Cited on page 19.)

[116] D. Rastetter. Managing Diversity in Teams. Erkenntnisse aus der Gruppenforschung. *Diversity Management. Impulse aus der Personalfuehrung*, pages 81–108, 2006. (Cited on page 75.)

[117] V. Reding. The EU's Data Protection rules and Cyber Security Strategy: two sides of the same coin, 2013. http://europa.eu/rapid/press-release_SPEECH-13-436_en.htm, last visited: August 2013. (Cited on page 45.)

[118] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011. (Cited on page 17.)

[119] C. Romero, P. Espejo, A. Zafra, J. Romero, and S. Ventura. Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, 21(1):135–146, 2013. (Cited on page 26.)

[120] B. Rowe, D. Wood, A. Link, and D. Simoni. Economic Impact Assessment of NIST's Text Retrieval Conference (TREC) Program. Technical report, National Institute of Standards and Technology, 2010. (Cited on page 19.)

[121] L. Schiebinger. *The Mind Has No Sex? Women in the Origins of Modern Science*. Harvard University Press, 1989. (Cited on page 52.)

[122] R. Sickles. Energy economics. In S. N. Durlauf and L. E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, 2008. (Cited on page 23.)

[123] Statistik Austria. *Bildung in Zahlen 2011/12. Schlüsselindikatoren und Analysen.* Statistik Austria, 2013. (Cited on pages 51 and 76.)

[124] V. Stodden. The legal framework for reproducible scientific research: Licensing and copyright. *Computing in Science & Engineering*, 11(1):35 –40, Feb. 2009. (Cited on page 42.)

[125] Y. Su, G. Agrawal, J. Woodring, K. Myers, J. Wendelberger, and J. Ahrens. Taming massive distributed datasets: data sampling using bitmap indices. In *Proceedings of HPDC*, pages 13–24. ACM, 2013. (Cited on page 38.)

[126] L. Sweeney. K-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002. (Cited on page 37.)

[127] A. Tashakkori and C. Teddlie. Introduction to mixed method and mixed model studies in the social and behavioral sciences. In V. L. P. Clark and J. W. Creswell, editors, *The mixed methods reader*, pages 7–26. Sage, Thousand Oaks, 2008. (Cited on page 7.)

[128] K. Temple. What Happens in an Internet Minute?, 2013. http://scoop.intel.com/what-happens-in-an-internet-minute/, last visited: August 2013. (Cited on page 1.)

[129] The Economist. Big data: Crunching the numbers, 2012. http://www.economist.com/node/21554743, last visited: December 2013. (Cited on page 27.)

[130] J. Thomas and K. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics.* IEEE Press, 2005. (Cited on page 18.)

[131] D. Tinholt. The Open Data Economy: Unlocking Economic Value by Opening Government and Public Data. CapGemini Consulting, 2013. http://www.capgemini-consulting.com/resource-file-access/resource/pdf/opendata_pov_6feb.pdf, last visited: 20 December 2013. (Cited on pages 46 and 47.)

[132] J. W. Tukey. Sunset salvo. *The American Statistician*, 40(1):72–76, 1986. (Cited on page 1.)

[133] D. Turner, M. Schroeck, and R. Shockley. Analytics: The real-world use of Big Data in financial services. IBM Global Business Services, May 2013. Executive Report. (Cited on page 27.)

[134] H. van de Sompel and C. Lagoze. All aboard: Toward a machine-friendly scholarly communication system. In T. Hey, S. Tansley, and K. Tolle, editors, *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, 2009. (Cited on page 24.)

[135] F. van Harmelen, G. Kampis, K. Börner, P. van den Besselaar, E. Schultes, C. Goble, P. Groth, B. Mons, S. Anderson, S. Decker, C. Hayes, T. Buecheler, and D. Helbing. Theoretical and technological building blocks for an innovation accelerator. *Eur. Phys. J. Special Topics*, 214:183–214, 2012. (Cited on page 24.)

[136] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press, 2005. (Cited on page 19.)

[137] A. Wetterer. *Arbeitsteilung und Geschlecherkonstruktion. 'Gender at Work' in theoretischer und historischer Perspektive.* Universitätsverlag Konstanz, 2002. (Cited on page 52.)

[138] Wikipedia. Data science, 2013. http://en.wikipedia.org/wiki/Data_science, last visited: December 2013. (Cited on page 50.)

[139] Wikipedia. Networked readiness index, 2013. http://en.wikipedia.org/wiki/Networked_Readiness_Index, last visited: December 2013. (Cited on page 35.)

[140] Wikipedia. Predictive analytics, 2013. http://en.wikipedia.org/wiki/Predictive_analytics, last visited: December 2013. (Cited on page 66.)

[141] C. Yiu. Big data opportunity. making government faster, smarter and more personal, 2012. Policy Exchange. (Cited on page 28.)

[142] ZIT Die Technologieagntur der Stadt Wien. Zit fempower studie 2012: Effekte der zit fempower massnahmen auf die karriereverlaeufe von frauen in der betrieblichen forschung, 2012. http://www.zit.co.at/fileadmin/user_upload/ZIT/Dienstleistungen/ZIT_Studie_2012_Frauenkarrieren.pdf, last visited: December 2013. (Cited on pages 75 and 76.)

Appendix A

Online Survey Questionnaire

Background information

Question 1: What is your current activity environment? (Provide comments if you wish to):

- Academia (University)
- Non-University Research
- Industry
- Public Office

Question 2: How many years of experience do you have in your activity area?

- 1-3
- 4-8
- 9 or more

Question 3: Would you consider yourself a... (Provide comments if you wish to)

- Researcher
- Service Provider
- Policy Maker
- User of Data Analytics technology
- Other:

Question 4: Gender

- Male
- Female

Question 5: In which country do you work?

- List of countries

Data Analytics Definition

Question 6: What is your understanding of Data Analytics?

- Free-test answer

Research and Development Focus

Question 7: Which of the following sub-fields do you focus on? (Provide specific details if you wish to)

- Search and Analysis
- Semantic Processing
- Cognitive Systems
- Visualisation and Presentation
- Other:

Question 8: Which of the following Application Domains do you find important today? (i.e. Application Domains you might already be working on. Provide specific details if you wish to)

- Healthcare
- Commerce
- Manufacturing and Logistics
- Transportation
- Energy and Utilities
- Public Sector / Government
- Education
- Tourism
- Telecommunications
- eScience (incl. Life Science)
- Law Enforcement
- Finance and Insurance

Question 9: Other Application Domains you find important

- Free-test answer

Challenges

Question 10: Which challenges do you see in Data Analytics?

- Free-test answer

Question 11: Following your previous answer, please judge if the following challenges will be important in the short, medium or long term.

- Privacy and Security – short term, medium term, long term, not important, don't know
- Algorithm Issues (e.g. Scalability) – short term, medium term, long term, not important, don't know
- Qualified Personnel – short term, medium term, long term, not important, don't know
- Data Preservation and Curation – short term, medium term, long term, not important, don't know
- Evaluation and Benchmarking – short term, medium term, long term, not important, don't know
- Data Ownership and Open Data – short term, medium term, long term, not important, don't know

Question 12: Which challenges do you see in Data Analytics?

- Healthcare – short term, medium term, long term, not important, don't know
- Commerce – short term, medium term, long term, not important, don't know
- Manufacturing and Logistics – short term, medium term, long term, not important, don't know
- Transportation –¿ (short term, medium term, long term, not important, don't know)
- Energy and Utilities – short term, medium term, long term, not important, don't know
- Public Sector / Government – short term, medium term, long term, not important, don't know
- Education – short term, medium term, long term, not important, don't know
- Tourism – short term, medium term, long term, not important, don't know
- Telecommunications – short term, medium term, long term, not important, don't know
- eScience (incl. Life Science) – short term, medium term, long term, not important, don't know

- Law Enforcement – short term, medium term, long term, not important, don't know
- Finance and Insurance – short term, medium term, long term, not important, don't know

Question 13: Other Application Domains you find important (please indicate Short/Medium/Long Term)

- Free-test answer

Public Funding

Question 14: Which research areas or topics in the Data Analytics field are most important and should be prioritized by public funding (name 3 and rank)

- Top Priority: Free-test answer
- Second Priority: Free-test answer
- Third Priority: Free-test answer

Question 15: Please complete the following news headline: 10,000,000 Euro for...

- Free-test answer

Question 16: Other comments you might have about data analytics

- Free-test answer

Appendix B

Interview Guideline

Einstieg

Vielen Dank, dass Sie sich Zeit genommen haben. Wie bereits telefonisch erwaehnt, möchten wir mit Ihnen darüber sprechen, wie in Ihrem Unternehmen mit Daten umgegangen wird. Wir, das ist die max.recall information systems GmbH und das Institut für Softwaretechnik und Interaktive Systeme der TU Wien, führen eine Roadmap Studie im Bereich Data Analytics durch. Das heisst wir beschaeftigen uns damit, welche Herausforderungen es in diesem Bereich gibt und in Zukunft geben wird und dazu wuerde ich gerne mit Ihnen sprechen. Ich moechte das Interview aufzeichnen, wenn das für Sie in Ordnung ist. Selbstverstaendlich werden personenbezogenen Daten anonymisiert und das Ergebnis dieses Gesprächs wird nur in Verbindung mit anderen Gesprächen, also als kumuliertes Ergebnis in unsere Studie einfliessen.

Hintergrund

Laut Intel werden im Internet pro Minute 100.000 Tweets abgesetzt, 277.000 Facebook-Logins durchgefuehrt, 204 Millionen E-Mails ausgetauscht, und mehr als 2 Millionen Suchanfragen abgesetzt. Wir, die Bewohner dieses digitalen Universums, generieren pro Jahr mehr als 200 Exabyte an Daten. Dies entspricht etwa 20 Millionen Mal der Library of Congress. Daten werden primaer digital erfasst, mit einer Vielzahl von Datenstroemen unterschiedlicher Formate verknuepft und abgeglichen. Analysen passieren in Echtzeit; ausgewertet und visualisiert wird interaktiv. Vorhersagen sind zuverlaessiger den je; Reagieren wird in Sekundeschnelle möglich.

Der zunehmende Einsatz von Sensoren aller Art führt zu einer Flut an maschinengenerierten Daten. Kontinuierliche Verbesserungen der Sensortechnologien haben zur Folge, dass Daten in immer hoeherer zeitlicher und rauumlicher Aufloesung zur Verfügung stehen. Das “Internet der Dinge” wird zunehmend zur Realitaet, und damit werden Geraete und Objekte zukuenftig mehr Daten beisteuern als je zuvor. Die dadurch entstehenden Datenmengen koennen manuell nicht verarbeitet werden und bestaerken damit den Bedarf an innovativen Ansaetzen zu deren automatisierten Verarbeitung.

Doch welche Herausforderung stellen diese Entwicklungen an Forschung, Industrie und Gesellschaft? Welche technologischen Barrieren gilt es zu ueberbruecken? Welche neuen Geschaeftsmodelle entstehen aus diesen Moeglichkeiten? Wie sieht die rechtliche Situation in diesem Umfeld aus? Diese und andere Themen sollen im Rahmen des Interviews diskutiert werden. Die Ergebnisse fließen in eine, von der Oesterr. Forschungsfoerderungsgesellschaft

FFG und dem Bundesministerium für Verkehr, Innovation und Technologie – bmvit in Auftrag gegebenen Roadmap-Studie zum Thema "IKT der Zukunft: Daten durchdringen – Intelligente Systeme" ein. Ziel dieser Studie ist es die Entscheidungsgrundlage für die kurz-, mittel- bis langfristigen Zielsetzungen des Foerderprogramms im Themengebiets „Daten durchdringen" zu liefern.

Datenart & Umgang mit den Daten

Es ist ja davon auszugehen, dass in ihrem Unternehmen verschiedene Daten anfallen. Uns interessiert jetzt besonders in welcher Form Daten vorhanden sind und wie Sie diese weiter verwenden.

- Um welche Daten handelt es sich konkret?
- Wie werden Daten gespeichert? Daten Silos? Zentral?
- Handelt es sich um strukturierte Daten oder unstrukturierte Daten (oder beides)?
- (Wie) werden sie bearbeitet und weiterverwendet? Liegen diese brach?
- Werden die Daten in irgendeiner Form systematisch analysiert und aufbereitet?

Welche Tools verwenden Sie dafuer?

- Verwenden Sie (innerbetriebliche) Lösungen, ein eigenes Data-Warehouse?
- Welche Business Intelligence-Loesungen verwenden Sie?
- Wozu setzen Sie diese Tools ein? Welchen Mehrwert sollten sie generieren?
- Engagieren Sie evtl. externe Firmen um Daten zu strukturieren und zu bearbeiten? Wenn ja, welche?

Wie ist Ihre Zufriedenheit mit diesen Tools bzw. Firmen?

- Sehen Sie ungenuetztes Potential in den Daten? Nachholbedarf bei Technologien? Datenstruktur?
- Welche Komponenten einer solchen Applikation sind fuer Sie besonders relevant?

Herausforderungen

Was sind Ihrer Meinung nach die groessten Herausforderungen in Ihrem Bereich?

- was davon kurzfristig, was langfristig?

Einige Bereiche andiskutieren:

- Gesellschaftlich / Oeknonomisch
- Wie ist es mit dem Bereich Security,

- Algorithmen? Performance? Heterogenitaet der Daten? Datensilos? Datenfusion? Big-Data?
- Open-Data?
- Geschaeftsmodelle
- wie schwierig ist es, qualifiziertes Personal zu finden?
- Data Preservation

Forschung und Entwicklung

Wie war das als Sie begonnen haben, sich mit diesem Thema auseinanderzusetzen? Wie hat sich der Bereich weiterentwickelt?

- Welche Themen wurden schwerpunktmaessig behandelt?
- Haben Sie in diesem Bereich auch geforscht/Diplomarbeiten vergeben etc.?

Nehmen Sie bzw. haben Sie bereits Foerdergelder in Anspruch genommen?

- Wenn ja: In welchem Ausmass und Umfang wurden Foerdergelder in Anspruch genommen? (Wie) Waren Sie zufrieden?
- Wenn nein: Was ist der Grund, warum Sie das nicht gemacht haben?
- Ist das Prozedere, sind es die Vorgaben, ist hier zu wenig unternehmensspezifische Foerderung vorhanden, gibt es Bedarf?
- Für welche Bereiche in Data Analytics sollten, Ihrer Meinung nach, umfangreiche Forschungsgelder in Zukunft eingesetzt werden?

Was glauben Sie, in welchen Bereichen wird das Thema Data Analytics in Zukunft eine gross Rolle spielen?

- Bei Ihnen im Unternehmen
- generell

Thema Vernetzung: Für Sie relevant?
Wo sollten Ihrer Meinung nach Gelder von der oeffentlichen Hand investiert werden? Projekte, Struktur, Anwendungsgebiete, Start-ups?
Wuerden Sie Daten für die Forschung zur Verfügung stellen?

Data Analytics Landschaft

Welche Unternehmen sehen Sie als die „Großen“ in Österreich, für die Data Analytics wichtig ist und warum? In House Entwicklung?

- Kennen Sie weitere Unternehmen?
- Was genau sind deren Geschaeftsbereiche?
- Zu welchen Unternehmen hatten Sie bereits Kontakt?

Appendix C

World Café Discussion Topics

- Welche Herausforderungen muessen im Bereich der Daten Visualisierung und Repräsentation bewaeltigt werden?
- Vor welchen Herausforderungen steht man im Bereich der Kognitiven Systeme und Prediction?
- Wie sehen die wichtigsten Herausforderungen im Bereich der Semantischen Datenverarbeitung aus?
- Welche Herausforderungen muessen in den Bereichen der Datenintegration & Datenfusion bewaeltigt werden? (bspw. multi-modal, multi-lingual, multi-spectral data)
- Vor welche Herausforderungen steht man aus algorithmischer Sicht? (bspw. Parallelisierung, incompletness of data, statistical models for big data, etc)
- Wie sehen die groeßten Herausforderungen im Bereich „Suche und Analyse“ aus?
- Wie sehen Sie die groeßten Herausforderungen im Bereich der rechtlichen Themen (z.B., Data Privacy and Security, Compliance issues, Data ownership, Service Levels, Reliability and other warranties, Indemnification and Limitations of Liabilities)?
- Bedarf an qualifiziertem Personal: Welche Qualifikationen benoetigt Personal um in Zukunft national und international erfolgreich zu sein?
- Welche neuen Geschaeftsmodelle ermöglicht die Verfuegbarkeit neuer Verfahren zur Daten-Durchdringung? Welche neuen Services erwarten Sie sich?
- Welche Staerken und Schwaechen besitzt Oesterreich auf dem Gebiet der Handhabbarmachung von Daten?
- In welchen Anwendungsgebieten wird in Zukunft die Handhabbarmachung von Daten die groeßte Rolle für Oesterreich spielen? (Bspw. Healthcare, Commerce, Manufacturing, etc.)
- Welche Massnahmen sollen zur nachhaltigen Vernetzung der nationalen StakeholderInnen gesetzt werden?

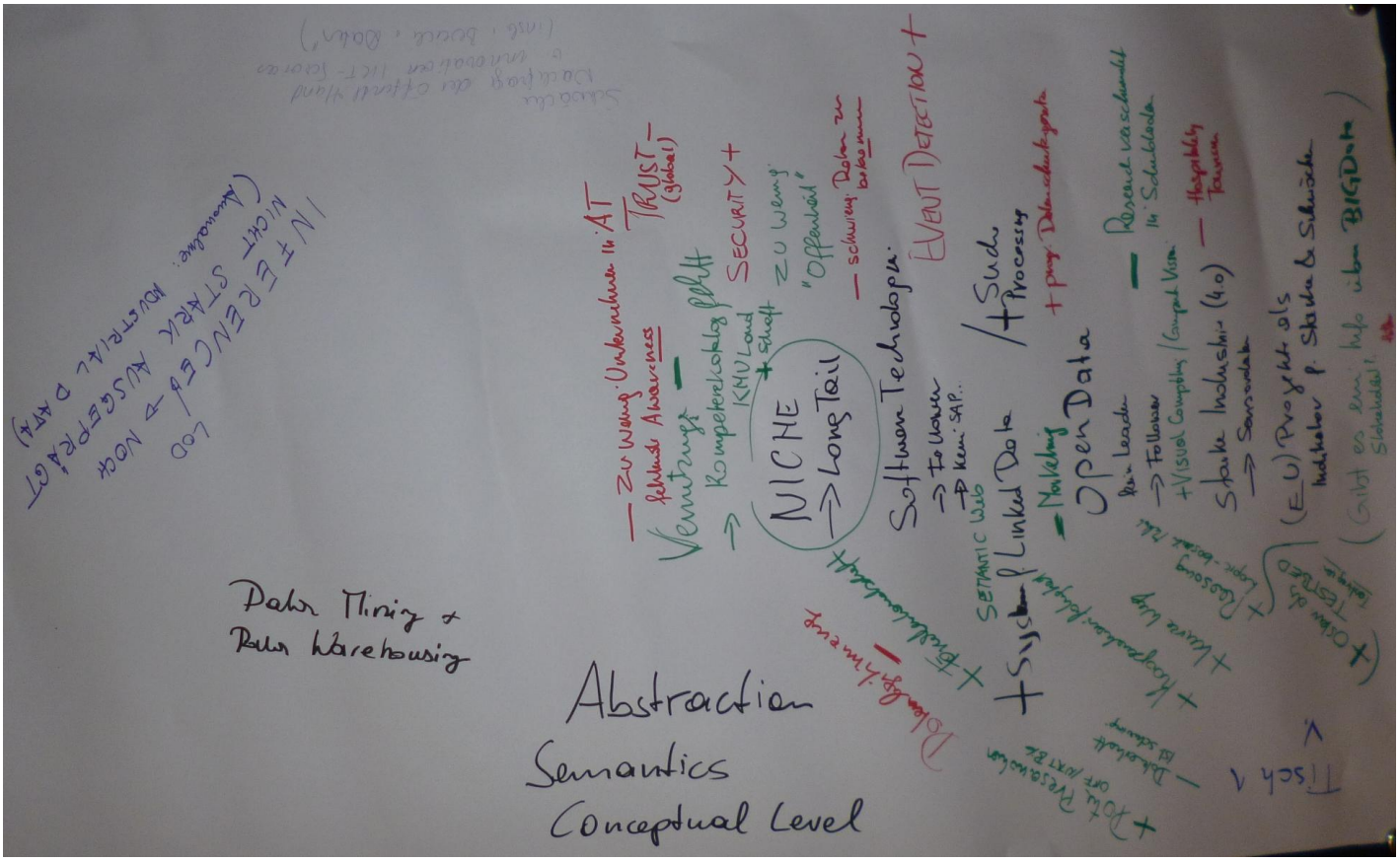
- Welche gesellschaftlichen und oekonomischen Auswirkungen erwarten Sie aufgrund der allgegenwaertigen Datenanalyse für Oesterreich?
- In welches Forschungsprojekt bzw. welchen Anwendungsbereich sollten jetzt 10 Mio. Euro investiert werden?
- Wie werden wir im Jahre 2025 mit Daten hantieren?

Appendix D

World Café Table Cloths

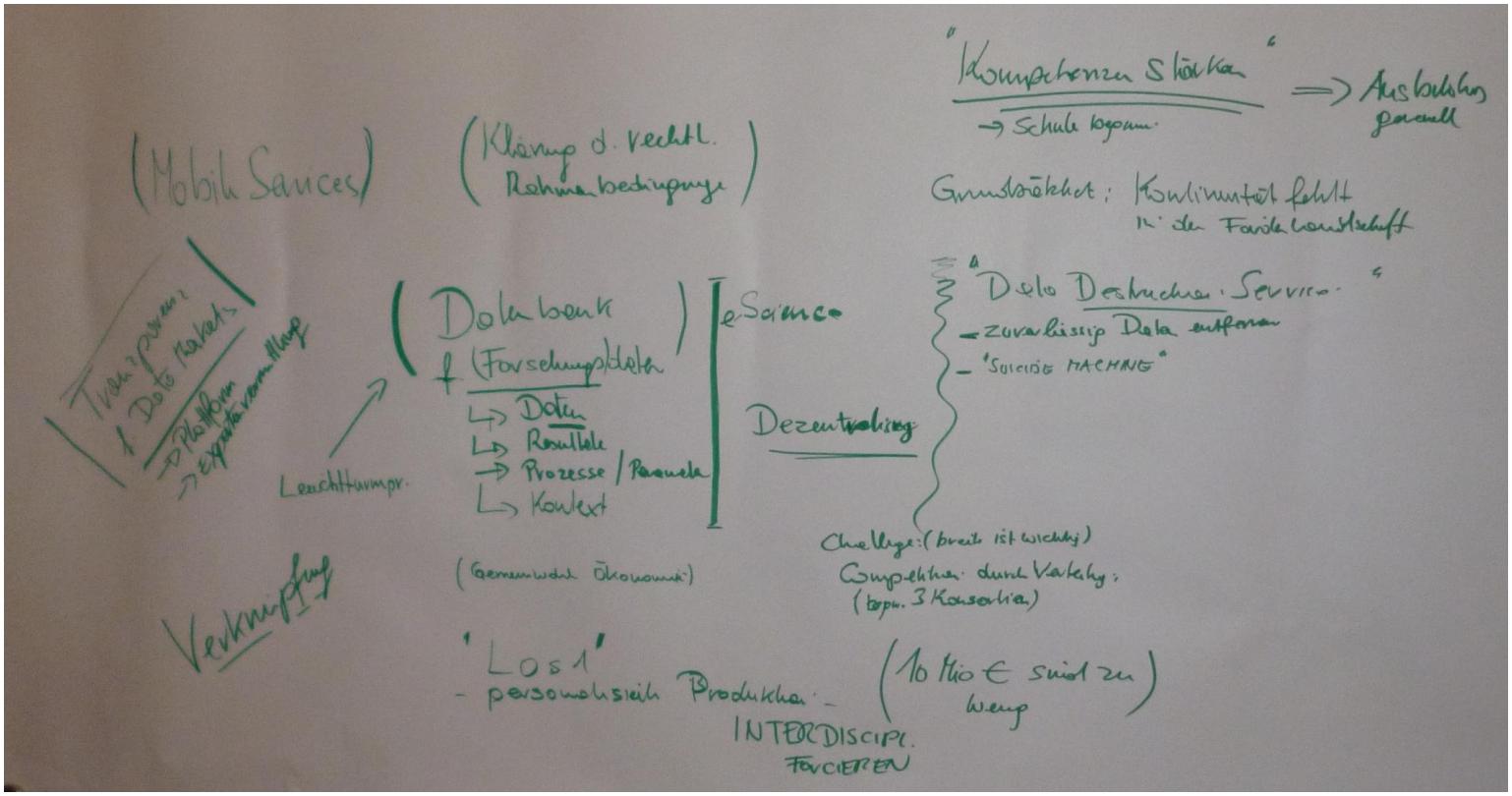
All of the table cloths on which notes were made during the World Cafés at the workshops are reproduced on the following pages.

Salzburg, Table-1, morning session:
„Welche Stärken und Schwächen besitzt Österreich auf dem Gebiet der Handhabbarmachung von Daten?“

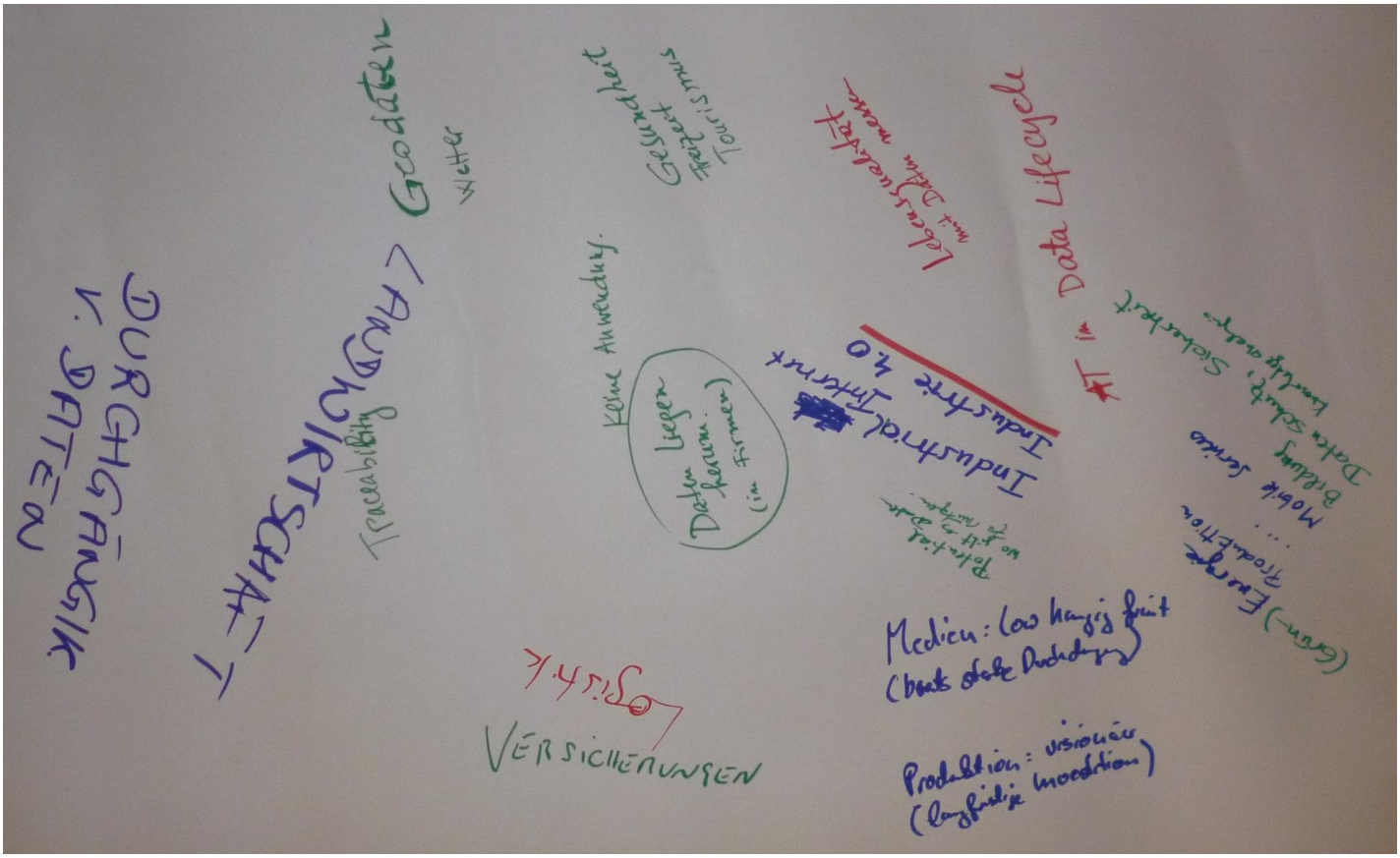


Salzburg, Table-1, afternoon session:
„Welche Herausforderungen müssen im Bereich der Daten Visualisierung und Repräsentation bewältigt werden?“

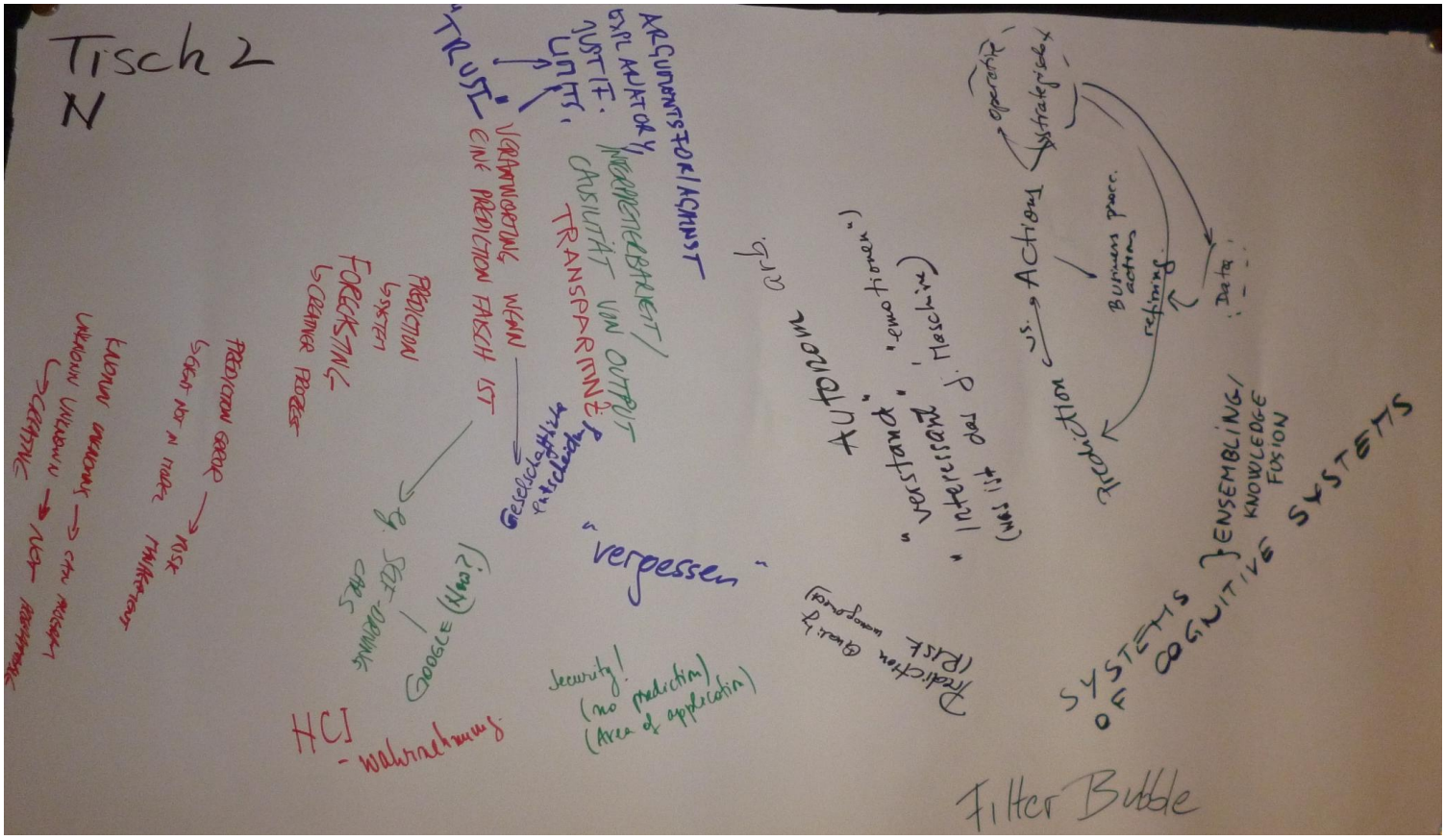
Salzburg, Table-1, afternoon tea session:
„In welches Forschungsprojekt bzw. welchen Anwendungsbereich sollten jetzt € 10 Mio. investiert werden? “



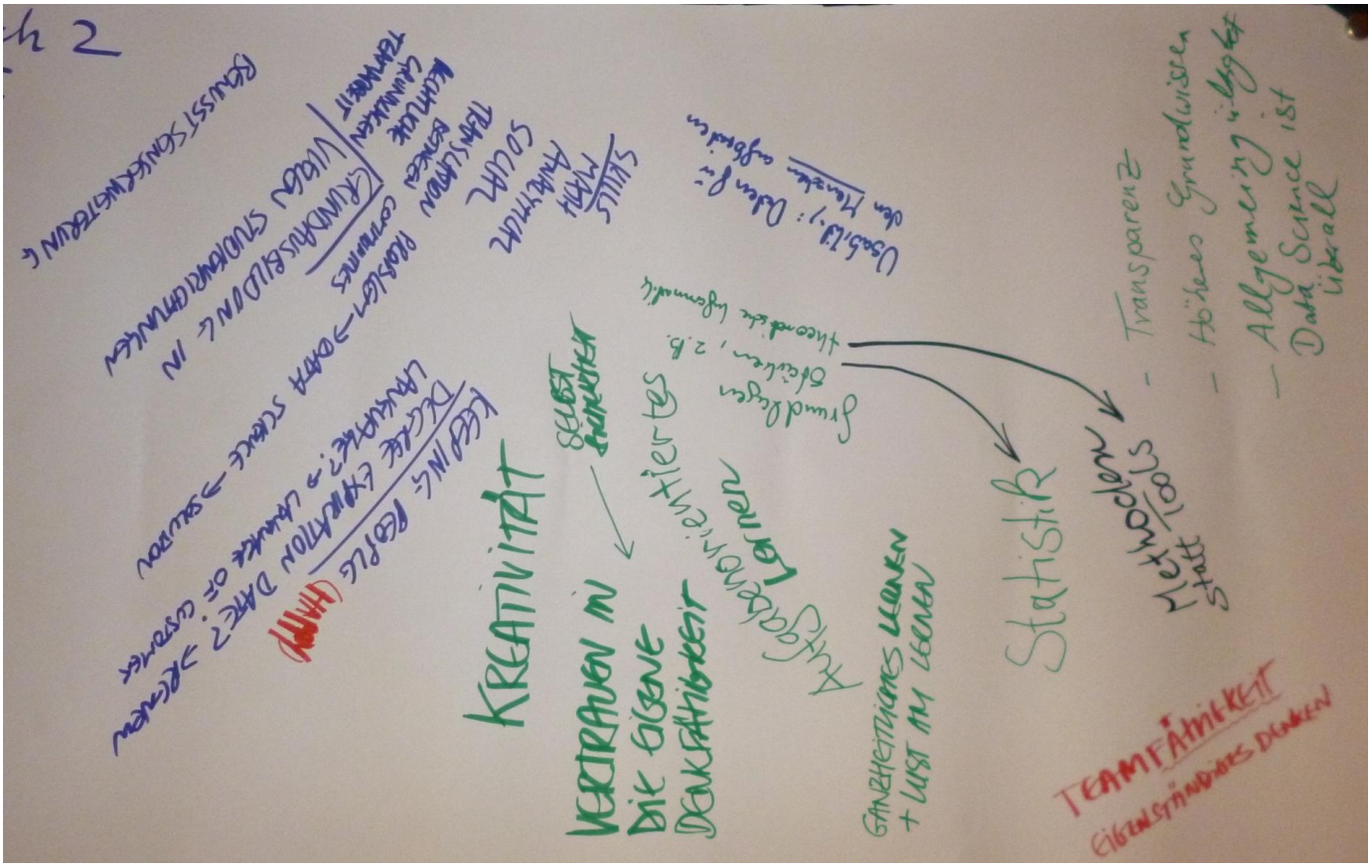
Salzburg, Table-2, morning session:
„In welchen Anwendungsgebieten wird in Zukunft die Handhabbarmachung von Daten die größte Rolle für Österreich spielen?“



Salzburg, Table-2, afternoon session:
„Vor welchen Herausforderungen steht man im Bereich der Kognitiven Systeme und Prediction?“

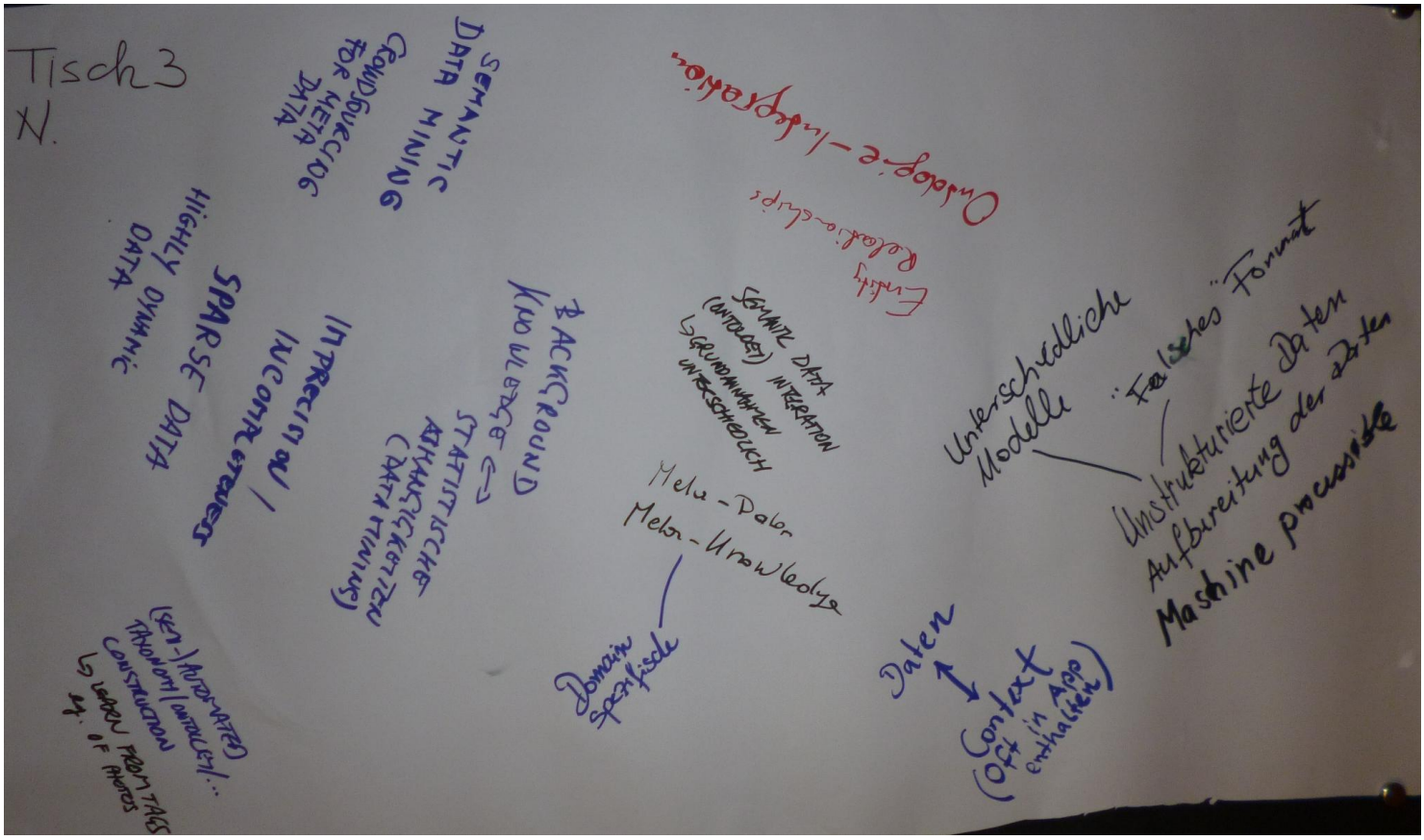


Salzburg, Table-2, afternoon tea session:
„Stichwort „Bedarf an Qualifiziertem Personal“: Welche Qualifikationen benötigt Personal um in Zukunft national und international erfolgreich zu sein?“

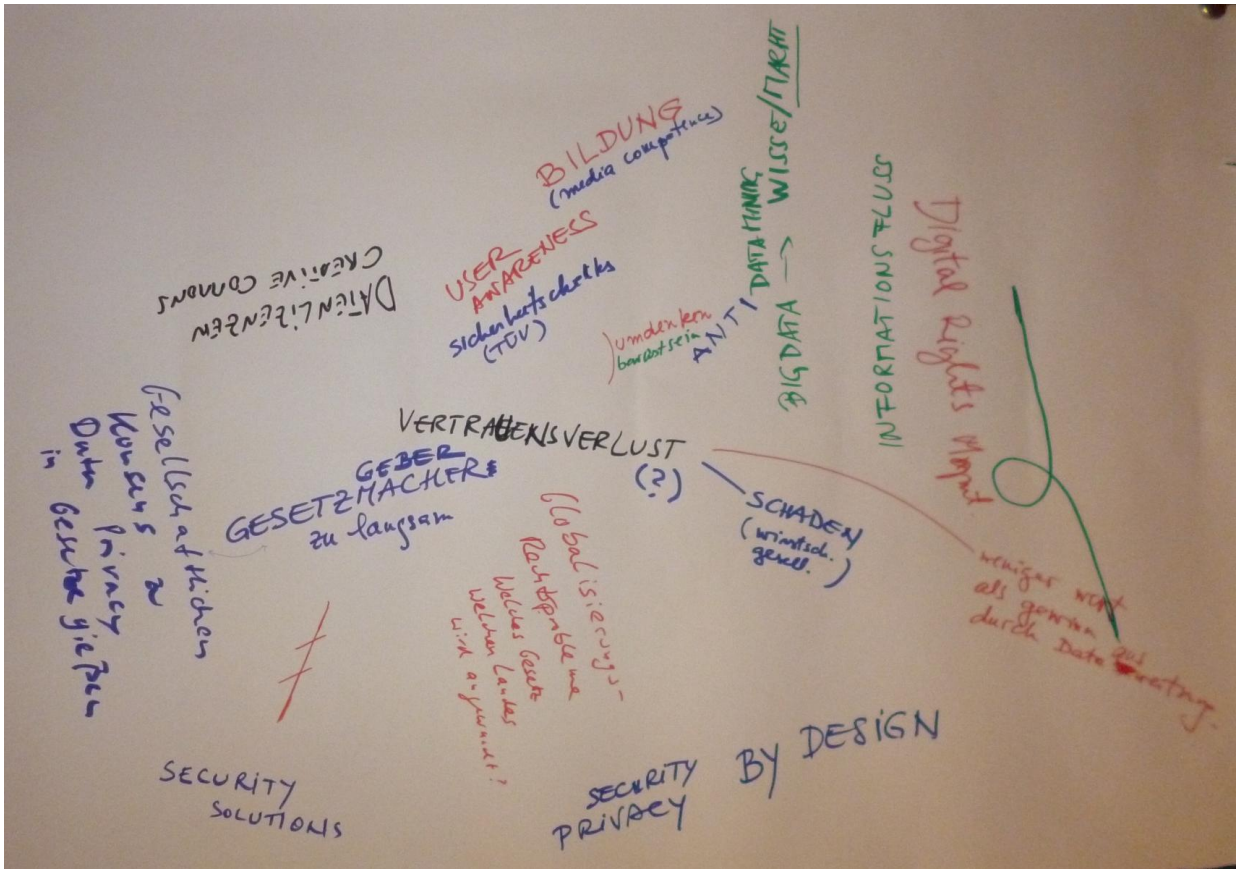


Salzburg, Table-3, morning session:
„Wie sehen die größten Herausforderungen im Bereich „Suche und Analyse“ aus?“

Salzburg, Table-3, afternoon session:
„Wie sehen die wichtigsten Herausforderungen im Bereich der Semantischen Datenverarbeitung?“

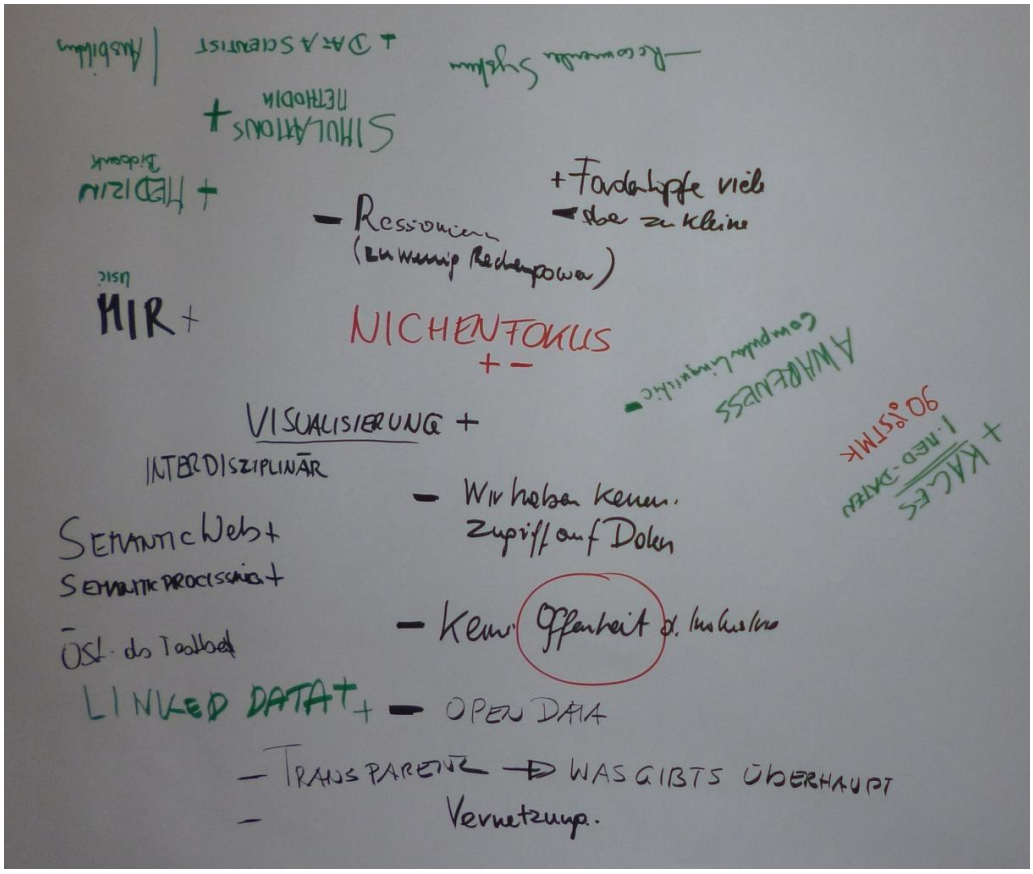


Salzburg, Table-3, afternoon tea session:
„Wie sehen Sie die größten Herausforderungen im Bereich der rechtlichen Themen aus?“

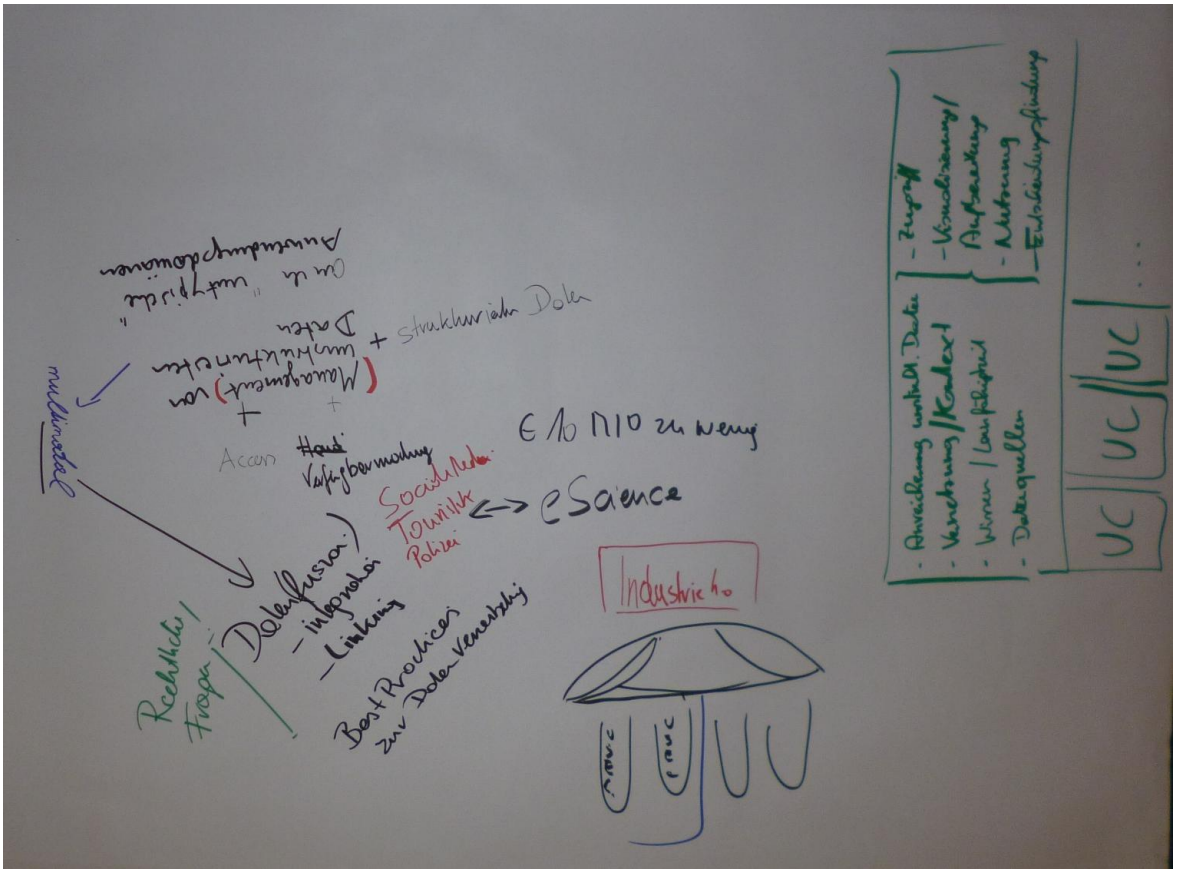


Graz, Table-1, morning session:
„Welche Herausforderungen müssen im Bereich der Daten Visualisierung und Repräsentation bewältigt werden?“

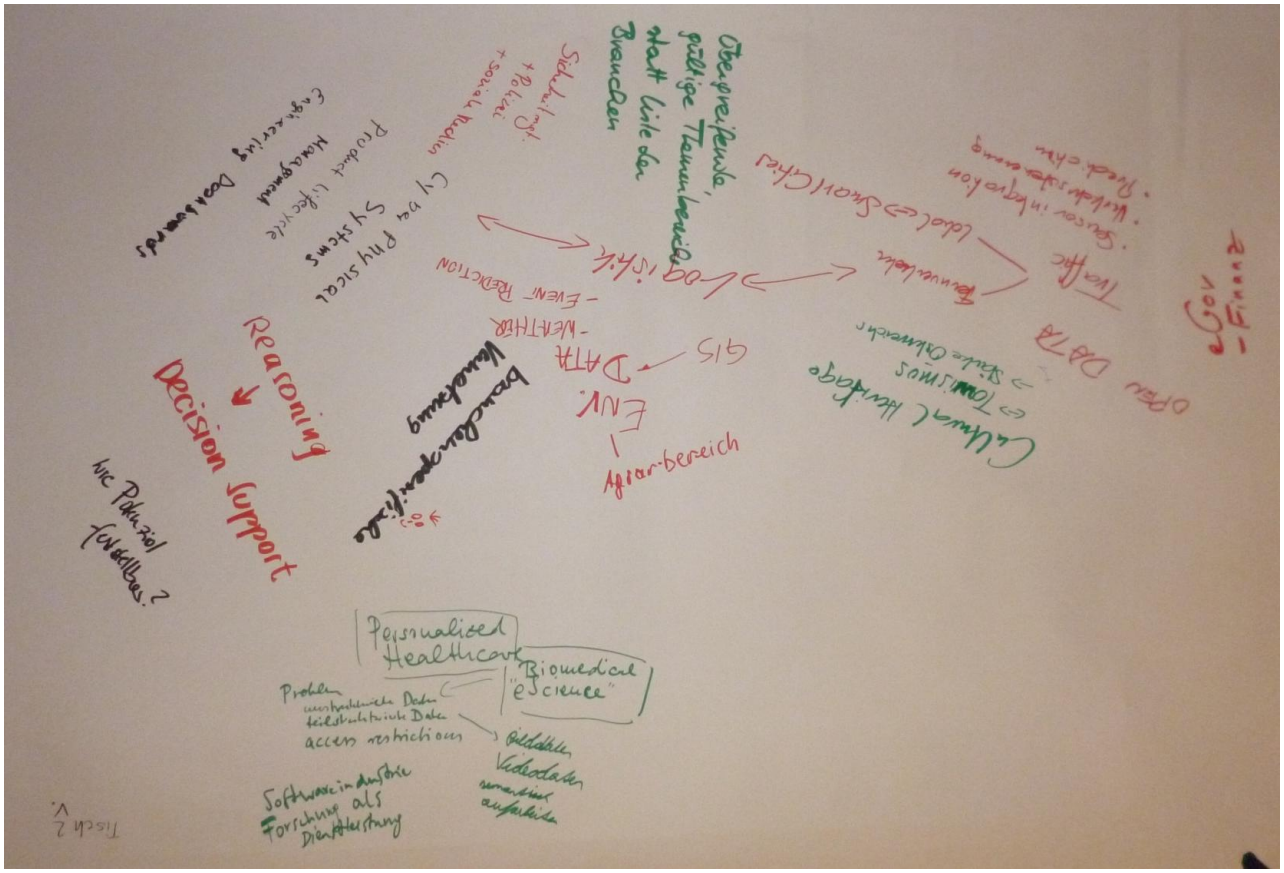
Graz, Table-1, afternoon session:
„Welche Stärken und Schwächen besitzt Österreich auf dem Gebiet der Handhabbarmachung von Daten?“



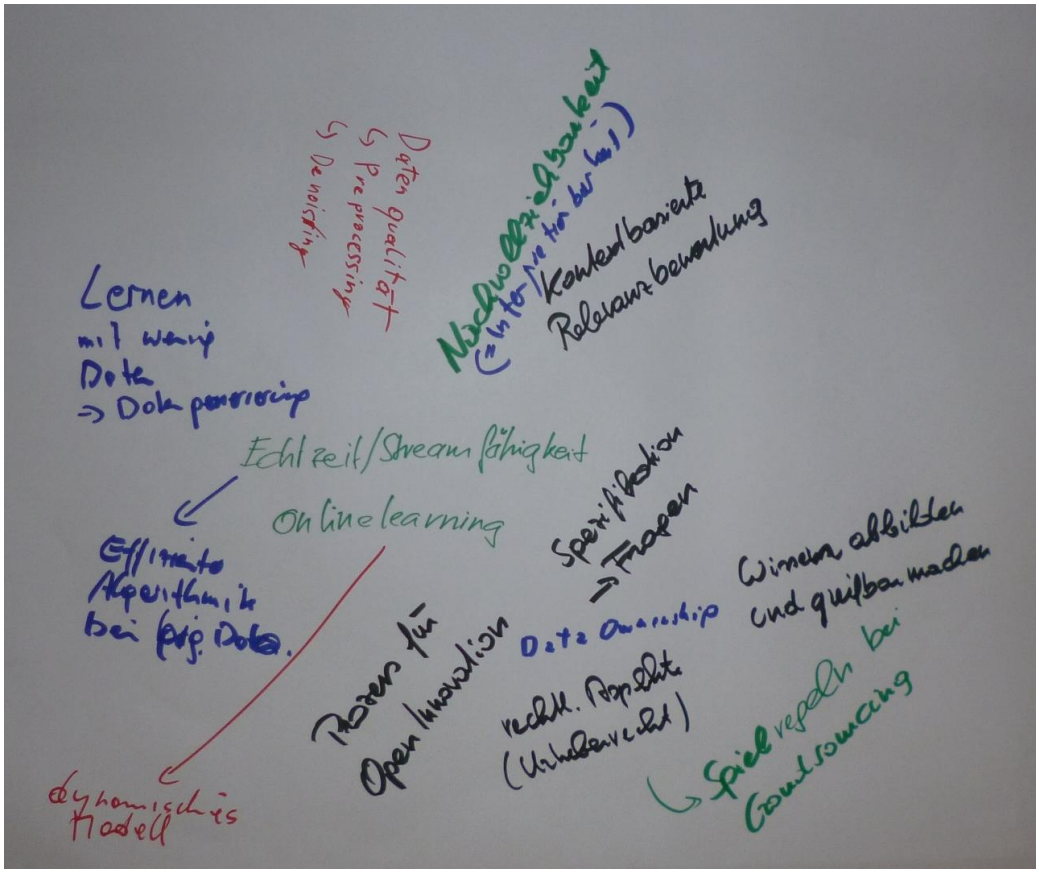
Graz, Table-1, afternoon tea session:
„In welches Forschungsprojekt bzw. welchen Anwendungsbereich sollten jetzt € 10 Mio. investiert werden? “



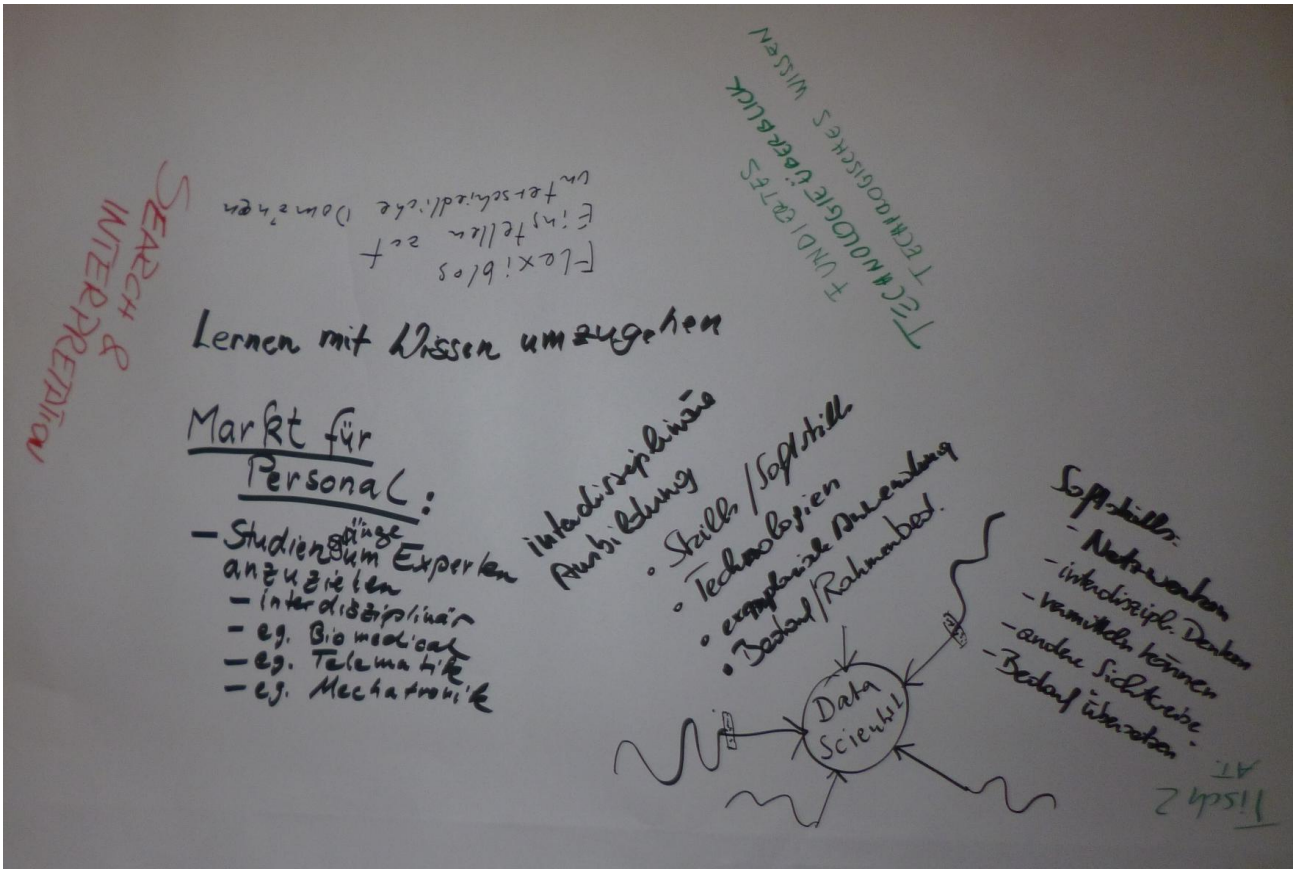
Graz, Table-2, morning session:
„In welchen Anwendungsgebieten wird in Zukunft die Handhabbarmachung von Daten die größte Rolle für Österreich spielen?“



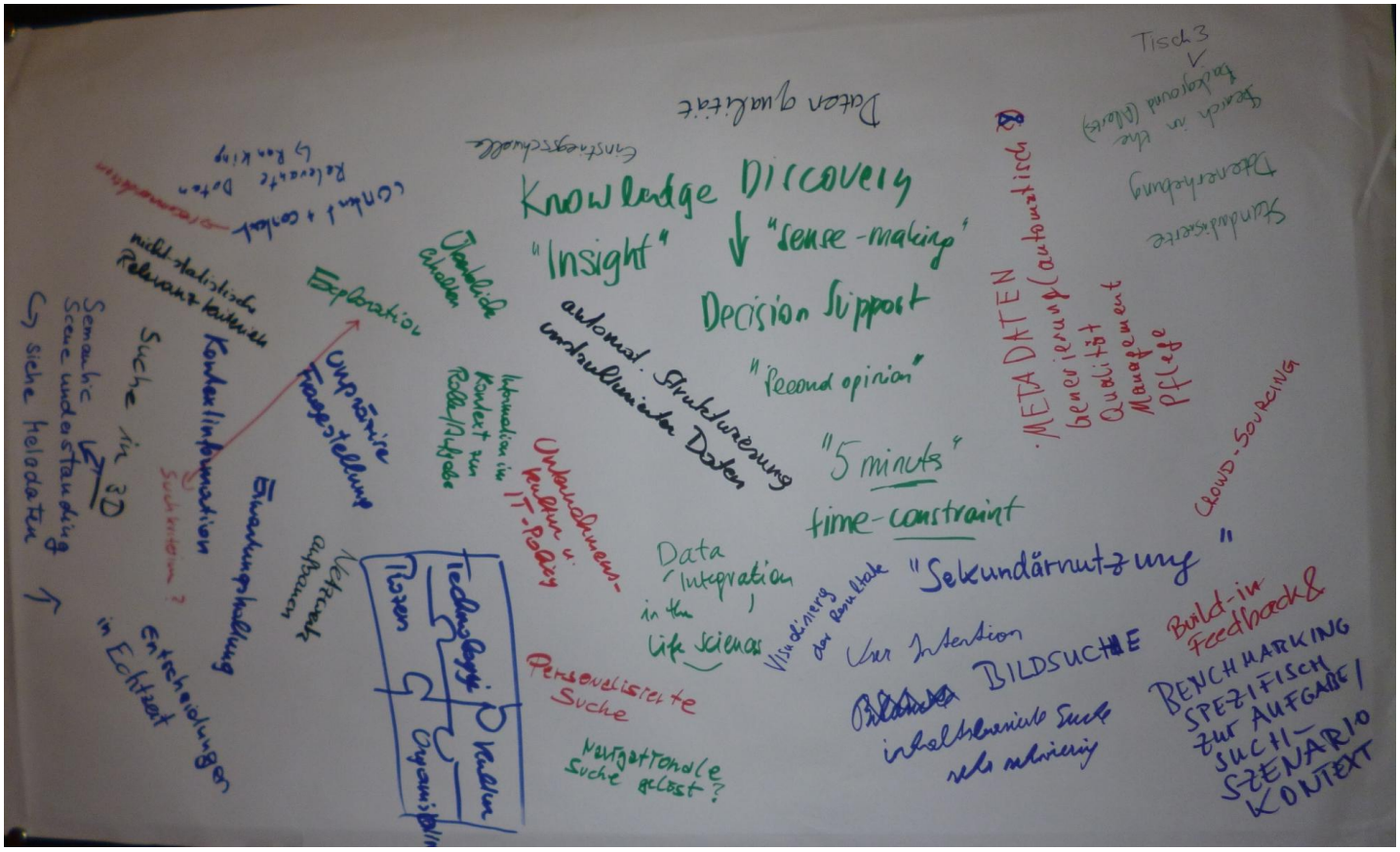
Graz, Table-2, afternoon session:
„Vor welchen Herausforderungen steht man im Bereich der Kognitiven Systeme und Prediction?“



Graz, Table-2, afternoon tea session:
„Stichwort „Bedarf an Qualifiziertem Personal“: Welche Qualifikationen benötigt Personal um in Zukunft national und international erfolgreich zu sein?“

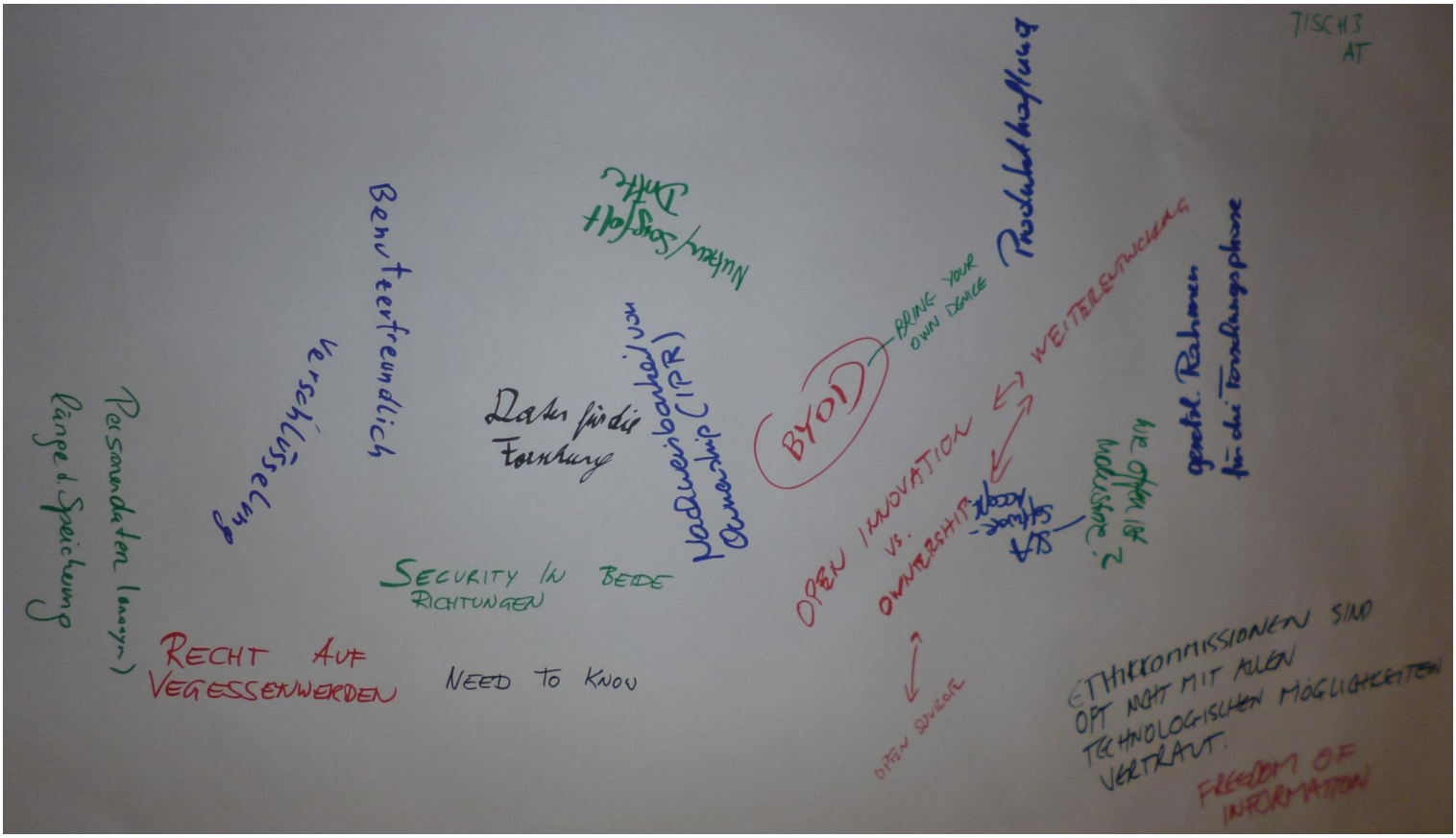


Graz, Table-3, morning session:
„Wie sehen die größten Herausforderungen im Bereich „Suche und Analyse“ aus?“



Graz, Table-3, afternoon session:
„Wie sehen die wichtigsten Herausforderungen im Bereich der Semantischen Datenverarbeitung?“

Graz, Table-3, afternoon tea session:
„Wie sehen Sie die größten Herausforderungen im Bereich der rechtlichen Themen aus?“

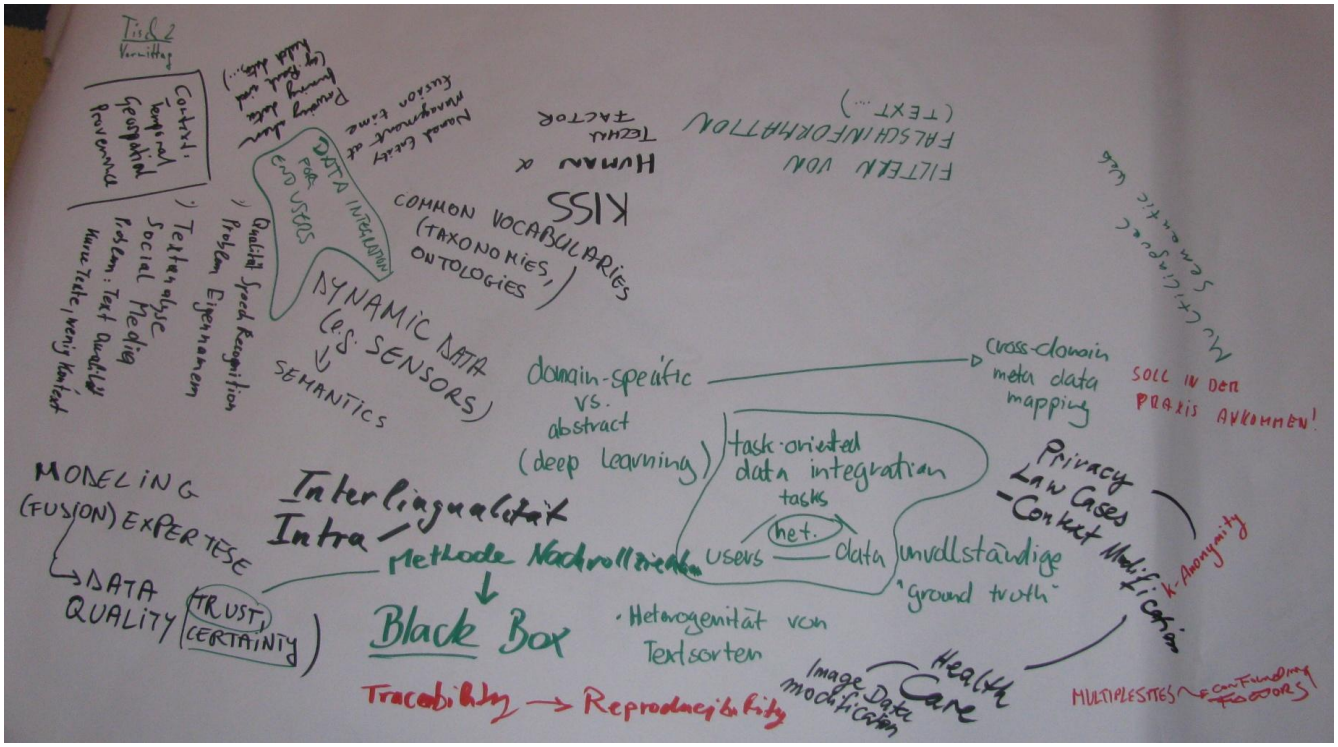


Vienna, Table-1, morning session:
„Welche Herausforderungen müssen im Bereich der Daten Visualisierung und Repräsentation bewältigt werden?“

Vienna, Table-1, afternoon session:
„Vor welchen Herausforderungen steht man im Bereich der Kognitiven Systeme und Prediction?“

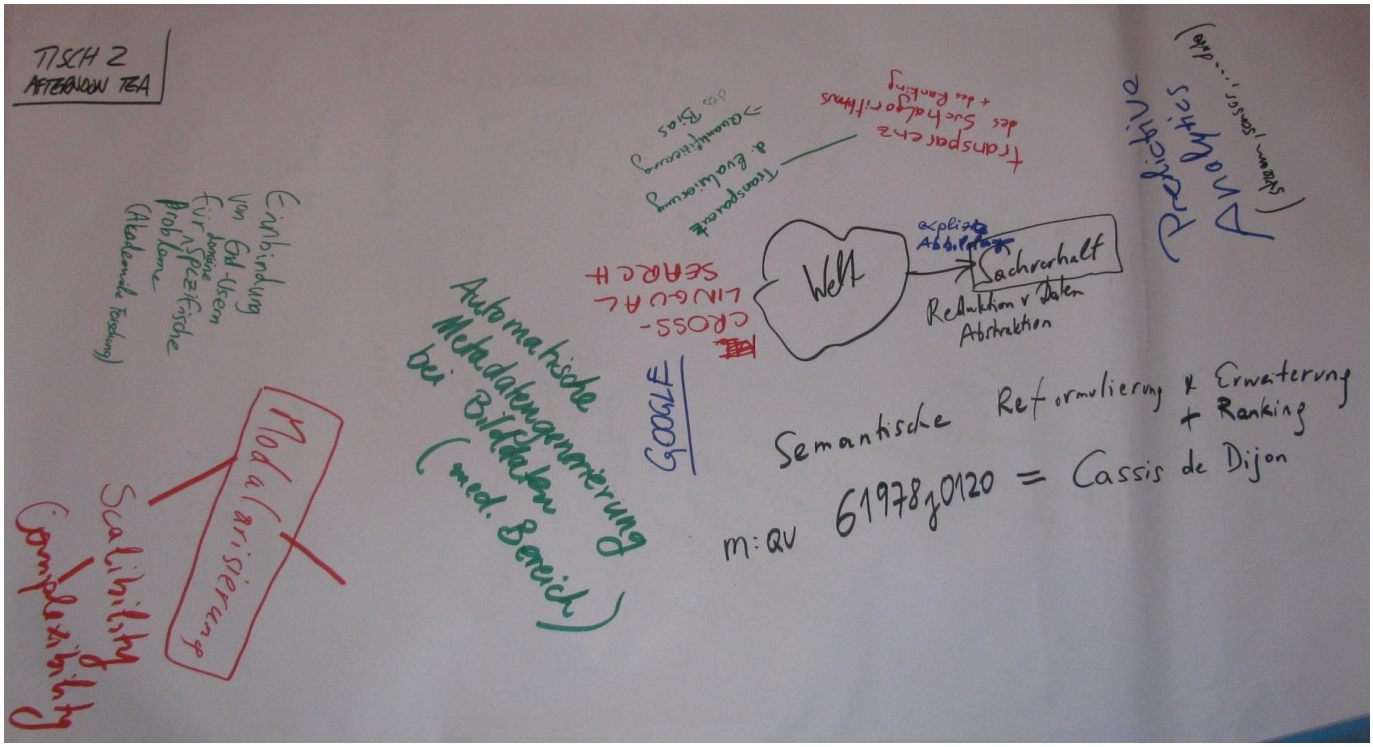
Vienna, Table-1, afternoon tea session:
„Wie sehen die wichtigsten Herausforderungen im Bereich der Semantischen Datenverarbeitung aus?“

Vienna, Table-2, morning session:
„Welche Herausforderungen müssen in den Bereichen der Datenintegration & -fusion bewältigt werden? (eg. multi-modal, multi-lingual, multi-spectral data)“



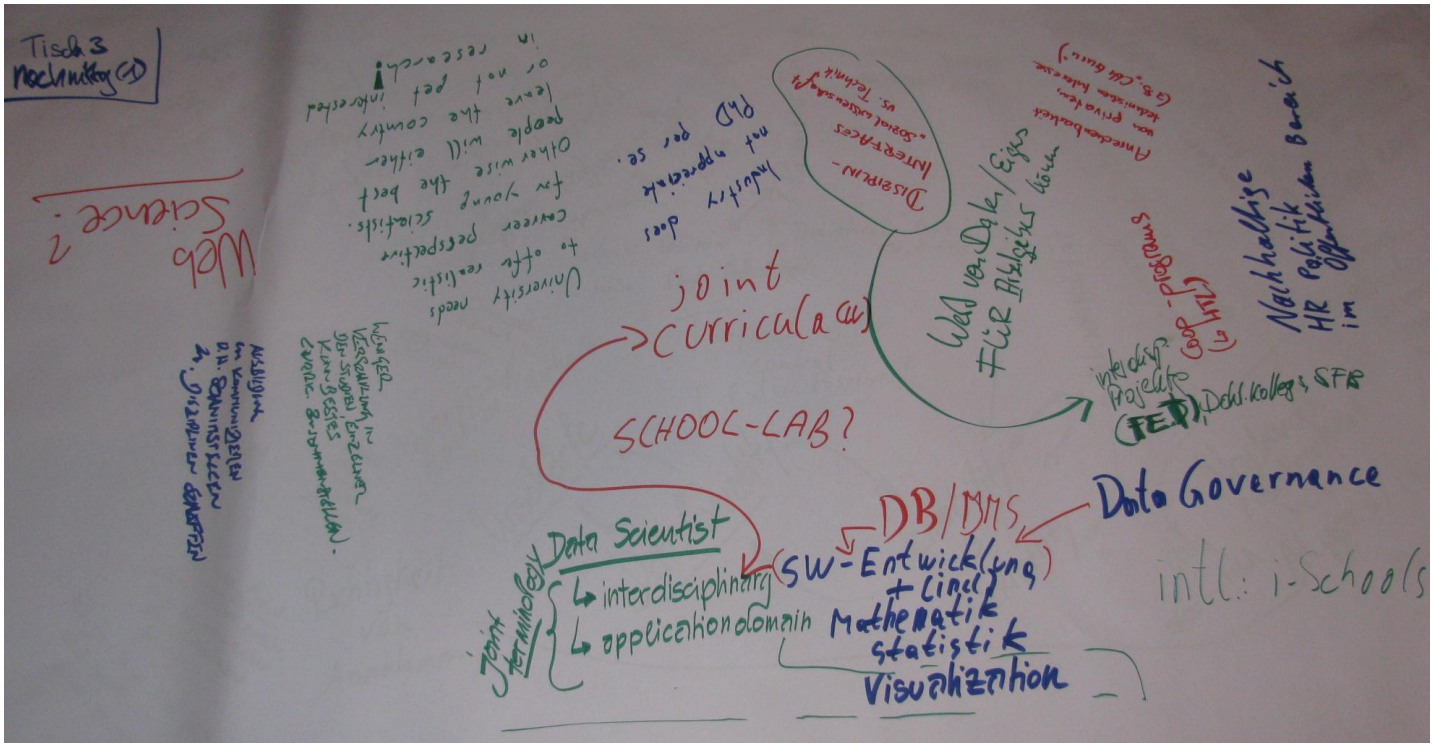
Vienna, Table-2, afternoon session:
„Vor welchen Herausforderungen steht man aus algorithmischer Sicht? (bspw. Parallelisierung, incompletness of data, statistical models for big data, etc)?“

Vienna, Table-2, afternoon tea session:
„Wie sehen die größten Herausforderungen im Bereich Suche und Analyse aus?“

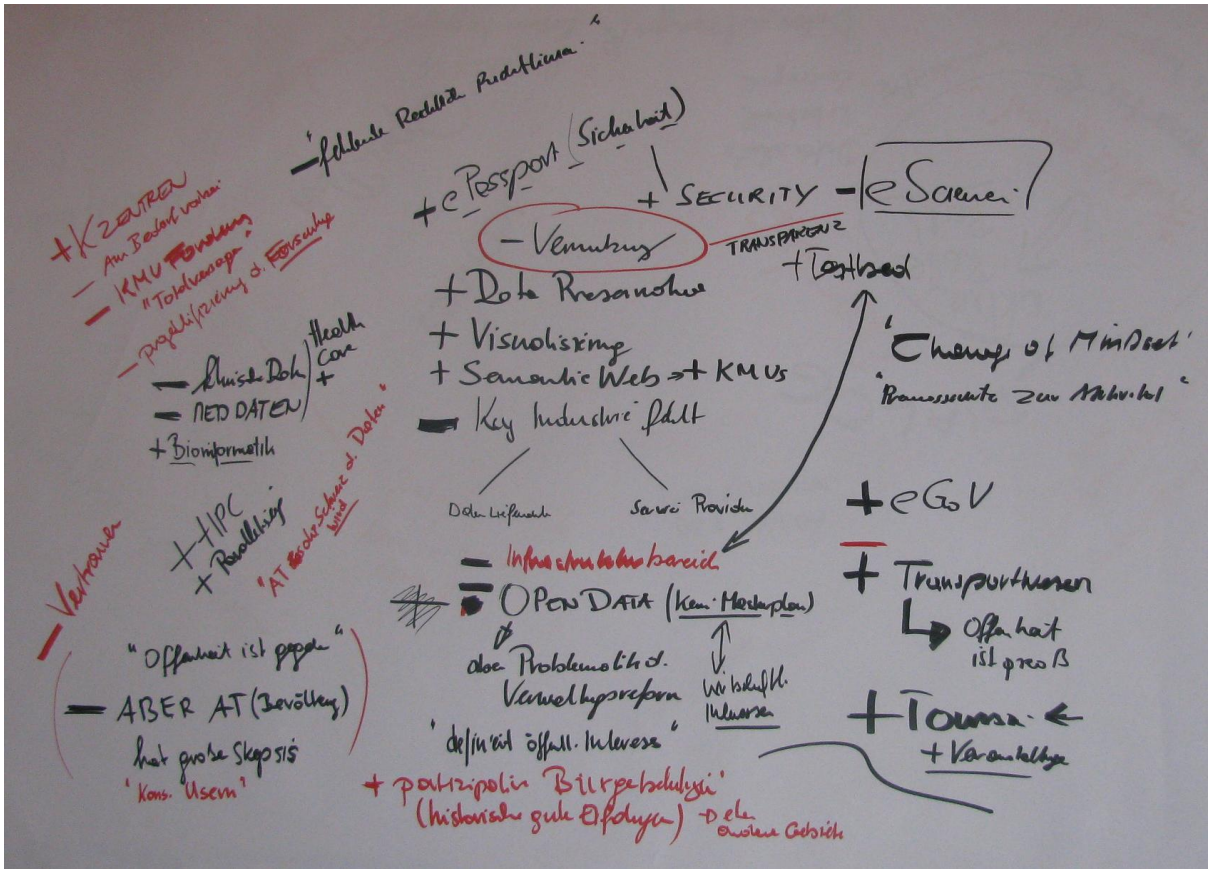


Vienna, Table-3, morning session:
„Wo sehen Sie die größten Herausforderungen im Bereich der rechtlichen Themen (e.g. Data Privacy and Security, Compliance, Ownership, Service Levels, Reliability, Indemnification and Limitations of Liabilities)?

Vienna, Table-3, afternoon session:
„Bedarf an qualifiziertem Personal: Welche Qualifikationen benötigt Personal um in Zukunft national und international erfolgreich zu sein?“



Vienna, Table-4, morning session:
„Welche Stärken und Schwächen besitzt Österreich auf dem Gebiet der Handhabbarmachung von Daten?“

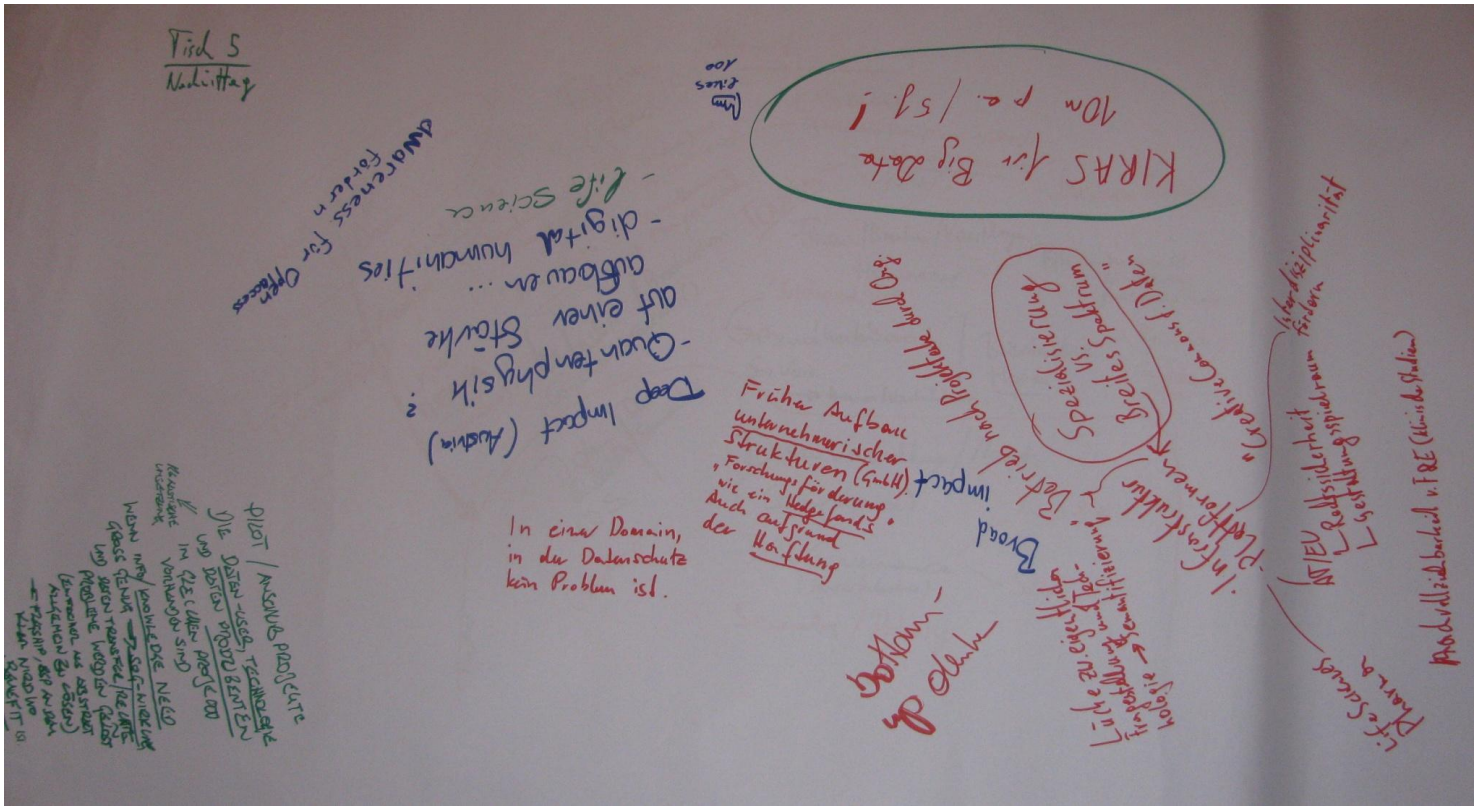


Vienna, Table-4, afternoon session:
„In welchen Anwendungsgebieten wird in Zukunft die Hand-habbarmachung von Daten die größte Rolle für Österreich spielen? (e.g., Healthcare, Commerce, Manufacturing…)“

Vienna, Table-4, afternoon tea session:
„ Welche Maßnahmen sollen zur nachhaltigen Vernetzung der nationalen StakeholderInnen gesetzt werden? “

Vienna, Table-5, morning session:
„Welche gesellschaftlichen und ökonomischen Auswirkungen erwarten Sie aufgrund der Möglichkeit zur allgegenwärtigen Datenanalyse für Österreich?“

Vienna, Table-5, afternoon session:
„In welches Forschungsprojekt bzw. welchen Anwendungsbereich sollten jetzt € 10 Mio. investiert werden?“



Vienna, Table-5, afternoon tea session:
„Wie werden wir im Jahre 2025 mit Daten hantieren?“