Conquering Data:
The State of Play in Intelligent Data Analytics

Ralf Bierig, Florina Piroi, Mihai Lupu, Allan Hanbury
Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria
Helmut Berger, Michael Dittenbach, Marita Haas
max.recall information systems GmbH, Vienna, Austria

12 September 2013

This paper is the basis for the creation of a research roadmap for *Conquering Data* in Austria. It summarises techniques for Intelligent Data Analytics, presents applications in which the analysis of data has the potential to generate significant value, discusses research challenges in the Intelligent Data Analytics area, and presents an Intelligent Data Analytics Competence Landscape of Austria. Further information on this work is available online at `http://conqueringdata.at/`

Contents

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. According to Intel, each internet minute sees 100,000 tweets, 277,000 Facebook logins, 204 million email exchanges, and more than 2 million search queries fired [60]. 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 extra 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 and analyse this data in order to obtain insight to guide action or decision making. Hence there is a clear need for automation [24].

There is no dearth of data for today's enterprises. On the contrary, they are mired in data and quite deeply at that. Today, therefore, the focus is on discovery, integration, exploitation and analysis of this overwhelming information. 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 that hopefully proves to be beneficial for the data owner, the business and the society at large. But with huge amounts of data, it is easy to find correlations that may not be related in a causal way.

These changes 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 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 [18].

We take the position that effective Intelligent Data Analytics has the potential to greatly benefit the Austrian society and economy and that it is essential for a successful innovation economy to understand how to effectively analyze such data resources. Nevertheless, 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.

This position paper is the first step in the development of an Intelligent Data Analytics research roadmap for Austria and forms the basis for further discussion in workshops and interviews with stakeholders. During these workshops and interviews, the applications and challenges discussed at a more international level in this paper will be made more specific to the context in Austria.

We begin this paper with a definition of Intelligent Data Analytics and list the techniques that fall into this domain. We then summarise the application areas in which Intelligent Data Analytics has made an impact internationally or has the potential to make an impact. The open issues and challenges for Intelligent Data Analytics, particularly those in which additional research is required, are listed and discussed. Finally, we give an overview of the Intelligent Data Analytics Research and Development (R&D) landscape in Austria.

2 Conquering Data with Intelligent Data Analytics

The art of Conquering Data with Intelligent Systems includes all 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 from (raw) data. This requires methods ranging from learning, inference, prediction, knowledge discovery and visualisation that are applicable on 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 automated hypothesis generation, event correlation, and anomaly detection and helps in explaining phenomena and inferring results that would otherwise remain hidden [38]. Intelligent Data Analytics is a cornerstone in modern Big Data, amplifying perhaps its most important aspect: Value.

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 Representation**. 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 [42], but also based on an analysis of the R&D in data analysis currently 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. As the quantitative evaluation of algorithm and system performance is important, **evaluation and benchmarking** is also discussed. Figure 1 presents a graphical overview of these groups, along with the application areas from Section 3. The techniques and their grouping form the basis for the analysis of the Austrian R&D landscape presented in Section 5.1. 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.

2.1 Search and Analysis

Search and Analysis is the domain of searching and analyzing multimodal data (text, image, audio and voice, and video), and the aggregation and fusion of such multimodal data streams in real time. The results of the analysis range from preparing the data for further semantic processing or as input for cognitive systems, to discovering interesting patterns or relationships in the data.

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.

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, organization, 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.

Bioinformatics is an interdisciplinary field that develops and improves on methods for storing, retrieving, organizing and analyzing biological data.

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.

Algorithmic efficiency refers to developing efficient algorithms able to process larger amounts of data in a shorter time. It includes the development of **parallel algorithms**, **optimisation** and **grid computing**.

2.2 Semantic Processing

Semantic Processing adds structure and “meaning” to data, whilst making it accessible for machine learn-

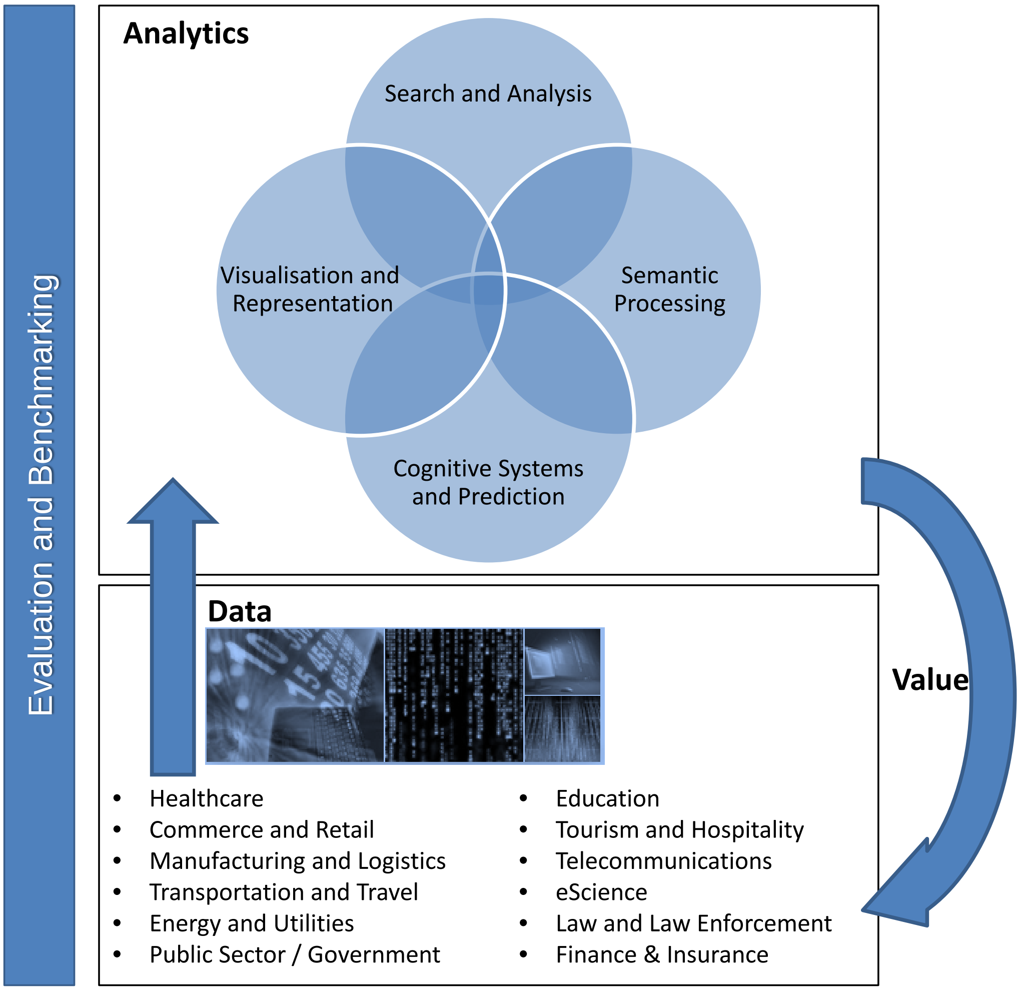

Figure 1: *Relationship of Intelligent Data Analytics techniques and application domains.*

ing 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 [26].

The Semantic Web aims at the creation of a "web of data" by encouraging the inclusion of semantic content in web pages, allowing users to find, share, and combine information more easily.

Natural Language Processing uses algorithms to analyze human natural language [42]. It also includes the area of **sentiment analysis**.

2.3 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 given 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.

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 [22]. It is distinct from machine learning in 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 information gathering and research.

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) [54].

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.

2.4 Visualisation and Representation

Visualisation creates a human-accessible interface to Data Analytics. It supports exploring and understanding data more easily (especially when volumes grow) and helps discovering patterns, thus making data more accessible for specialists and, most importantly, the general public. Non-visual representations are also possible.

Visualisation is any technique for creating images, diagrams, or animations to communicate a message. Areas of interest for Data Analytics include **scientific visualisation**, **information visualisation** and **knowledge visualisation**.

Visual analytics focuses on human interaction with visualisation systems as part of a larger process of data analysis.

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 [40]. Auditory perception has advantages in temporal, amplitude, and frequency resolution as an alternative or complement to visualisation techniques.

2.5 Evaluation and Benchmarking

All types of algorithms above need to be evaluated or benchmarked in order for their suitability of purpose to be quantitatively measured, and hence to assist in the choice of a good algorithm for a given practical problem. Algorithms can be evaluated individually or as systems of multiple algorithms. It is usually difficult to impossible to extrapolate from the performance of individual algorithms to the performance of a system using these algorithms, due to dependencies between the algorithms. *Empirical evaluation* looks only at the performance of the algorithm or system without human intervention, while *user-centred evaluation* measures the performance of an algorithm or system while being used by an end-user. Many methodologies exist for evaluation, and methods from statistics are commonly used to estimate the significance of the results obtained.

3 Intelligent Data Analytics Applications

Austrian companies focus on a range of economic areas. For the position paper, we reviewed some of these areas with respect to how they currently handle and use Intelligent Data Analytics for Conquering Data. We chose an international viewpoint to provide a more holistic overview to the reader.

3.1 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, the economic environment, etc.

One of the first areas to 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 [57].

Data Analytics is having a significant effect on retail [42]. 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 price transparency, and allows the customers to put pressure on prices. Available data is growing also on the retailer side, not least 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 and the use of weather and social media information to adjust offer and pricing to trends.

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.

3.2 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 [42] and is continuously growing at the speed of RFID tags sold[1] Data is mostly used to ensure quality and efficiency in the production process.

Based on these large data sets, analytics has also become a long-standing tradition in manufacturing. Harding, in [35], summarises research in analytics[2] that dates back to the mid-80s 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 organizational boundaries [42]. 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 lifecycle by better integrating the customer and by combining into a better whole.

3.3 Transportation and Travel

The travel and transportation industry faces challenges and opportunities as the economy moves into an information age [23]. 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, organizational, process, and decision management frameworks will be required to cope with the volume of data generated across the industry [25]. This includes applications such as: price optimization for the transportation and travel industry commodities (e.g. airplane seats); offer personalisation based on customer purchasing history and preferences; efficient booking and travel management for corporations and organizations; 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.

Data Analytics of primarily large volume, unstructured data plays a vital role in delivering a more efficient and tailored travel experience with benefits

[1] RFID tag sales are projected to rise from 12 million pieces to 209 billion between 2011 and 2021 according to [42].

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

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; called TPF for Transaction Processing Facility, they were developed by IBM in the 1960s and 70s, and have been refined ever since.

3.4 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 [33]. 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 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 to not so much measure customer satisfaction and respond to complaints as to design, implement and assess entire customer travel experiences.

Data Analytics suggest that 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. Other industries have shown, however, that in a digital world, past success is no guarantee of future business growth and vitality. Identifying preferences and affinities may become as important to travel and hospitality companies seeking customer loyalty as being able to provide these services themselves [10].

3.5 Healthcare

The healthcare sector consists of many stakeholders, including the pharmaceutical and medical products industries, healthcare providers, health insurers and patients. Each generates pools of data, which have typically remained disconnected [42]. 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 [6]. The cost of sequencing a human genome has already dropped below US$1000, 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 [42], especially through so-called secondary use of this information (uses beyond providing direct health care). The following are some of the ways in which the use Data Analytics in health can lead to significant savings in health care expenditure [42]: *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 Modelling* 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 [48]. 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" [28]. 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.

3.6 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 [61]. Many traditional systems store this data for years without being able to analyse it. In the last years, though, financial institutions recognised the competitive advantage they can gain when making use of the information they store on their customers. A recent IBM study [61] 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. In Austria, paysafecard, for example, successfully introduced last year the use of big Data Analytics solutions to be able to respond in real-time to transaction requests and to recognise patterns of fraud attempts.

For both commerce institutions and banks, with plenty of customer data (derived from, e.g., credit card transactions), and insurance companies, with less frequent customer interactions [19], predictive models based on data analysis enrich the customer experience and improve the communication relevancy between financial institutions and customers.

Financial institutions make use of their internal data (collected by the institution) and, in the recent years, do real-time data analysis on social networks to watch news about companies, to evaluate prices and opinion trends, or to do damage control, 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.

3.7 Public Sector and Government

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 restated), 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. Powerful data analytic tools can help speed up retrieving relevant offers and matching them with the profile of the job seeker.

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 [42]. A recent report of the British Policy Exchange initiative [65] 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 3.5. The McKinsey study [42] 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, are often inefficient in publicizing their data and communicating them 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. Bundesrechenzentrum GmbH has been awarded the 2012 EuroCloud Europe Award for its E-Gov Portal Services. The City of Vienna provides much of its data as part of the Open Government Initiative so that companies can develop applications that ease daily life.

3.8 Education

The education sector is another potential beneficiary of Data Analytics in the near future. Educational Data Mining [15] 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 [55] with potential for dramatic change in the educational landscape of the future.

3.9 Telecommunications

Telecommunication providers traditionally work with large data sets. With five billion mobile phones in active use in 2010 [42], 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, there is information about approximate positioning to nearby cell antennas, 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 center of their business model and participate more strongly in the movement than other businesses. Unlike other types of businesses, they focus more strongly on the real-time aspects of data, and less on the large volume of the data, probably because large data volumes were part of their business model long before Big Data existed. In [42], it is found that communication businesses focus on the customer as the main objective for their big data analytics efforts, but that they are at a pilot stage and use their internal, pre-existing data sources.

3.10 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 basis (or even less frequently), modern meters now record constantly on a 15-Minute basis. 96 million readings per day, for every 100 million meters, results in a 3000 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, trends published in a recent report. 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 Data Analytics already starts informing us about the large-scale transformations in our society.

3.11 Law and Law Enforcement

This area is related to the government and public sector as described in Section 3.7. Traditionally, a police department collects millions of service calls, archives thousands of reports in combination with months of audio and video recordings. Data Analytics helps 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 5000 crimes dating back to 2006 and used it for predictive crime forecast [47]. 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 [45] 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.

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" [51]. 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.

3.12 eScience

Science is moving into a new, data-intensive computing era that has been called the *fourth paradigm for science* [36], 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 and genomics [43]. But smaller datasets are also making an impact, for example the digitisation of extensive cultural heritage, which allows analyses at an unprecedented scale in the humanities [34].

The European Commission Report *Riding the Wave* [6] lists the requirements and challenges in creating e-Infrastructures that will allow the practice of eScience. These 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 [36]. 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 programme supported the creation of e-Infrastructures, and projects have been funded in research areas including physics, astronomy, earth sciences, biology, seismology and agriculture.

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" [62]. Due to the volume of scientific papers published, augmenting the papers with machine-readable meta-data 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 workflow descriptions. A proposal for an ICT-based system, incorporating these concepts and more, to revolutionise scholarly communication and hence create an innovation accelerator is presented in [63].

3.13 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 [39]. Sport teams turn to Big Data to evaluate players, analyse strategies or past games.

Media and Entertainment. In [42, 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 mind set. Interestingly, at the same time, this area is one of the most IT intensive. Others [49] recognise the benefits the entertainment industry has had from the digitization of media. The main barriers in making the most out of the entertainment data are the siloed organizational structures, disconnected cross-functional and multi-business unit processes [49]. The strategies taken to increase the companies' competitiveness use Data Analytics and Big Data algorithms to manage customer experience, consolidate the hardware systems management to address the multitude of devices, services and networks currently in use, or develop new, on-demand delivery business and consumer products.

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 frontline 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, gathered 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 analyzed quickly, can provide invaluable customer insights that can help gaming companies effectively segment their customers, identify the 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 [5]. Through loyalty programs companies understand the behavior patterns borne out in hundreds of thousands of interactions per day [7].

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.

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, which is then consolidated with, for example, climate data and market conditions, provided by public and private institutions, help farmers do their business planning [17]. Opportunities of Data Analytics in agriculture include higher efficiency, reduced machine operation costs, better crop productivity, healthier livestock with impacts on food safety and environment.

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 [11]. 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.

4 Challenges and Open Issues

This section presents an initial overview of important open research challenges in the Intelligent Data Analytics area. It covers not only technical challenges related to conquering the data, but also challenges to be faced in a wider legal or societal framework. The challenges will be extended and further developed through input received from an online questionnaire, workshops with researchers and interviews with stakeholders.

4.1 Privacy and Security

Privacy is the selective ability of individuals or groups to withhold themselves or information about themselves entirely or in part[3]. 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[4]" 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 [20]. A recent paper from the Cloud Security Alliance presents the top ten Big Data security and privacy challenges [4].

Privacy concerns increase with high data volumes and advanced analytics applications. Whereas some sectors have established regulations (e.g. the health sector about sharing health records electronically), others rely far too often on inconsistent and unstable company policies paired with very general privacy laws that have not been updated to match the new technological landscape. 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. In January 2012, the European Commission unveiled a draft of the European General Data Protection Regulation that will supersede the Data Protection Directive. The directive 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.

This effort, 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 reindentify people and draw the entire legal effort into 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 [46]. User names and IP numbers were suppressed in an effort to anonymise the data set. However, AOL decided to maintain the connection 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 query (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[5]. This risk naturally increases with bigger data sets and better tools.

Another issue is that some types of information are very easy to exploit for re-identification (e.g. user location) and at the same time are required for an entire category of services that we all like (e.g. location-based services). A malicious location-based service can infer the identity of the query source from collecting a sequence of locations and then finally associating this "anonymous" user to a residence or an office building. At this point, the user has lost the anonymity and is exposed. Several other types of surprisingly private information such as health issues (e.g., presence in a cancer treatment center) or religious preferences (e.g., presence in a church) can also be revealed by just observing anonymous users' movement and usage pattern over time. It has been shown that there is a close correlation between people's identities and their movement patterns [31].

This paralyses almost everything that currently exists in terms of data privacy law and calls for revision and the need to address these issues both legally and technologically. 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 or 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

[3] Paraphrased from `http://en.wikipedia.org/wiki/Privacy`

[4] `http://www.en.wikipedia.org/wiki/Information_security`

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

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 [13].

Accumulo, infamously known for its use in the NSA PRISM project, is ironically one of the few tools that includes security in its core. The tool is based on the BigTable framework by Google, and built on top of Apache Hadoop. One of Accumulo's strengths is finding connections among seemingly unrelated data sets. Accumulo's major feature is a security filter that operates at the cell level unlike HBase or Cassandra. Each data cell can carry a range of security labels that need to be satisfied at query time. This allows data with divergent security clearance to be stored in the same table and the cell can still be controlled with respect to a security policy on a large, distributed scale with many users and many simultaneous tasks.

4.2 Algorithm and Data Issues

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 (Section 2). The choice of (sets of) algorithms 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.

One of the main challenges observed in Intelligent Data Analytics relates to the data fed into the analytical processes. Algorithms have to deal with data that is dynamic (like financial market data); or 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 [13]. 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. Among the currently most popular approaches to overcome the problem of scale, we mention parallel algorithms [21] to split the workload among several computing units, extracting representative samples of data to reduce the computing workload [59], streaming algorithms that process the data as it is collected [30], where the data is unknown before execution. Observing how data is generated and collected we expect that future solution implementations will use adaptable algorithms. Changes in data format and representations, in hardware used, etc. are very likely, therefore algorithms should be designed having in mind that the context in which they are used may change.

It is generally a requirement that, as the amount of data to be processed increases, the algorithms' output is given in a reasonable and useful amount of time. Though not explicitly mentioned, 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 [16]. Research into creating better methods to analyze and extract knowledge from large and massive data repositories is rich with improvements and novel ideas (see for example [12, 27]).

It is mentioned in [13] 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 optimization 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.

4.3 Evaluation

As is clear from Section 2, there are many algorithms 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. Whether the optimal solution is found can often depend on the aptitudes and skills of the people solving the problem.

Kaggle (`http://kaggle.com`) 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.

In the academic community, the emphasis has been more on removing the human influence on solutions by careful benchmarking of algorithms. An important tool in achieving this is the organisation of evaluation campaigns, also known as challenges, benchmarks or competitions. Such events are common, for example, in Information Retrieval [64] and Machine Learning [52]. In such events, the organisers make available data and associated tasks, 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 algorithms submitted in order to advance scientific knowledge, and 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 [56] 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 evaluation campaigns are carried out, allowing, for example, the use of larger datasets and real-time evaluation independent of a cycle of workshops.

This can be seen as work towards increasing the reproducibility of experimental results in the computational sciences, in order to overcome concerns about their reliability [29]. The work underway ranges from presenting the case for open computer programs [37], through creating infrastructures to allow reproducible computational research [29] to considerations about the legal licensing and copyright frameworks for computational research [58].

4.4 Data Ownership and Open Data

The problems faced when talking about *data ownership* are illustrated in a 2003 Information Management column by D. Loshin [41]. 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 [53]. 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 [9].

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 from the "new" data. However, the notorious AOL search log data release incident [14] (also mentioned in Section 4.1) 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 in [32].

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 this assessment.

4.5 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.

With the value of data having been recognised, 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 [50]. This requires completely new machine-actionable models of process management plans [44].

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 [8, 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 lose their support, tools need to be developed that efficiently and as far as possible losslessly transform these data and move them into newer data repositories. At the semantic level, data changes its meaning as terminology evolves. Ensuring correct and authentic interpretation 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 capture data semantics in a machine-processable manner [3].

4.6 Qualified Personnel

According to the McKinsey report [42], there are three types of talent required to take full advantage of Intelligent Data Analytics: *deep analytical talent* — 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; and *supporting technology personnel* who develop, implement, and maintain the hardware and software tools such as databases and analytic programs needed. Although this report looked at the requirements from the view of the industry, a recent paper in Nature [43] 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.

The 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.

McKinsey [42] estimates, based on 2008 figures, that there will be a shortage in the USA of 140,000–190,000 people with deep analytical talent by 2018. In 2008, it is estimated that 24,730 students with deep analytical skills graduated in the USA, while in Austria in 2008, the estimate is that 370 such students graduated. It is likely that the US will continue to rely partly on immigration to cover their talent shortage in this area, increasing the pressure on countries that tend to export talent. As developing deep analytic skills requires mathematical aptitude and extensive training, it will take time to increase the rate of production of data scientists, and universities should consider creating the necessary interdisciplinary degree programmes to achieve this. 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" (Harvard Business Review, October 2012).

McKinsey [42] also predict that there will be a shortage of 1.5 million data-savvy managers and analysts in the USA in 2018, and hence also likely in other countries. The level of training and mathematical aptitude for this group is lower than for the deep analytical talent, but these people need enough conceptual knowledge and quantitative skills to be able to frame and interpret analyses in an effective way. Filling this gap requires enhancing relevant university curricula with Intelligent Data Analytics courses, but also training the existing workforce again.

5 Austrian Intelligent Data Analytics Competence Landscape

This section summarises the Austrian Intelligent Data Analytics competence landscape. As with other aspects of this study, it will be extended through interaction and consultation with the Intelligent Data Analytics community in Austria, and the final version will appear in the final study report. The Intelligent 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 Intelligent Data Analytics challenges for clients; and *end users* — companies with data and challenges to be solved by Intelligent Data Analytics.

5.1 Research Landscape

For the universities, research groups working in the areas 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[6], 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, more emphasis was placed on obtaining the research foci from the web pages, due to a usually smaller number of publications. The Universities of Applied Science are currently not included in the list, due to their larger emphasis on teaching — it is however planned to include relevant Universities of Applied Science in the final study report.

Two visualisations of the data obtained are presented. For Table 2, all research groups are clustered by their host organisation, and their research foci are classified into the groups listed in Section 2, leading to a matrix showing the Intelligent Data Analytics research competences available in the Austrian universities and research centres. The colours indicate the province, according to the legend in Table 1. While it is clear that all groups evaluate their algorithms, the evaluation column indicates groups that

Table 1: *Colour legend.*

Colour	Meaning
🟪	Carinthia
🟦	Lower Austria
🟩	Styria
🟪	Salzburg
🟦	Tyrol
🟧	Upper Austria
🟥	Vienna
🟥	Company with headquarters outside Austria

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

Table 2: *Research foci of universities and research centres. Colour meanings are described in Table 1.*

	Search and Analysis															**Sem. Proc.**				**Cognitive Systems**								**Visualisation**				
Organization	text search	image search	music search	video/multimedia search	computer vision	speech/audio processing	process analysis	network science	ubiquitous computing	info. integration/fusion	statistics	data mining	bioinformatics	digital preservation	algorithmic efficiency	information extraction	knowledge engineering	semantic web	natural language proc.	machine learning	comp. neuroscience	reasoning	recommender systems	decision support systems	simulation	crowdsourcing	brain com interfaces	visualisation	visual analytics	rendering	sonification	Evaluation
Universität Wien	🟥			🟥			🟥				🟥	🟥	🟥		🟥		🟥		🟥									🟥		🟥		
TU Wien	🟥	🟥	🟥	🟥	🟥						🟥	🟥		🟥	🟥	🟥	🟥	🟥		🟥		🟥						🟥	🟥	🟥		🟥
Wirtschaftuni. Wien							🟥		🟥		🟥	🟥						🟥		🟥		🟥		🟥								
MODUL Universität																		🟥											🟥			
BOKU Wien							🟥				🟥									🟥												
Medizinische Uni. Wien	🟥				🟥			🟥			🟥					🟥				🟥		🟥										
Max F. Perutz Labs													🟥																			
Forschungszentrum Telekommunikation Wien (FTW)									🟥																							
Zentrum für Virtual Reality und Visualisierung (VRVIS)																												🟥	🟥	🟥		
Austrian Institute of Technology (AIT)				🟥	🟥				🟥	🟥		🟥	🟥	🟥	🟥					🟥					🟥							
Austrian Institute for Artificial Intelligence (OFAI)			🟥																🟥													
Österreichische Akademie der Wissenschaften													🟥						🟥													
Österreichische Nationalbibliothek														🟥																		
TU Graz					🟩	🟩		🟩			🟩					🟩	🟩	🟩		🟩	🟩						🟩	🟩				
Universität Graz											🟩				🟩																	
Medizinische Uni. Graz	🟩										🟩							🟩				🟩										🟩
Montanuni. Leoben																				🟩												
Fraunhofer Austria					🟩																											
Joanneum DIGITAL				🟩	🟩	🟩												🟩														
Know Center Graz	🟩								🟩							🟩	🟩											🟩				
Universität Innsbruck					🟦						🟦				🟦		🟦	🟦		🟦			🟦									🟦
Med. Uni. Innsbruck											🟦									🟦												
UMIT					🟦			🟦									🟦															
Universität Salzburg					🟪						🟪				🟪					🟪												
Salzburg Research												🟪						🟪														
Universität Linz			🟧				🟧		🟧	🟧	🟧	🟧				🟧	🟧			🟧		🟧	🟧									🟧
Software Competence Center Hagenberg					🟧											🟧				🟧												
IST Austria					🟦						🟦									🟦												
Universität Klagenfurt		🟪		🟪					🟪		🟪							🟪	🟪				🟪									
TOTAL	**5**	**2**	**2**	**5**	**11**	**2**	**4**	**3**	**6**	**2**	**14**	**6**	**4**	**3**	**6**	**6**	**7**	**9**	**4**	**13**	**1**	**5**	**3**	**1**	**1**	**0**	**1**	**5**	**3**	**3**	**0**	**4**

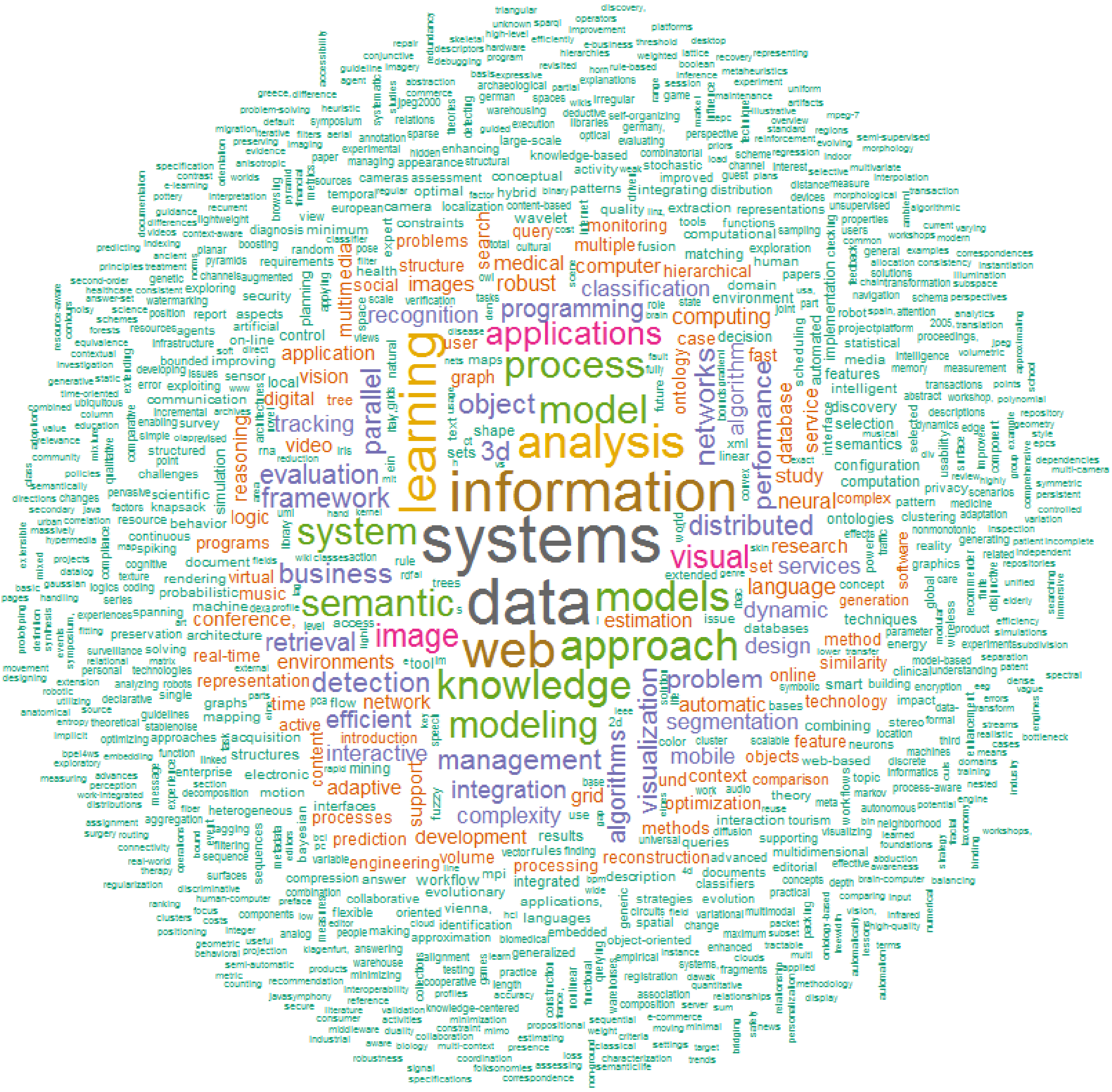

Figure 2: *Publication titles word cloud for research groups working in data analysis in Austria.*

have organised a larger-scale evaluation activity, such as an evaluation campaign or challenge. Figure 2 shows a word cloud created by processing the words in all publication titles extracted as described above from all identified research groups. Interestingly, the word occurring most in the titles is “using,” which has been removed to allow a better distribution of word sizes for the words more related to research areas.

5.2 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 i) identified the application domains and industries the companies operate in and ii) tried to elicit the technological foundations of their offerings. Table 4 summarises the result 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, *Tourism and Hospitality*, *Education* as well as *Law* are less prominently addressed by the reviewed providers (see Table 3). A summary of the technological foundations of the offerings of reviewed

Table 3: *Distribution of service providers across application domains.*

Manufacturing & Logistics	28
Finance & Insurance	24
Transportation & Travel	23
Public Sector/Government	23
Healthcare	17
Commerce & Retail	17
Telecommunications	17
Energy & Utilities	14
eScience	14
Tourism & Hospitality	7
Education	6
Law Enforcement & Legal	4

service providers is given in Figure 3. The technological foundations are depicted as a tag cloud, which was derived by collecting the technologies mentioned in the context of the respective offering. A content analytics tool (`http://smartcoder.at/`) was used to pre-code and harmonise the technologies.

5.3 End Users

This study focusses on creating a Intelligent Data Analytics research 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 3. A further study focussing on the end users of Data Analytics solutions and big data has been commissioned.

6 Conclusion

This position paper provides a definition of the field of *Intelligent Data Analytics* to be applied to the challenge of *Conquering Data* identified by the Austrian Research Promotion Agency (FFG). It identifies the techniques required for Conquering Data, divided into four (interacting) groups: *Search and Analysis*, *Semantic Processing*, *Cognitive Systems and Prediction*, and *Visualisation and Representation. Evaluation and Benchmarking* techniques are required in all four groups. For each of the techniques, research groups in Austria with relevant competencies were identified. The five competencies represented in the largest number of research organisations (universities and research centres) in Austria are: statistics, machine learning, computer vision, semantic web and knowledge engineering.

The most important Intelligent Data Analytics application areas are summarised, and companies in Austria providing solutions by applying Intelligent Data Analytics in these areas are identified. The largest number of companies are applying Intelligent Data Analytics in the area of Manufacturing and Logistics, with the areas of Finance and Insurance, Transportation and Travel, and Public Sector/Government also being well covered.

Finally, important challenges in Intelligent Data Analytics, both from a technological and societal point of view, are identified. Technological challenges can be divided into: *algorithmic challenges*, ensuring: that algorithms are able to effectively analyse the ever increasing amount of heterogeneous and potentially incomplete data, that the capabilities of the algorithms are well understood, and that the analyses lead to sensible results; and *data challenges*, ensuring that useful data is not lost and remains accessible over time. Some challenges require legal expertise, such as the questions on *data ownership*, while *privacy and security* should be treated from both the technological and societal sides — a legal framework and corresponding technological capabilities to implement it are required. Finally, it is not surprising that there is a predicted worldwide shortage of not only people with the training and expertise to perform Intelligent Data Analytics, but also of people capable of taking advantage of the potential of Intelligent Data Analytics — relevant educational measures are required to counter this prognosis.

Table 4: *Application domain by company. Colour meanings are described in Table 1.*

Company	**Healthcare**	**Commerce & Retail**	**Manufact. & Logistics**	**Transport & Travel**	**Energy and Utilities**	**Public Sector/Govt.**	**Education**	**Tourism & Hospitality**	**Telecommunications**	**eScience**	**Law & Law Enf.**	**Finance & Insurance**	**Further App Domain**	**n/a**
AltaPlana			▨			▨		▨				▨	IT (High Tech); Market Research	
ANDATA Entwicklungstech.				🟪										
APA-IT	🟥		🟥			🟥						🟥	Media and Entertainment	
BIConcepts		🟥	🟥			🟥				🟥		🟥		
Booz & Company														▨
Business Intelligence Accelerator				🟧										
Capgemini		▨	▨	▨	▨				▨					
Catalysts	🟧	🟧	🟧	🟧	🟧	🟧			🟧	🟧		🟧	Gaming	
cept systems														🟥
CogVis software & consulting			🟥	🟥									Media and Entertainment; Gaming	
Connex.cc DI Hadek									🟥					
CSC	▨			▨								▨		
Data Technology	🟥			🟥		🟥							Gaming	
DGR			🟧										Construction	
diamond:dogs														🟥
EBCONT														🟦
EMC2	▨				▨	▨							Media and Entertainment	
Evolaris			🟩									🟩		
FABASOFT // Mindbreeze			🟧		🟧	🟧					🟧			
Fluidtime Data Services				🟥		🟥		🟥						
Frequentis				🟥					🟥			🟥	Defense	
Gnowsis			🟥											
HP	▨	▨	▨		▨	▨			▨	▨	▨	▨	Media and Entertainment	
IBM	▨	▨	▨	▨	▨	▨	▨		▨	▨		▨	IT (High Tech); Media and Entertainment	
IDC - International Data Corp.	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨		
imposult		🟥		🟥					🟥	🟥		🟥		
InterXion Österreich													IT (High Tech)	
KiwiSecurity Software	🟥	🟥	🟥	🟥		🟥	🟥					🟥		
Lixto		🟥	🟥	🟥				🟥						
M2N	🟧		🟧			🟧					🟧	🟧		
max.recall information systems						🟥				🟥			Market Research	
MediaServices						🟪								
Microsoft AT	▨	▨	▨	▨	▨	▨			▨	▨		▨	IT (High Tech); Media and Entertainment	
Oracle AT	▨			▨	▨				▨				Media and Entertainment; IT (High Tech)	
Plaut Consulting Austria			🟥		🟥							🟥		
pmOne			▨	▨				▨		▨		▨		
Profactor GmbH	🟧		🟧	🟧						🟧			IT (High Tech)	
SanData Technology														🟥
SAP Austria		▨	▨	▨								▨	IT (High Tech)	
SAS Austria	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨		▨	Media and Entertainment	
Semantic Web Company			🟥		🟥	🟥	🟥		🟥	🟥		🟥	Media and Entertainment; Gaming	
Semanticlabs		🟥	🟥						🟥				Media and Entertainment	
Siemens					▨									
Software AG		▨		▨				▨	▨			▨	Media and Entertainment	
Spectralmind														🟥
Talend	🟧	🟧	🟧			🟧	🟧		🟧			🟧	Media and Entertainment	
Teradata	▨	▨	▨	▨		▨			▨	▨		▨	Media and Entertainment; Gaming	
Tricentis														🟥
UBIMET			🟥	🟥		🟥						🟥	Media and Entertainment	
uma information technology		🟥	🟥		🟥	🟥			🟥	🟥				
Umweltbundesamt													Environmental Services	
Unisys Österreich														
WebLyzard	🟥					🟥						🟥		
Total	**17**	**17**	**28**	**23**	**14**	**23**	**6**	**7**	**17**	**14**	**4**	**24**		**7**

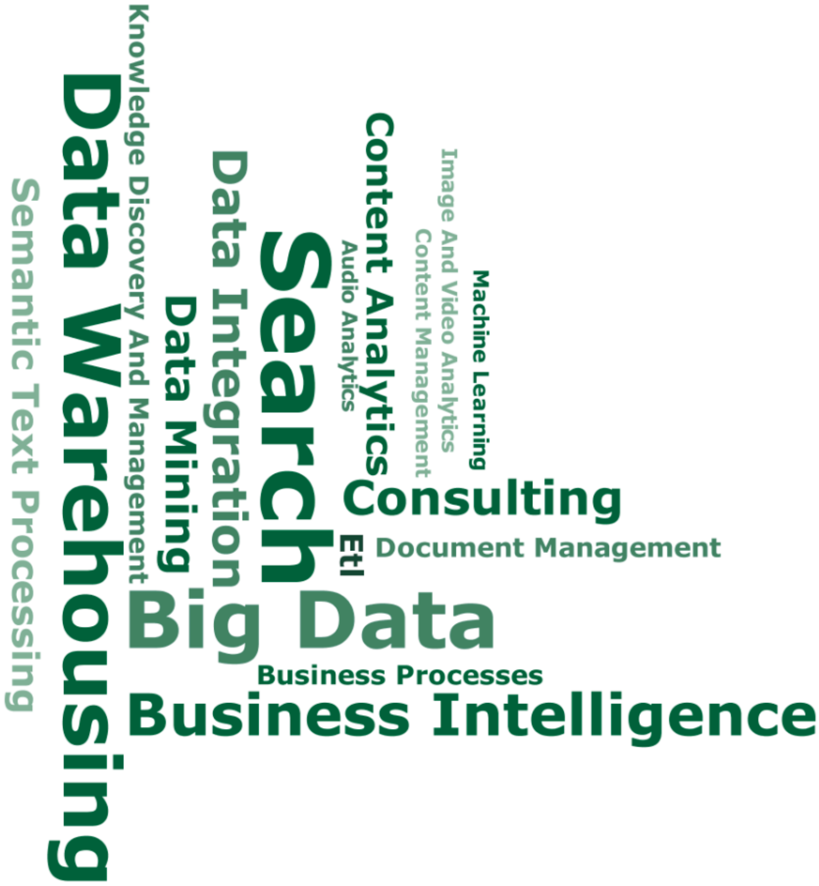

Figure 3: *Technological foundations of offerings of service providers.*

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 Andreas Rauber for his valuable input.

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

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.

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.

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.

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, Multi-modal Search and IR Evaluation. He has more than 25 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.

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.

References

[1] FORCE11 Data Citation Synthesis Group. Online: http://www.force11.org/node/4432 (last visited: September 2013).

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

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

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

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

[6] Riding the Wave: How Europe can gain from the rising tide of scientific data. European Commission, 2010.

[7] The big data opportunity. Birst Inc., 2012. online: http://tdwi.org/~/media/9836C350D9F641669563B19DD563B6E1.pdf (last visited: August 2013).

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

[9] Data protection laws of the world, 2013. online: http://bit.ly/16GAKEl (last visited: August 2013).

[10] Now arriving: Big data in the hospitality, travel and tourism sector. SOCAP International, 2013. online: https://scpelc.egnyte.com/h-s/20130510/c61863288b5f4c68 (last visited: August 2013).

[11] Österreichisches Bundesheer investiert 10 Millionen Euro in Technologie und Forschung. APA-OTS Presse sendung OTS0195, 2013. 2013-08-22.

[12] 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.

[13] 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. online: http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf (last visited: August 2013).

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

[15] 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.

[16] M. Barlow. *Real-Time Big Data Analytics: Emerging Architecture*. O'Reilly, 2013.

[17] C. Bauckhage and K. Kersting. Data Mining and Pattern Recognition in Agriculture. *KI - Künstliche Intelligenz*, pages 1–12, 2013.

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

[19] E. Brat, S. Heydorn, M. Stover, and M. Ziegler. Big Data: The Next Big Thing For Insurers? The Boston Consulting Group Perspectives, March 2013.

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

[21] 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.

[22] 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.

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

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

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

[26] E. A. Feigenbaum and P. McCorduck. *The fifth generation*. Addison-Wesley, 1983.

[27] 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.

[28] M. A. Feufel, G. Antes, J. Steurer, G. Gigerenzer, J. A. M. Gray, M. Mäkelä, J. A. G. Mulley, 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*. MIT Press, 2011.

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

[30] J. Gama. *Knowledge Discovery From Data Streams*. Chapman & Hall/CRC, 2010.

[31] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding Individual Human Mobility Patterns. *Nature*, 453(7196):779–782, June 2008.

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

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

[34] E. Hand. Word play. *Nature*, 474:436–440, 2011.

[35] J. A. Harding, M. Shahbaz, and A. Kusiak. Data Mining in Manufacturing: A Review. *J. of Manufacturing Science and Engineering*, 128, 2006.

[36] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, 2009.

[37] D. C. Ince, L. Hatton, and J. Graham-Cumming. The case for open computer programs. *Nature*, 482:485–488, 2012.

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

[39] 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/, 2013. Accessed: 2013-09-01.

[40] 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.

[41] D. Loshin. Who owns data? Information Management, 2003.

[42] 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.

[43] C. A. Mattman. A vision for data science. *Nature*, 493:474–475, 2013.

[44] 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.

[45] 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.

[46] P. Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, 57:1701, 2009. Available at SSRN: http://ssrn.com/abstract=1450006.

[47] A. Olesker. White paper: Big data solutions for law enforcement, 2012. IDC White paper.

[48] T. O'Reilly, M. Loukides, J. Steele, and C. Hill. *How Data Science is Transforming Health Care.* O'Reilly Media, 2012.

[49] D. Osswald and G. Girard. Improving Business Outcomes with Big Data and Analytics in Communications, Media, and Entertainment, 2012. IDC White paper.

[50] 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.

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

[52] 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.

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

[54] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook.* Springer, 2011.

[55] 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.

[56] 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.

[57] U. Securities, E. Commission, and the Commodity Futures Trading Commission. Findings regarding the market events of may 6, 2010. http://www.sec.gov/news/studies/2010/marketevents-report.pdf, 2010.

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

[59] 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.

[60] K. Temple. What happens in an internet minute?, 2013. online: http://scoop.intel.com/what-happens-in-an-internet-minute/ (last visited: August 2013).

[61] 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.

[62] 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.

[63] F. van Harmelen, G. Kampis, K. B?rner, 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.

[64] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press, 2005.

[65] C. Yiu. Big data opportunity. making government faster, smarter and more personal, 2012. Policy Exchange.

Copyright

Copyrighted material used under Fair Use. If you are the copyright holder and believe your material has been used unfairly, or if you have any suggestions, feedback, or support, please email to `contact@conqueringdata.at`.

Except where otherwise indicated, permission is granted to copy, distribute, and/or modify all images in this document under the terms of the GNU Free Documentation license, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled “GNU Free Documentation license” at `http://commons.wikimedia.org/wiki/Commons:GNU_Free_Documentation_License`.