A Researcher's View on (Big) Data Analytics in Austria Results from an Online Survey

Ralf Bierig[1(✉)], Allan Hanbury[1], Florina Piroi[1], Marita Haas[2], Helmut Berger[2], Mihai Lupu[1], and Michael Dittenbach[2]

[1] Institute of Software Technology and Interactive Systems, Vienna University of Technology, Favoritenstr. 9-11/188-1, 1040 Vienna, Austria
{bierig,hanbury,piroi,lupu}@ifs.tuwien.ac.at
[2] Max.recall Information Systems GmbH, Künstlergasse 11/1, 1150 Vienna, Austria
{m.haas,h.berger,m.dittenbach}@max-recall.com

Abstract. We present results from questionnaire data that were collected from leading data analytics researchers and experts across Austria. The online survey addresses very pressing questions in the area of (big) data analysis. Our findings provide valuable insights about what top Austrian data scientists think about data analytics, what they consider as important application areas that can benefit from big data and data processing, the challenges of the future and how soon these challenges will become relevant, and become potential research topics of tomorrow. We visualize results, summarize our findings and suggest a roadmap for future decision making.

Keywords: Data analysis · Data analytics · Big data · Questionnaire · Survey · Austria

1 Introduction

We are living in a data-saturated time. Continuous and large-scale methods for data analytics are needed as we now generate the impressive amount of 200 exabytes of data each year. This is equivalent to the volume of 20 million Libraries of Congress [1]. In 2012 each Internet minute has witnessed 100,000 tweets, 277,000 Facebook logins, 204 million email exchanges, and more than 2 million search queries fired to satisfy our increasing hunger for information [2].

This trend is accelerated technologically by devices that primarily generate digital data without the need for any intermediary step to first digitize analog data (e.g. digital cameras vs. film photography combined with scanning). Additional information is often automatically attached to the content (e.g. the exchangeable image file format 'Exif') that generates contextual metadata on a very fine-grained level. This means, when exchanging pictures, one also exchanges his or her travel destination, time (zone), specific camera configuration and the light conditions of the place, with more to come as devices evolve. Such sensors lead to a flood of machine-generated information that create a

M. Helfert et al. (Eds.): DATA 2014, CCIS 178, pp. 45–61, 2015.
DOI: 10.1007/978-3-319-25936-9_4

much higher spatial and temporal resolution than possible before. This 'Internet of Things' turns previously data-silent devices into autonomous hubs that collect, emit and process data at a scale that make it necessary to have automated information processing and analysis [1] to extract more value from data than possible with manual procedures. Today's enterprises are also increasing their data volumes. For example, energy providers now receive energy consumption readings from Smart Meters on a quarter-hour basis instead of once or twice per year. In hospitals it is becoming common to store multidimensional medical imaging instead of flat high-resolution images. Surveillance cameras and satellites are increasing in numbers and generate output with increasingly higher resolution and quality. Therefore, the focus today is on discovery, integration, consolidation, exploitation and analysis of this overwhelming data [1]. Paramount is the question of how all this (big) data should be analyzed and put to work. Collecting data is not an end but a means for doing something sensible and beneficial for the data owner, the business and the society at large. Technologies to harvest, store and process data efficiently have transformed our society and interesting questions and challenges have emerged of how society should handle these opportunities. While people are generally comfortable with storing large quantities of personal data remotely in a cloud there is also rising concern about data ownership, privacy and the dangers of data being intercepted and potentially misused [3].

In this paper, we present results of a study [4] that was conducted between June 2013 and January 2014 on the topic of (big) data analysis in Austria. Specifically, we present and highlight results obtained from an online survey that involved leading data scientists from Austrian companies and the public sector. The questionnaire was targeted to identify the status quo of (big) data analytics in Austria and the future challenges and developments in this area. We surveyed opinion from 56 experts and asked them about their understanding of data analytics and their projections on future developments and future research.

The paper first discusses related work and a status-quo of the potential application areas of (big) data in the next section. We then describe the method that was used for creating the questionnaire and for collecting and analyzing the feedback in Sect. 3. Results are presented and discussed in Sect. 4. In Sect. 5 we conclude and summarize our findings and suggest actions, in the form of a roadmap, that are based on our findings.

2 Related Work and State of (Big) Data Applications

Countries across the globe are eagerly developing strategies for dealing with big data. Prominent examples are the consultation process to create a Public-Private Partnership in Big Data currently underway in Europe,[1,2] work by the National Institute of Standards and Technology (NIST) Big Data Public Working Group[3]

[1] http://europa.eu/rapid/press-release_SPEECH-13-893_en.htm.
[2] All links have been accessed and validated in December 2014.
[3] http://bigdatawg.nist.gov.

as well as other groups [5] in the USA, and the creation of the Smart Data Innovation Lab[4] in Germany.

The recent and representative McKinsey report [6] estimates the potential global economic value of Big Data analytics between $3.2 trillion to $5.4 trillion every year. This value arises by intersecting open data with commercial data and thus providing more insights for customised products and services and enabling better decision making. The report identified the seven areas of education, transportation, consumer products, electricity, oil and gas, healthcare and consumer finance. We expanded this selection by specifically focusing on the Austrian market and its conditions before prompting participants with a comprehensive selection of application areas as described in Sect. 4.3. The remainder of this section will briefly review the scope of these application areas and why (big) data analytics could be interesting and helpful.

Healthcare has many stakeholders (i.e. patients, doctors, and pharmaceutical companies) each of which generate a wealth of data that often remains disconnected. In Austria there are also significant differences in how various federal states deal with patient data. There is great potential benefit in combining disconnected medical data sets for advancing healthcare, such as predictive modeling to improve drug development and personalized medicine to tailor treatments to individual genetic dispositions of patients [7, p. 37].

The *Energy and Utilities* sector now starts implementing the Smart Grid technology that allows recording energy consumption constantly every 15-minutes (i.e. Smart Metering) compared with manual monthly readings. This enables energy companies to better understand usage patterns and to adapt services and the energy grid to changing demands. Ecova[5], an energy and sustainablity management company, recently evaluated its energy data and published some interesting trends in consumer energy consumption [8]. They found that between 2008 and 2012, energy consumption across the US dropped by almost 9 % while water prices increased by almost 30 %. This demonstrates how (big) data analytics can help inform about large-scale changes in our society. There are however still many technical problems, primarily about privacy and security, and the issue that customers have only limited and slow access to their own consumption information. This were some of the reasons why Austria enabled consumers to opt-out from smart meters[6].

EScience is commonly seen as the extended and more data-intensive form of science supported by its increased capacity to generate and transform data [9]. An European Commision Report [10] describes the general infrastructure requirements, such as providing a generic set of tools for capture, validation, curation, analysis, and the archiving of data, the interoperability between different data sources, the (controlled and secured) access to data. Furthermore, incentives should be given to contribute data back to the infrastructure strengthened with

[4] http://www.sdil.de/de/.
[5] http://www.ecova.com.
[6] http://www.renewablesinternational.net/smart-meter-rollout-in-europe-rollback-in-austria/150/537/72823/

a range of financial models to ensure the sustainability of the infrastructure. A first basic e-Infrastructure is currently created in Austria to provide archiving services for scientific publications and data.[7]

Manufacturing and Logistics traditionally generates large data sets and requires sophisticated data analysis methods. This has developed further in recent years when moving to the fourth industrial revolution[8]. There are many applications of data analytics, such as Manufacturing Systems, Fault Detection, Decision Support, Engineering Design or Resource Planning, as described in [11]. These applications remain relevant and now further expand into big data applications that are more complex, more (socially) connected, and more interactive. The McKinsey report [6] foresees the (big) data applications in manufacturing to be in R&D, supply chain and in functions of production as a modern expansion of Harding's categories.

The sector of *Telecommunications* generally works with large data — as of May 2014, we have nearly 7 billion mobile subscriptions worldwide that represent 95.5 percent of the world population [12]. Each of these mobile phones transmit data every time we call, text, or access the internet. Moreover, mobile phones send passive information, such as when handing over between antennas and when estimating geo-positions. For this reason, telecommunications have long established big data at the core of their business model although they focus more on the real-time aspect of data and the customer as the target of their data analytics.

Transportation undergoes many changes as the economy moves into the information age where product-related services turn into information-related services. The growing amount of data, especially real-time data, needs to be managed and applied in such applications as price optimisation, personalisation, booking and travel management, customer relationship management [13]. In Austria, predictive data analysis is increasingly used for transportation incident management and traffic control, however at an initial stage.

Education can benefit from intelligent data analytics through Educational Data Mining [14] that develops methods for exploring data from educational settings to better understand students. These methods span across disciplines like machine learning, statistics, information visualization and computational modelling. There are initial efforts to include them in Moodle [15] in an attempt to make them broadly accessible.

Finance and Insurance deal with large data volumes of mostly transactional data from customers. The goal is to move such data from legacy storage-only solutions to distributed file systems, like Hadoop. This allows fraud detection [16] and custom-tailored product offers. An IBM study for example identified a 97 % increase in the number of financial companies that have gained competitive advantages with data analysis between 2010 and 2012 [17].

The gains for the *Public Sector and Government Administration* are mostly in tax and labour services and are benefitting both the people and the governments

[7] http://www.bibliothekstagung2013.at/doc/abstracts/Vortrag_Budroni.pdf.
[8] http://en.wikipedia.org/wiki/Industry_4.0.

that represent them. Citizens benefit from shorter paper trails (e.g. for re-entering data) and waiting times (e.g. for tax returns). At a European level, the public sector could reduce administrative costs by 15 % to 20 % and create an extra of €150 to €300 billion value according to [7]. Austria has taken essential steps for a more efficient public sector. The cities of Vienna, Linz and Graz for example provide much of their data as part of the Open Government Initiative[9] so that companies can develop applications that ease daily life. Electronic filing started in 2001 (ELAK-der elektronische Akt) to replace paper based filing and archiving in all Austrian ministries [18]. The system shortens the paper processing times by up to 15 % and already has over 9,500 users. However, these solutions often exist in isolation and there is much need to integrate these data silos and strengthen them with additional analytical capabilities.

While *Commerce and Retail* is traditionally involved in activities of business intelligence it now faces a scenario that shifts from the problem of actively collecting data to selecting the relevant from too much data available. One pioneer use of data analytics was the stock market where it made some a fortune and possibly also took part in stock market crashes [19]. Retail benefits significantly from (big) data analytics [7, p. 64–75] where customers have more options to compare and shop in a transparent form. Retailers make informed decisions by mining customer data (e.g. from loyalty cards and mobile applications).

Tourism and Hospitality deals with big data and can be an application for data analytics as tourists leave a rich digital trail of data. Those businesses who best anticipate the customers expectations will be able to sell more "experiences". In [20], an example is given of how tourism statistics combined with active digital footprints (e.g. social network postings) provide a representation of the tourist dispersion among sites in Austria.

Law and Law Enforcement deals with millions of cases and service calls every year. It was in Santa Cruz, California, where data from 5000 crimes was first applied to predict, and then prevent, upcoming offences. This might remind some of 'Minority Report' and Hunter S. Thompson would certainly describe it as a situation of "a closed society where everybody's guilty [and] the only crime is getting caught" [21, PartI,9]. However, the programme did, in a first round, reduce crime incidents by up to 11 % [22]. Financial crime detection also benefits from data analytics by processing and applying predictive methods on transactions [23]. In Austria concrete steps in this direction where taken by deciding that courts of law must use the electronic file system ELAK, as described in [18].

There are other areas that engage in (big) data analytics that have not been considered in the scope of this survey: In *Earth Observation*, many terabytes of data are generated and processed every day to forecast weather, predict earthquakes and estimate land use. *Agriculture* has drastically changed in the past half century by combining weather/environment data and data from machines and agricultural business processes for increased efficiency. In [7], *Media and Entertainment* is described as less capable for capturing the value of big data

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

despite being IT intense. Big data can also be found in *Sports*, *Gaming*, and *Real Estate*. Let us not forget *Defence and Intelligence* — an area that most likely started the idea of collecting and correlating data from the masses.

Many other surveys have been conducted on the topic of big data and big data analytics by consulting companies, but these surveys usually concentrate on large enterprises.[10] A summary of the 2013 surveys is available [24]. A survey among people defining themselves to be Data Scientists has also been conducted to better define the role of Data Scientists [25]. Here, we consider the views of mostly academic scientists working in multiple areas related to data analytics, and hence we provide an unusual "academic" view of this emerging new field.

3 Method

Surveys are powerful tools when collecting opinion from the masses. Our main objective was to further specify our understanding of data analytics in Austria and to identify future challenges in this emerging field.

We followed the strategy of active sampling. The identification of Austrian stakeholders in data analytics formed the starting point: We first scanned and reviewed Austrian industry and research institutions based on their activities and research areas. We then identified key people from these institutions and asked about their opinions, attitudes, feedback and participation during a roadmapping process.

Our final contact list comprised 258 experts, all of them senior and visible data scientists, that we contacted twice and invited them to complete our questionnaire. This means our contact list has consensus-quality and represents the current situation and strength of senior data scientists in Austria. The survey was online between the beginning of September 2013 until the middle of October 2013. A total of 105 people followed the link to the survey resulting in a general response rate of 39 %. However, several of them turned down the questionnaire or cancelled their efforts after only one or two questions. We took a strict measure and removed those incomplete cases from the list of responses to increase the quality of the data. This reduced the original 105 responses (39 %) further down to 56 responses (21.7 %).

The general advantages of online surveys, such as truthfulness, increased statistical variation and improved possibilities for data analysis (e.g. [26,27]), unfortunately suffer from the problems of limited control, a higher demand on participants in terms of time and patience and the potential that people may be engaged in other, distracting activities that alter the results and increase the dropout rate (e.g. [28]). While our response rate of nearly 40 % is normal for

[10] Some examples: http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html, http://www.sas.com/resources/whitepaper/wp_58466.pdf, the Computing Research Association (CRA) http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf and SAS http://www.sas.com/resources/whitepaper/wp_55536.pdf.

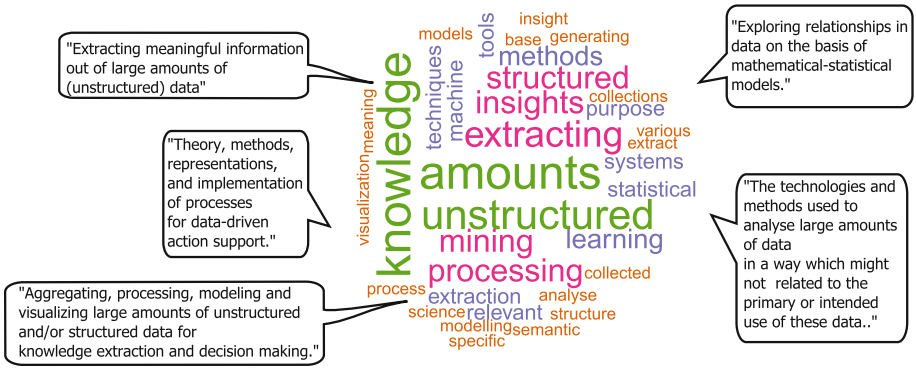

Fig. 1. What participants understood as data analytics.

online surveys [26], the high dropout rate in our specific case can be attributed to the complex nature of the subject.

The data of the survey was collected anonymously with LimeSurvey[11] and was analyzed with R, a statistical software package[12].

4 Survey Results and Discussion

This section highlights the results we obtained from the data and focuses on four areas. First, the demographic information about the participants (e.g. their age, area of work, and their work experience) that helps us to get a better understanding of the characteristics of a typical data scientist in Austria. Second, we look at the application areas of data analytics and how participants projected the future relevance of these areas. Third, we investigated the future challenges of data analytics. Fourth, we analysed free text submissions from questions about research priorities and the need for immediate funding to get a better understanding of possible future directions, interests and desires. We omitted replies for questions that only had a very limited text response that would not be meaningful to analyze statistically. The survey questions are provided in detail in the appendix of [29].

4.1 Participants: Our Data Scientists

The data presented in this paper is based on the opinions of 56 people who completed the questionnaire — four female (7.1 %) and 52 male (92.9 %). This gender distribution is similar in the original contact list — 26 female (10.1 %) and 232 male (89.9 %) — and therefore represents the current gender situation in

[11] http://www.limesurvey.org/de/.

[12] http://www.r-project.org/. The survey questions are provided in the appendix of [29].

the data science profession in Austria. Participants were mostly Austrians (96 %) and the majority of them were working in the research and academic sector. About a fifth (21.4 %) of all responses came from the industry. The larger part worked for academic (55.4 %) or non-industry (33.9 %) research organisations.[13] The majority of participants (80.3 %) had an extended experience of nine or more years. This defines our sample as a group of mostly academic, male, and Austrian data scientists.

4.2 What is Data Analytics? A Definition

We asked participants to describe the term 'Data Analytics' in their own words as an open question to get an idea about the dimensions of the concept and the individual views on the subject. Figure 1 depicts a summary word cloud from the collected free-text responses for all those terms that repeatedly appeared in the response.[14] It further depicts a small set of representative extracts from the comments and definitions that participants submitted. Overall, the comments were very much focused on the issue of large data volumes, the process of knowledge extraction with specific methods and algorithms and the aggregation and combination of data in order to get new insights. Often it was related to machine learning and data mining but as a wider and more integrative approach. Only very few respondents labeled Data Analytics to be simply a modern and fashionable word for data mining or pattern recognition.

4.3 Important Application Areas

Based on the literature review that preceded this survey, we identified the main application areas of data analytics in Austria as healthcare, commerce, manufacturing and logistics, transportation, energy and utilities, the public sector and the government, education, tourism, telecommunication, e-science, law enforcement, and finance and insurance. Figure 2 shows the relative importance of these areas as attributed by participants. Selections were made in binary form with multiple selections possible. The figure shows that the area of healthcare is perceived as a strong sector for data analytics (66.1 %) followed by energy (53.6 %), manufacturing and e-science (both 50.0 %). As a sector that is perceived to benefit only little from (big) data analytics are tourism and commerce (both 23.2 %). This is despite the fact that these areas are large in Austria based on demographic data as provided by Statistics Austria.[15]

[13] Multiple selections were possible which means that these numbers do not add up to 100 %.

[14] We only included terms that appeared a least three times and we filtered with an english and a topical stop word list (e.g. terms like 'and' or 'etc' and terms like 'data' or 'analytics').

[15] In 2010, 19.3 % of the employed worked in commerce and 9.1 % in the gastronomical and leisure sector (source: 'Ergebnisse im Ueberblick: Statistik zur Unternehmensdemografie 2004 bis 2010', available at http://www.statistik.at/web_de/statistiken/unternehmen_arbeitsstaetten/arbeitgeberunternehmensdemografie/index.html, extracted 09-12-2014.

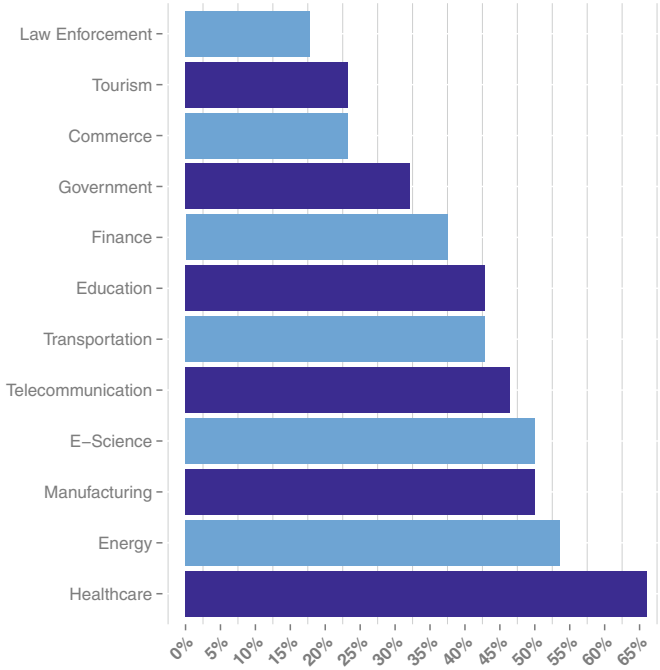

Fig. 2. Important application areas for data analytics.

We additionally asked participants to provide future projections for these application areas w.r.t. how they think these areas would become important for data analytics in the future (see Fig. 3). Here, participants rated the application areas based on their relevance for the *short*, *middle* and *long term* future. The diagram also visualizes the amount of uncertainty in these projections as participants could select if they were unsure or even declare an application area as unimportant. The figure shows that application areas that are perceived as strong candidates (e.g. healthcare, energy and telecommunication) are all marked as relevant for the short term future with decreasing ratings on the longer timeline. Less strongly perceived application areas, such as law enforcement and tourism have results that are less clearly expressed with a stronger emphasis on a longer time frame. The amount of uncertainty about these areas is also much higher. Law enforcement is perceived as both less important and not benefiting from data analytics. This is conceivable as law enforcement may not be perceived as an independent sector, as this is the case in the United States [30] where data analytics already assists the crime prediction process with data mining, e.g. with the use of clustering, classification, deviation detection, social network analysis, entity extraction, association rule mining, string comparison and sequential pattern mining. It comes as a surprise that tourism in Austria was both perceived as rather unimportant and also as an area that would only benefit from data analytics in the mid- and long-term future. The large proportion of uncertainty shows that experts seem to be rather unsure about the future of these two sectors.

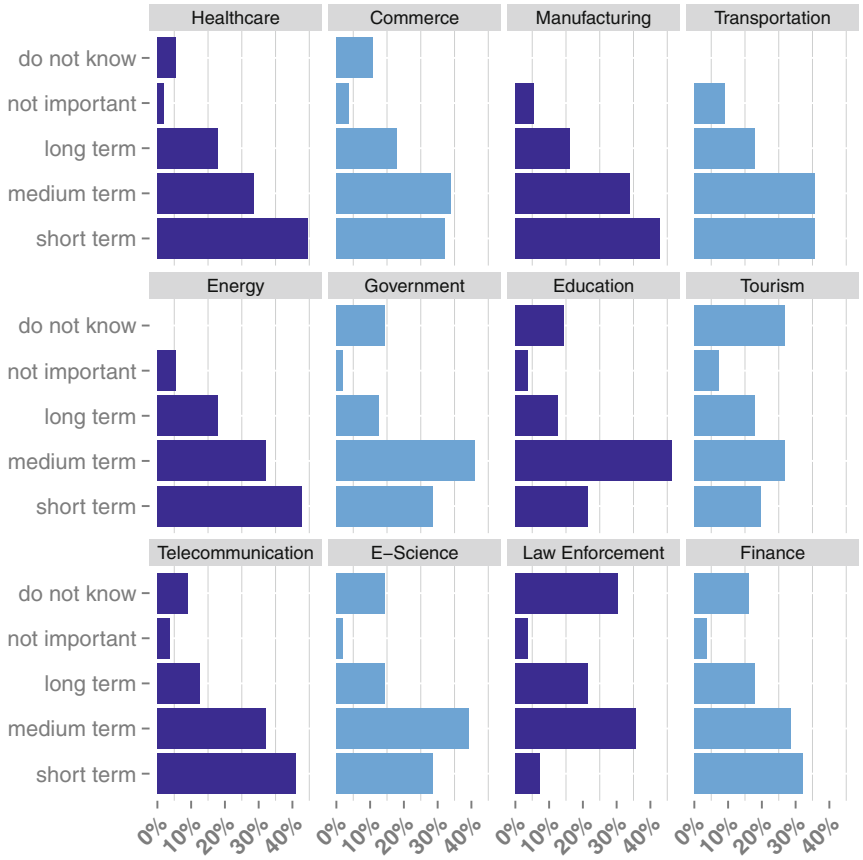

Fig. 3. Application areas where data analytics will become important in the short, middle and long term.

4.4 Current and Future Challenges

Based on the literature review, we identified the main challenges of data analytics in the areas of privacy and security, algorithms and scalability, getting qualified personnel, the preservation and curation process, the evaluation and benchmarking, and data ownership and open data. We now asked participants to categorize these challenges into three groups: *Short term* if they see it as an issue of the very near future, *medium term* if there is still time and *long term* if this might become an issue some time in the far future. Our intent was to obtain a priority that can help us to identify possible actions and recommendations for decision making. Figure 4 depicts all responses for all categories and also includes the amount of uncertainty (*do not know*) and how unimportant people thought it to be (*not important*).

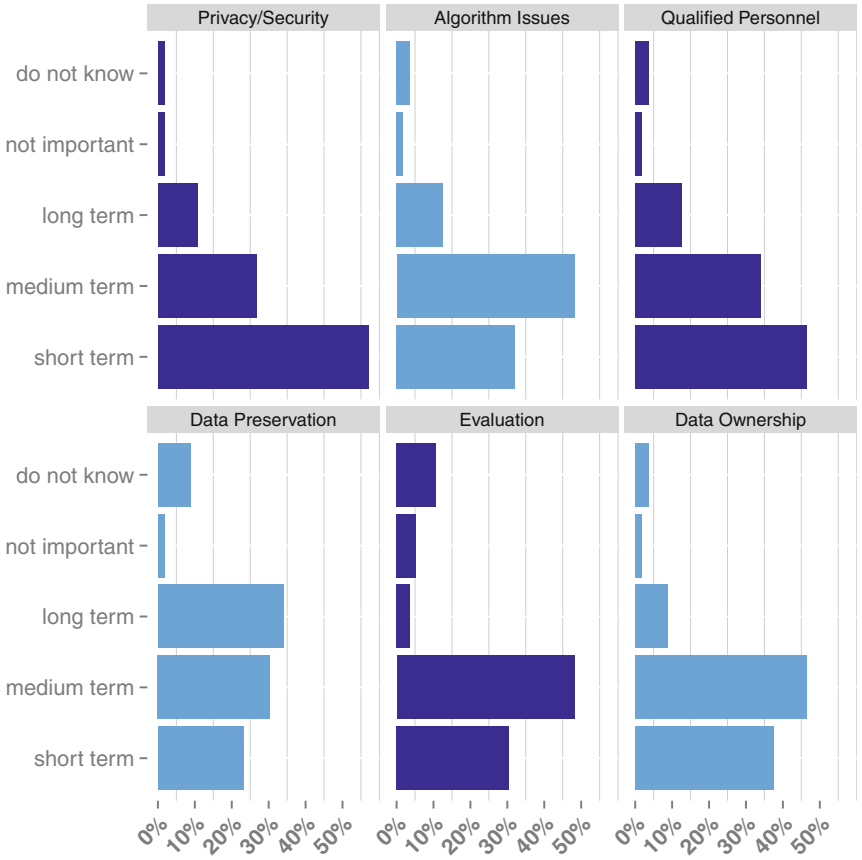

Fig. 4. Future challenges in data analytics for the short, middle and long term.

All challenges share that they are perceived as all being relevant in the short and middle-term future, with high certainty throughout. Upon closer investigation, the response can then be further divided into two groups.

The first group of challenges consists of privacy and security and the issue of qualified personnel. These issues are perceived as considerably more important and pressing in the short term future than the middle and the long term future. It was especially striking how important *privacy and security* was emphasised throughout the entire study — including this online questionnaire. This strong impact might be partially attributed to the very recent NSA scandal. However, it might also be that our target group quite naturally possesses a heightened sensitivity to the potential dangers of (big) data analytics and the often unprotected flow of personal data on the open web. *Qualified personnel* is a problem of the near future and has been discussed in the literature throughout many studies. This is well confirmed in our own findings and an important issue to address in future decision making.

Table 1. Comparison of Big Data Challenges as identified in three different studies.

CRA study	SAS study	Our study
Lack of skills/Experience	-	Personnel
Accessing data/Sharing	Human collaboration	Data ownership/Data preservation
Effective use	-	-
Analysis and understanding	Data heterogenity	Algorithm issues
Scalibility	Scalibility/Timeliness	Evaluation
-	Privacy	Privacy and security

The second group of challenges covers algorithmic and scalability issues, data preservation and curation, evaluation, and data ownership and open data. All of these were attributed more frequently to be issues of the future. Ironically, data preservation and curation has been attributed with being more relevant in the long-term future than the mid- and short-term with the highest amount of uncertainty in the entire response. This should ideally be the opposite. We would have also expected that data ownership and open data issues would be categorized very similar to the privacy and security response and that the algorithmic issues are more relevant on a short term scale as data is mounting very fast. The responses nevertheless demonstrate the feeling that the privacy and security and qualified personnel challenges need to be solved before progress can be made in the field.

We additionally compared our list of challenges with those that were identified in two related, recent studies: One study hosted by the Computing Research Association (CRA) that focused on Challenges and Opportunities of Big Data in general[16] and one study on Big Data visualization by SAS.[17] In Table 1, we refer to them as 'CRA Study', 'SAS Study' and 'Our Study' and compare six challenge categories that were identified across these studies. The challenges are presented in no particular order, however, the reader can compare challenge categories horizontally in the table. A dash (-) means that a particular challenge was not identified by a study. We related the categories with each other to give the reader an overview about the similarities and differences from three perspectives. Naturally, the categories did not always represent a perfect match. For example, the challenge of data access and data sharing was addressed as the need for human collaboration in the SAS study and our own study identified the challenge of data ownership and the challenge of preserving data in this category. However, the issue of a lack of personnel was also identified as a lack of skills and experience in the CRA study. Whereas privacy was clearly addressed in the SAS study and more comprehensively combined with security in our study, the CRA study did not consider it a challenge at all. Overall, this comparison shows that there is considerable agreement between studies with respect to future

[16] http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf.
[17] http://www.sas.com/resources/whitepaper/wp_55536.pdf.

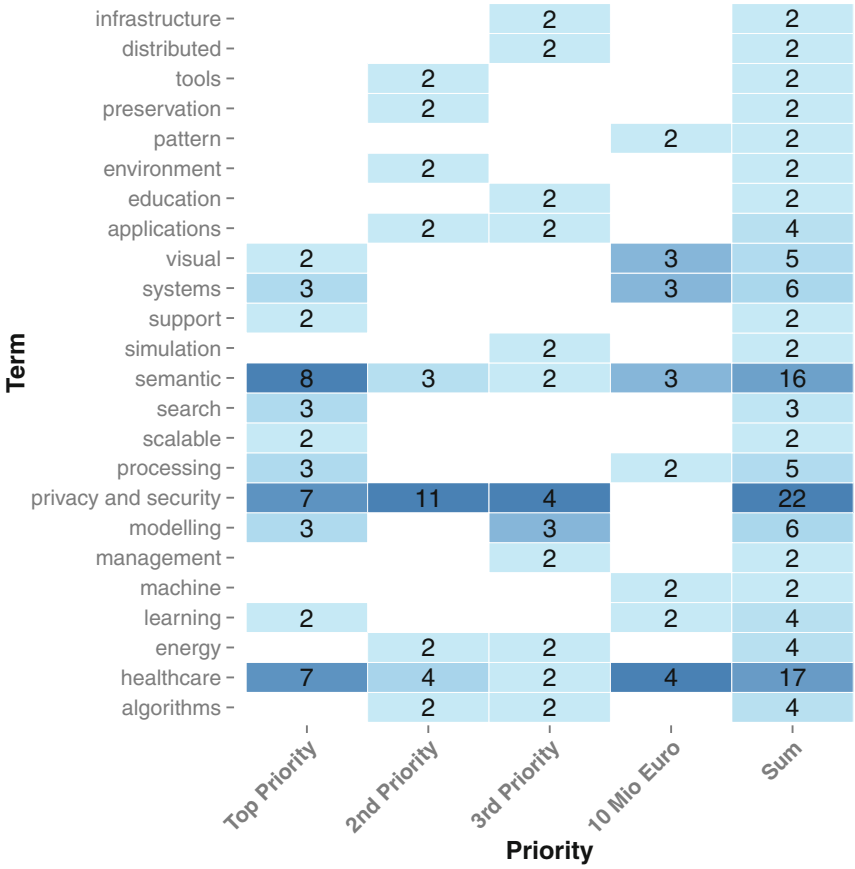

Fig. 5. Priorities for future research as expressed by participants.

challenges. It would be interesting to further extend this comparison to a much wider range of studies in future work.

4.5 Future Research Topics

We prompted participants with two questions about future research in (big) data analytics. The first question asked them to enter free text on topics of their preferred future research which they had to prioritise by three levels (top priority, 2nd priority, 3rd priority). The second question can be seen as a refinement of the top priority level of the previous question and asked them to describe which research topic they would like to see publicly funded with 10 Million €. Again, this was submitted as free text allowing participants to contribute their ideas in a completely free and unrestricted form. Figure 5 shows the term frequencies of those texts for all three priorities and also the text for the

10 Million € research topic. This allows for easy comparisons. A sum across all four columns of frequencies provides an overview about the entire topic space.

The most frequent themes are privacy/security (mentioned 22 times) and healthcare (mentioned 17 times) which coincides with the findings from the previous questions. The importance of privacy and security was found to be the most pressing future challenge (see Fig. 4). Likewise, healthcare was also perceived as the most promising application domain (see Fig. 2) with the most pressing time line that strongly leans toward the short-term future (see Fig. 3). The third most frequent keyword were semantic issues (mentioned 16 times) that were more extensively investigated in a number of workshops and an expert interview that is documented in more detail in [4].

5 Conclusions and Technology Roadmap

This paper presented results from a study on (big) data analysis in Austria. We summarized the opinions of Austrian data scientists sampled from both industry and academia and some of the most pressing and current issues in the field.

We found that data analytics is understood as dealing with large data volumes, where knowledge is extracted and aggregated to lead to new insights. It was interesting to see that it was often related to data mining but viewed more widely and more highly integrated. Healthcare was seen as the most important application area (66.1 %), followed by energy (53.6 %), manufacturing and logistics (50.0 %) and e-science (50.0 %) with big potential in the short term future. Other areas were judged less important, such as tourism (23.2 %) and law enforcement (17.9 %), with high uncertainty. We found that our literature-informed list of challenges were confirmed by our respondents, however only privacy/security and the challenge to get qualified personnel was strongly attributed to the very near future. Algorithm issues, data preservation, evaluation and data ownership were seen as challenges that become more relevant only in the longer run. Research priorities and funding requests where strongly targeted to privacy/security (mentioned 22 times), healthcare (mentioned 17 times) and semantic issues (mentioned 16 times). This result conforms largely to the findings in the other parts of the study.

Based on the results of the survey presented in this paper, along with the outcomes of three workshops and interviews, a technology roadmap consisting of a number of objectives was drawn up. The roadmap actions are described in much greater detail in [4] as part of the complete report that focuses on all parts of the study. This is outside the scope of this paper that focuses on the details of the online survey. In summary, the identified challenges, together with their careful evaluation, have led to three categories of actions that are manifested in this roadmap.

First, to meet the challenges of data sharing, evaluation and data preservation, an objective in the roadmap is to create a “Data-Services Ecosystem” in Austria. This is related to an objective to create a legal and regulatory framework that covers issues such as privacy, security and data ownership, as such a framework is

necessary to have a functioning Ecosystem. In particular, it is suggested to fund a study project to develop the concept of such an Ecosystem, launch measures to educate and encourage data owners to make their data and problems available, and progress to lighthouse projects to implement and refine the Ecosystem and its corresponding infrastructure. Furthermore, it is recommended to develop a legal framework and create technological framework controls to address the pressing challenges of privacy and security in data analytics.

Second, technical objectives are to overcome challenges related to data integration and fusion and algorithmic efficiency, as well as to create actionable information and revolutionise the way that knowledge work is done. We suggest to fund research that focuses on future data preservation, to develop fusion approaches for very large amounts of data, to create methods that assure anonymity when combining data from many sources, to enable real time processing, and to launch algorithmic challenges. A full list of suggestions are described in more detail in [4].

The third and final objective in the roadmap is to increase the number of data scientists being trained. We suggest a comprehensive approach to create these human resources and competences through educational measures at all levels: from schools through universities and universities of applied sciences to companies. The issue of having more and highly skilled data scientists soon is an issue that requires immediate action to secure the future prosperity of the Austrian (big) data analytics landscape.

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) as FFG ICT of the Future project number 840200. We thank Andreas Rauber for his valuable input. Information about the project and access to all deliverables are provided at http://www.conqueringdata.com.

References

1. Dandawate, Y. ed.: Big Data: Challenges and Opportunities. Infosys Labs Briefings., vol. 11, Infosys Labs (2013). http://www.infosys.com/infosys-labs/publications/Documents/bigdata-challenges-opportunities.pdf. Last visited, December 2014
2. Temple, K.: What Happens in an Internet Minute? (2013). http://scoop.intel.com/what-happens-in-an-internet-minute/. Last visited, December 2014
3. Boyd, D., Crawford, K.: Critical questions for big data. Inf. Commun. Soc. **15**, 662–679 (2012)
4. Berger, H., Dittenbach, M., Haas, M., Bierig, R., Hanbury, A., Lupu, M., Piroi, F.: Conquering Data in Austria. bmvit (Bundesministerium für Verkehr, Innovation and Technology, Vienna, Austria (2014)
5. Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J., Jagadish, H.V., Labrinidis, A., Madden, S., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., Ross, K., Shahabi, C., Suciu, D., Vaithyanathan, S., Widom, J.: Challenges and Opportunities with Big Data (2012). http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf. Last visited, December 2014

6. Manyika, J., Chui, M., Groves, P., Farrell, D., Kuiken, S.V., Doshi, E.A.: Open Data: Unlocking Innovation and Performance with Liquid Information. McKinsey Global Institute, Kenya (2013)
7. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, Kenya (2011)
8. Hardesty, L.: 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 2014
9. 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 2014
10. Riding the Wave: How Europe can gain from the rising tide of scientific data. European Commission (2010)
11. Harding, J.A., Shahbaz, M., Kusiak, A.: Data mining in manufacturing. Rev. J. Manuf. Sci. Eng. **128**(4), 969–976 (2006)
12. ICT: The world in 2014: ICT facts and figures (2014)
13. Ellis, S.: 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, December 2014
14. Baker, R., Yacef, K.: The state of educational data mining in 2009 a review and future visions. J. Educ. Data Min. J. Educ. Data Min. **1**, 3–17 (2009)
15. Romero, C., Espejo, P., Zafra, A., Romero, J., Ventura, S.: Web usage mining for predicting final marks of students that use Moodle courses. Comput. Appl. Eng. Educ. **21**, 135–146 (2013)
16. The Economist: Big data: Crunching the numbers (2012). http://www.economist.com/node/21554743. Last visited, December 2014
17. Turner, D., Schroeck, M., Shockley, R.: Analytics: The real-world use of Big Data in financial services. IBM Global Business Services, Executive report (2013)
18. Müller, H.: ELAK, the e-filing system of the Austrian Federal Ministries (2008). http://www.epractice.eu/en/cases/elak. Last visited, November 2013
19. 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 2014
20. Koerbitz, W., Önder, I., Hubmann-Haidvogel, A.: Identifying tourist dispersion in austria by digital footprints. In: Cantoni, L., Xiang, Z.P. (eds.) Information and Communication Technologies in Tourism 2013, pp. 495–506. Springer, Heidelberg (2013)
21. Thompson, H.S.: Fear and Loathing in Las Vegas: A Savage Journey to the Heart of the American Dream. Modern Library, Vintage Books (1971)
22. Olesker, A.: White Paper: Big Data Solutions For Law Enforcement, IDC White paper (2012)
23. Mehmet, M., Wijesekera, D.: Data analytics to detect evolving money laundering 71–78. In: Laskey, K.B., Emmons, I., Costa, P.C.G. (eds.) Proceedings of the Eighth Conference on Semantic Technologies for Intelligence, Defense, and Security, STIDS 2013, CEUR Workshop Proceedings, vol. 1097, pp. 71–78 (2013)
24. Press, G.: The state of big data: What the surveys say (2013). http://www.forbes.com/sites/gilpress/2013/11/30/the-state-of-big-data-what-the-surveys-say/. Last visited, December 2014

25. Harris, H., Murphy, S., Vaisman, M.: Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work. O'Reilly, USA (2013)
26. Batinic, B.: Internetbasierte befragungsverfahren. Österreichische Zeitschrift für Soziologie **28**, 6–18 (2003)
27. Döring, N.: Sozialpsychologie des Internet. Die Bedeutung des Internet für Kommunikationsprozesse, Identitäten, soziale Beziehungen und Gruppen. 2nd edn. Hogrefe, Göttingen (2003)
28. Birnbaum, M.H.: Human research and data collection via the internet. Annu. Rev. Psychol. **55**, 803–832 (2004)
29. Bierig, R., Hanbury, A., Haas, M., Piroi, F., Berger, H., Lupu, M., Dittenbach, M.: A glimpse into the state and future of (big) data analytics in austria - results from an online survey. In: DATA 2014 - Proceedings of 3rd International Conference on Data Management Technologies and Applications, Vienna, Austria, 29–31 August, 2014, pp. 178–188 (2014)
30. Norton, A.: Predictive policing - the future of law enforcement in the trinidad and tobago police service. Int. J. Comput. Appl. **62**, 32–36 (2013)