Chapter 5
Big Data Analysis

John Domingue, Nelia Lasierra, Anna Fensel, Tim van Kasteren, Martin Strohbach, and Andreas Thalhammer

5.1 Introduction

Data comes in many forms and one dimension to consider and compare differing data formats is the amount of *structure* contained therein. The more structure a dataset has the more amenable it will be to machine processing. At the extreme, semantic representations will enable machine reasoning. Big data analysis is the sub-area of big data concerned with adding structure to data to support decision-making as well as supporting domain-specific usage scenarios. This chapter outlines key insights, state of the art, emerging trends, future requirements, and sectorial case studies for data analysis.

The position of big data analysis within the overall big data value chain can be seen in Fig. 5.1. 'Raw' data which may or may not be structured and which will usually be composed of many different formats is transformed to be ready for data curation, data storage, and data usage. That is why without big data analysis most of the acquired data would be useless.

J. Domingue (✉)
STI International, Neubaugasse 10/15 A, 1070 Vienna, Austria

Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK
e-mail: john.domingue@open.ac.uk

N. Lasierra • A. Fensel
University of Innsbruck, Technikerstraße 21a, 6020 Innsbruck, Austria
e-mail: nelia.lasierra@sti2.at; anna.fensel@sti2.at

T. van Kasteren • M. Strohbach
AGT International, Hilpertstr, 35, 64295 Darmstadt, Germany
e-mail: TKasteren@agtinternational.com; MStrohbach@agtinternational.com

A. Thalhammer
Institute for Applied Informatics and Formal Description Methods, Karlsruhe Institute of Technology, Kaiserstraße 89, 76133 Karlsruhe, Germany
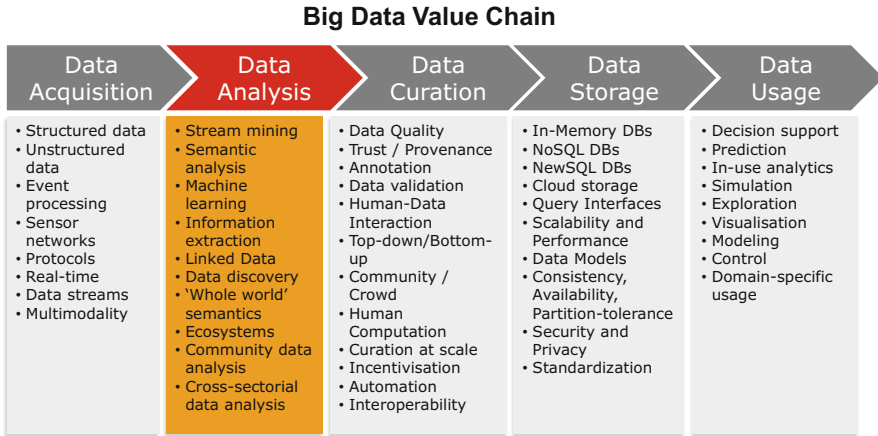e-mail: Thalhammer@kit.edu

Big Data Value Chain

Data Acquisition	Data Analysis	Data Curation	Data Storage	Data Usage
• Structured data • Unstructured data • Event processing • Sensor networks • Protocols • Real-time • Data streams • Multimodality	• Stream mining • Semantic analysis • Machine learning • Information extraction • Linked Data • Data discovery • 'Whole world' semantics • Ecosystems • Community data analysis • Cross-sectorial data analysis	• Data Quality • Trust / Provenance • Annotation • Data validation • Human-Data Interaction • Top-down/Bottom-up • Community / Crowd • Human Computation • Curation at scale • Incentivisation • Automation • Interoperability	• In-Memory DBs • NoSQL DBs • NewSQL DBs • Cloud storage • Query Interfaces • Scalability and Performance • Data Models • Consistency, Availability, Partition-tolerance • Security and Privacy • Standardization	• Decision support • Prediction • In-use analytics • Simulation • Exploration • Visualisation • Modeling • Control • Domain-specific usage

Fig. 5.1 Data analysis in the big data value chain

The analysis found that the following generic techniques are either useful today or will be in the short to medium term: reasoning (including stream reasoning), semantic processing, data mining, machine learning, information extraction, and data discovery.

These generic areas are not new. What is new however are the challenges raised by the specific characteristics of big data related to the three Vs:

- **Volume**—places scalability at the centre of all processing. Large-scale reasoning, semantic processing, data mining, machine learning, and information extraction are required.
- **Velocity**—this challenge has resulted in the emergence of the areas of stream data processing, stream reasoning, and stream data mining to cope with high volumes of incoming raw data.
- **Variety**—may take the form of differing syntactic formats (e.g. spreadsheet vs. csv) or differing data schemas or differing meanings attached to the same syntactic forms (e.g. 'Paris' as a city or person). Semantic techniques, especially those related to Linked Data, have proven to be the most successful applied thus far although scalability issues remain to be addressed.

5.2 Key Insights for Big Data Analysis

Interviews with various stakeholders related to big data analysis have identified the following key insights. A full list of interviewees is given in Table 3.1.

Old Technologies Applied in a New Context Individual and combinations of old technologies being applied in the Big Data context. The difference is the scale (volume) and the amount of heterogeneity encountered (variety). Specifically, in

Table 3.1 Big data analysis interviewees

No.	First name	Last name	Organization	Role/Position
1	Sören	Auer	Leipzig	Professor
2	Ricardo	Baeza-Yates	Yahoo!	VP of Research
3	François	Bancilhon	Data Publica	CEO
4	Richard	Benjamins	Telefoncica	Director Biz Intel
5	Hjalmar	Gislason	datamarket.com	Founder
6	Alon	Halvey	Google	Research Scientist
7	Usman	Haque	Cosm (Pachube)	Director Urban Project Division
8	Steve	Harris	Garlik/Experian	CTO
9	Jim	Hendler	RPI	Professor
10	Alek	Kołcz	Twitter	Data Scientist
11	Prasanna	Lal Das	World Bank	Snr Prog. Officer, Head of Open Financial Data Program
12	Peter	Mika	Yahoo!	Researcher
13	Andreas	Ribbrock	Teradata GmbH	Team Lead Big Data Analytics and Senior Architect
14	Jeni	Tennison	Open Data Institute	Technical Director
15	Bill	Thompson	BBC	Head of Partner Development
16	Andraž	Tori	Zemanta	Owner and CTO
17	Frank	van Harmelen	Amsterdam	Professor
18	Marco	Viceconti	University of Sheffield and the VPH Institute	Professor and Director
19	Jim	Webber	Neo	Chief Scientist

the web context a focus is seen on large semantically based datasets such as Freebase and on the extraction of high-quality data from the web. Besides scale there is novelty in the fact that these technologies come together at the same time.

Stream Data Mining This is required to handle high volumes of stream data that will come from sensor networks or online activities from high numbers of users. This capability would allow organizations to provide highly adaptive and accurate personalization.

'Good' Data Discovery Recurrent questions asked by users and developers are: Where can we get the data about X? Where can we get information about Y? It is hard to find the data and found data is often out of date and not in the right format. Crawlers are needed to find big datasets, metadata for big data, meaningful links between related datasets, and a dataset ranking mechanism that performs as well as Page Rank does for web documents.

Dealing with Both Very Broad and Very Specific Data A near feature about information extraction from the web is that the web is about everything so coverage is broad. Pre-web the focus was on specific domains when building databases and

knowledge bases. This can no longer be done in the context of the web. The whole notion of "conceptualizing the domain" is altered: Now the domain is everything in the world. On the positive side, the benefit is you get a lot of breadth, and the research challenge is how one can go deeper into a domain while maintaining the broad context.

Simplicity Leads to Adoptability Hadoop[1] succeeded because it is the easiest tool to use for developers, changing the game in the area of big data. It did not succeed because it was the best but because it was the easiest to use (along with HIVE).[2] Hadoop managed to successfully balance dealing with complexity (processing big data) and simplicity for developers. Conversely, semantic technologies are often hard to use. Hjalmar Gislason, one of our interviewees advocates the need for the "democratisation of semantic technologies".

Ecosystems Built around Collections of Tools Have a Significant Impact These are often driven by large companies where a technology is created to solve an internal problem and then is given away. Apache Cassandra[3] is an example of this initially developed by Facebook to power their inbox search feature until 2010. The ecosystem around Hadoop is perhaps the best known.

Communities and Big Data Will Be Involved in New and Interesting Relationships Communities will be engaged with big data in all stages of the value chain and in a variety of ways. In particular, communities will be involved intimately in data collection, improving data accuracy and data usage. Big data will also enhance community engagement in society in general.

Cross-sectorial Uses of Big Data Will Open Up New Business Opportunities The retail section of future requirements and emerging trends describes an example for this. O2 UK together with Telefónica Digital has recently launched a service that maps and repurposes mobile data for the retail industry. This service allows retailers to plan where to site retail outlets based upon the daily movement of potential customers. This service highlights the importance of internal big data (in this case mobile records) that is later combined with external data sources (geographical and preference data) to generate new types of business. In general aggregating data across organizations and across sectors will enhance the competitiveness of European industry.

The biggest challenge for most industries is now to incorporate big data technologies in their processes and infrastructures. Many companies identify the need for doing big data analysis, but do not have the resources for setting up an infrastructure for analysing and maintaining the analytics pipeline (Benjamins). Increasing the simplicity of the technology will aid the adoption rate. On top of this a large body of domain knowledge has to be built up within each industry on how

---

[1] http://hadoop.apache.org/

[2] https://hive.apache.org/

[3] http://cassandra.apache.org/

data can be used: What is valuable to extract and what output can be used in daily operations.

The costs of implementing big data analytics are a business barrier for big data technology adoption. Anonymity, privacy, and data protection are cross-sectorial requirements highlighted for big data technologies. Additional information can be found in the final analysis of sector's requisites (Zillner et al. 2014). Examples of some sectorial case studies can be found in Sect. 5.5.

5.3   Big Data Analysis State of the Art

Industry is today applying large-scale machine learning and other algorithms for the analysis of huge datasets, in combination with complex event processing and stream processing for real-time analytics. It was also found that the current trends on Linked Data, semantic technologies, and large-scale reasoning are some of the topics highlighted by the interviewed experts in relation to the main research challenges and main technological requirements for big data.

This section presents a state-of-the-art review regarding big data analysis and published literature, outlining a variety of topics ranging from working efficiently with data to large-scale data management.

5.3.1   Large-Scale: Reasoning, Benchmarking, and Machine Learning

The size and heterogeneity of the web precludes performing full reasoning and requires new technological solutions to satisfy the requested inference capabilities. This requested feature has also been extended to machine-learning technologies and these technologies are required in order to extract useful information from huge amounts of data. Specifically, François Bancilhon mentioned in his interview how machine learning is important for topic detection and document classification at Data Publica. Then, Ricardo Baeza-Yates highlighted in his interview the need for standards in big data computation in order to allow big data providers to compare their systems.

5.3.1.1   Large-Scale Reasoning

The promise of reasoning as promoted within the context of the semantic web does not currently match the requirements of big data due to scalability issues. Reasoning is defined by certain principles, such as soundness and completeness, which are far from the practical world and the characteristics of the web, where data is often

contradictory, incomplete, and of an overwhelming size. Moreover, there exists a gap between reasoning at web scale and the more tailored reasoning over simplified subsets of first-order logic, due to the fact that many aspects are assumed, which differ from reality (e.g. small set of axioms and facts, completeness and correctness of inference rules).

State-of-the-art approaches (Fensel 2007) propose a combination of reasoning and information retrieval methods (based on search techniques), to overcome the problems of web scale reasoning. Incomplete and approximate reasoning was highlighted by Frank van Harmelen as an important topic in his interview.

Querying and reasoning over structured data can be supported by semantic models automatically built from word co-occurrence patterns from large text collections (distributional semantic models) (Turney and Pantel 2010). Distributional semantic models provide a complementary layer of meaning for structured data, which can be used to support semantic approximation for querying and reasoning over heterogeneous data (Novacek et al. 2011; Freitas et al. 2013; Freitas and Curry 2014).

The combination of logic-based reasoning with information retrieval is one of the key aspects to these approaches and also machine-learning techniques, which provide a trade-off between the full-fledged aspects of reasoning and the practicality of these in the web context. When the topic of scalability arises, storage systems play an important role as well, especially the indexing techniques and retrieval strategies. The trade-off between online (backward) reasoning and offline (forward) reasoning was mentioned by Frank van Harmelen in his interview. Peter Mika outlined as well the importance of efficient indexing techniques in his interview.

Under the topic of large-scale systems, LarKC (Fensel et al. 2008) is a flagship project. LarKC[4] was an EU FP7 Large-Scale Integrating Project and the aim of it was to deal with large scalable reasoning systems and techniques using semantic technologies.

5.3.1.2 Benchmarking for Large-Scale Repositories

Benchmarking is nascent for the area of large-scale semantic data processing, and in fact currently they are only now being produced. Particularly, the Linked Data Benchmark Council (LDBC) project[5] aims to "create a suite of benchmarks for large-scale graph and RDF (Resource Description Framework) data management as well as establish an independent authority for developing benchmarks". A part of the suite of benchmarks created in LDBC is the benchmarking and testing of data integration and reasoning functionalities as supported by RDF systems. These benchmarks are focused on testing: (1) instance matching and Extract, Transform and Load that play a critical role in *data integration*; and (2) the *reasoning*

---

[4] LarKC Homepage, http://www.larkc.eu, last visited 3/03/2015.

[5] LDBC Homepage, http://www.ldbc.eu/, last visited 3/05/2015.

capabilities of existing RDF engines. Both topics are very important in practice, and they have both been largely ignored by existing benchmarks for Linked Data processing. In creating such benchmarks LDBC analyses various available scenarios to identify those that can best showcase the data integration and reasoning functionalities of RDF engines. Based on these scenarios, the limitations of existing RDF systems are identified in order to gather a set of requirements for RDF data integration and reasoning benchmarks. For instance, it is well known that existing systems do not perform well in the presence of non-standard reasoning rules (e.g. advanced reasoning that considers negation and aggregation). Moreover, existing reasoners perform inference by materializing the closure of the dataset (using backward or forward chaining). However, this approach might not be applicable when application-specific reasoning rules are provided and hence it is likely that improving the state of the art will imply support for hybrid reasoning strategies involving both backward and forward chaining, and query rewriting (i.e. incorporating the ruleset in the query).

5.3.1.3 Large-Scale Machine Learning

Machine-learning algorithms use data to automatically learn how to perform tasks such as prediction, classification, and anomaly detection. Most machine-learning algorithms have been designed to run efficiently on a single processor or core. Developments in multi-core architectures and grid computing have led to an increasing need for machine learning to take advantage of the availability of multiple processing units. Many programming interfaces and languages dedicated to parallel programming exist such as Orca MPI or OpenACC, which are useful for general purpose parallel programming. However, it is not always obvious how existing machine-learning algorithms can be implemented in a parallelized manner. There is a large body of research on distributed learning and data mining (Bhaduri et al. 2011), which encompasses machine-learning algorithms that have been designed specifically for distributed computing purposes.

Rather than creating specific parallel versions of algorithms, more generalized approaches involve frameworks for programming machine learning on multiple processing units. One approach is to use a high-level abstraction that significantly simplifies the design and implementation of a restricted class of parallel algorithms. In particular the MapReduce abstraction has been successfully applied to a broad range of machine-learning applications. Chu et al. (2007) show that any algorithm fitting the statistical query model can be written in a certain summation form, which can be easily implemented in a MapReduce fashion and achieves a near linear speed-up with the number of processing units used. They show that this applies to a variety of learning algorithms (Chu et al. 2007). The implementations shown in the paper led to the first version of the MapReduce machine learning library Mahout.

Low et al. (2010) explain how the MapReduce paradigm restricts users to using overly simple modelling assumptions to ensure there are no computational dependencies in processing the data. They propose the Graphlab abstraction that insulates

users from the complexities of parallel programming (i.e. data races, deadlocks), while maintaining the ability to express complex computational dependencies using a data graph.

The programming languages, toolkits, and frameworks discussed allow many different configurations for carrying out large-scale machine learning. The ideal configuration to use is application dependent, since different applications will have different sets of requirements. However, one of the most popular frameworks used in recent years is that of Apache Hadoop, which is an open-source and free implementation of the MapReduce paradigm discussed above. Andraž Tori, one of our interviewees, identifies the simplicity of Hadoop and MapReduce as the main driver of its success. He explains that a Hadoop implementation can be outperformed in terms of computation time by, for example, an implementation using OpenMP, but Hadoop won in terms of popularity because it was easy to use.

The parallelized computation efforts described above make it possible to process large amounts of data. Besides the obvious application of applying existing methods to increasingly large datasets, the increase in computation power also leads to novel large-scale machine-learning approaches. One example is the recent work from Le et al. (2011) in which a dataset of ten million images was used to teach a face detector using only unlabelled data. Using the resulting features in an object recognition task resulted in a performance increase of 70 % over the state of the art (Le et al. 2011). Utilizing large amounts of data to overcome the need for labelled training data could become an important trend. By using only unlabelled data, one of the biggest bottlenecks to the broad adoption of machine learning is bypassed. The use of unsupervised learning methods has its limitations though and it remains to be seen if similar techniques can also be applied in other application domains.

5.3.2 Stream Data Processing

Stream data mining was highlighted as a promising area of research by Ricardo Baeza-Yates in his interview. This technique relates to the technological capabilities needed to deal with data streams with high volume and high velocity, coming from sensors networks, or other online activities where a high number of users are involved.

5.3.2.1 RDF Data Stream Pattern Matching

Motivated by the huge amount of structured and unstructured data available on the web as continuous streams, streaming processing techniques using web technologies have recently appeared. In order to process data streams on the web, it is important to cope with openness and heterogeneity. A core issue of data stream processing systems is to process data in a certain time frame and to be able to query

for patterns. Additional desired features include static data support that will not change over time and can be used to enhance dynamic data. Temporal operators and time-based windows are also typically found in these systems, used to combine several RDF graphs with time dependencies. Some major developments in this area are C-SPARQL (Barbieri et al. 2010) ETALIS (Anicic et al. 2011), and SPARKWAVE (Komazec et al. 2012).

C-SPARQL is a language based on SPARQL (SPARQL Protocol and RDF Query Language) and extended with definitions for streams and time windows. Incoming triples are first materialized based on RDFS and then fed into the evaluation system. C-SPARQL does not provide true continuous pattern evaluation, due to the usage of RDF snapshots, which are evaluated periodically. However C-SPARQL's strength is in situations with significant amounts of static knowledge, which need to be combined with dynamic incoming data streams.

ETALIS is an event-processing system on top of SPARQL. As the pattern language component of SPARQL was extended with event-processing syntax, the pattern language is called EP-SPARQL. The supported features are temporal operators, out-of-order evaluation, aggregate functions, several garbage collection modes, and different consumption strategies.

SPARKWAVE provides continuous pattern matching over schema-enhanced RDF data streams. In contrast to the C-SPARQL and EP-SPARQL, SPARKWAVE is fixed regarding the utilized schema and does not support temporal operators or aggregate functions. The benefit of having a fixed schema and no complex reasoning is that the system can optimize and pre-calculate at the initialization phase the used pattern structure in memory, thus leading to high throughput when processing incoming RDF data.

5.3.2.2 Complex Event Processing

One insight of the interviews is that big data stream technologies can be classified according to (1) complex event-processing engines, and (2) highly scalable stream processing infrastructures. Complex event-processing engines focus on language and execution aspects of the business logic, while stream processing infrastructure provides the communication framework for processing asynchronous messages on a large scale.

Complex event processing (CEP) describes a set of technologies that are able to process events "in stream", i.e. in contrast to batch processing where data is inserted into a database and polled at regular intervals for further analysis. The advantages of CEP systems are their capability to process potentially large amounts of events in real time. The name complex event processing is due to the fact that simple events, e.g. from sensors or other operational data, can be correlated and processed generating more complex events. Such processing may happen in multiple steps, eventually generating an event of interest triggering a human operator or some business intelligence.

As Voisard and Ziekow point out, an event-based system "encompasses a large range of functionalities on various technological levels (e.g., language, execution, or communication)" (Voisard and Ziekow 2011). They provide a comprehensive survey that aids the understanding and classification of complex event-processing systems.

For big data stream analytics, it is a key capability that complex event-processing systems are able to scale out in order to process all incoming events in a timely fashion as required by the application domain. For instance the smart meter data of a large utility company may generate millions or even billions of events per second that may be analysed in order to maintain the operational reliability of the electricity grid. Additionally, coping with the semantic heterogeneity behind multiple data sources in a distributed event generation environment is a fundamental capability for big data scenarios. There are emerging automated semantic event-matching approaches (Hasan and Curry 2014) that target scenarios with heterogeneous event types. Examples of complex event-processing engines include the SAP Sybase Event Stream Processor, IBM InfoSphere Stream,[6] and ruleCore[7] to name just a few.

5.3.3   Use of Linked Data and Semantic Approaches to Big Data Analysis

According to Tim Berners-Lee and his colleagues (Bizer et al. 2009), "Linked Data is simply about using the Web to create typed links between data from different sources". Linked data refers to machine-readable data, linked to other datasets and published on the web according to a set of best practices built upon web technologies such as HTTP (Hypertext Transfer Protocol), RDF, and URIs (Uniform Resource Identifier).[8] Semantic technologies such as SPARQL, OWL, and RDF allow one to manage and deal with these. Building on the principles of Linked Data, a dataspace groups all relevant data sources into a unified shared repository (Heath and Bizer 2011). Hence, a dataspace offers a good solution to cover the heterogeneity of the web (large-scale integration) and deal with broad and specific types of data.

Linked data and semantic approaches to big data analysis have been highlighted by a number of interviewees including Sören Auer, François Bancilhon, Richard Benjamins, Hjalmar Gislason, Frank van Harmelen, Jim Hendler, Peter Mika, and Jeni Tennison. These technologies were highlighted as they address important challenges related to big data including efficient indexing, entities extraction and classification, and search over data found on the web.

---

[6] http://www-01.ibm.com/software/data/infosphere/streams, last visited 25/02/2014.

[7] RuleCore Homepage, http://www.rulecore.com/, last visited 13/02/2014.

[8] http://www.w3.org/standards/semanticweb/data

5.3.3.1   Entity Summarization

To the best of our knowledge, entity summarization was first mentioned in Cheng et al. (2008). The authors present Falcons which "... provides keyword-based search for Semantic Web entities". Next to features such as concept search, ontology and class recommendation, and keyword-based search, the system also describes a popularity-based approach for ranking statements an entity is involved in. Further, the authors also describe the use of the MMR technique (Carbonell and Jade 1998) to re-rank statements to account for diversity. In a later publication (Cheng 2011), entity summarization requires "... ranking data elements according to how much they help identify the underlying entity". This statement accounts for the most common definition of entity summarization: the ranking and selection of statements that identify or define an entity.

In Singhal (2012), the author introduces Google's Knowledge Graph. Next to entity disambiguation ("Find the right thing") and exploratory search ("Go deeper and broader"), the knowledge graph also provides summaries of entities, i.e. "get the best summary". Although not explained in detail, Google points out that they use the search queries of users for the summaries.[9] For the knowledge graph summaries, Google uses a unique dataset of millions of daily queries in order to provide concise summaries. Such a dataset is, however, not available to all content providers.

As an alternative, Thalhammer et al. (2012b) suggest using the background data of consumption patterns of items in order to derive summaries of movie entities. The idea stems from the field of recommender systems where item neighbourhoods can be derived by the co-consumption behaviour of users (i.e. through analysing the user-item matrix).

A first attempt to standardize the evaluation of entity summarization is provided by Thalhammer et al. (2012a). The authors suggest a game with a purpose (GWAP) in order to produce a reference dataset for entity summarization. In the description, the game is designed as a quiz about movie entities from Freebase. In their evaluation, the authors compare the summaries produced by Singhal (2012) and the summaries of Thalhammer et al. (2012b).

5.3.3.2   Data Abstraction Based on Ontologies and Communication Workflow Patterns

The problem of communication on the web, as well as beyond it, is not trivial, considering the rapidly increasing amount of channels (content sharing platforms, social media and networks, variety of devices) and audiences to be reached. To address this problem, technological solutions are being developed such as the one presented by Fensel et al. (2012) based on semantics. Data management via

---

[9] http://insidesearch.blogspot.co.at/2012/05/introducing-knowledge-graph-things-not.html

semantic techniques can certainly facilitate the communication abstraction and also increase automation and reduce the overall effort.

Inspired by the work of Mika (2005), eCommunication workflow patterns (e.g. typical query response patterns for online communication), which are usable and adaptable to the needs of the social web, can be defined (Stavrakantonakis 2013a, b). Moreover, there is an interest in social network interactions (Fuentes-Fernandez et al. 2012). The authors of the last work coined "social property" as a network of activity theory concepts with a given meaning. Social properties are considered as "patterns that represent knowledge grounded in the social sciences about motivation, behaviour, organization, interaction" (Fuentes-Fernandez et al. 2012). The results of this research direction combined with the generic work flow patterns described in Van Der Aalst et al. (2003) are highly relevant with the materialization of the communication patterns. The design of the patterns is also related to the collaboration among the various agents as described in Dorn et al. (2012) in the scope of the social workflows. Aside from the social properties, the work described in Rowe et al. (2011) introduces the usage of ontologies in the modelling of the user's activities in conjunction with content and sentiment. In the context of the approach, modelling behaviours enable one to identify patterns in communication problems and understand the dynamics in discussions in order to discover ways of engaging more efficiently with the public in the social web. Several researchers have proposed the realization of context-aware work flows (Wieland et al. 2007) and social collaboration processes (Liptchinsky et al. 2012), which are related to the idea of modelling the related actors and artefacts in order to enable adaptiveness and personalization in the communication patterns infrastructure.

5.4 Future Requirements and Emerging Trends for Big Data Analysis

5.4.1 Future Requirements for Big Data Analysis

5.4.1.1 Next Generation Big Data Technologies

Current big data technologies such as Apache Hadoop have matured well over the years into platforms that are widely used within various industries. Several of our interviewees have identified future requirements that the next generation of big data technologies should address:

- *Handle the growth of the Internet* (Baeza-Yates)—as more users come online big data technologies will need to handle larger volumes of data.
- *Process complex data types* (Baeza-Yates)—data such as graph data and possible other types of more complicated data structures need to be easily processed by big data technologies.

- *Real-time processing* (Baeza-Yates)—big data processing was initially carried out in batches of historical data. In recent years, stream processing systems such as Apache Storm have become available and enable new application capabilities. This technology is relatively new and needs to be developed further.
- *Concurrent data processing* (Baeza-Yates)—being able to process large quantities of data concurrently is very useful for handling large volumes of users at the same time.
- *Dynamic orchestration of services in multi-server and cloud contexts* (Tori)—most platforms today are not suitable for the cloud and keeping data consistent between different data stores is challenging.
- *Efficient indexing* (Mika)—indexing is fundamental to the online lookup of data and is therefore essential in managing large collections of documents and their associated metadata.

5.4.1.2   Simplicity

The simplicity of big data technologies refers to how easily developers are able to acquire the technology and use it in their specific environment. Simplicity is important as it leads to a higher adoptability of the technology (Baeza-Yates). Several of our interviewees have identified the critical role of simplicity in current and future big data technologies.

The success of Hadoop and MapReduce is mainly due to its simplicity (Tori). Other big data platforms are available that can be considered as more powerful, but have a smaller community of users because their adoption is harder to manage. Similarly, Linked Data technologies, for example, RDF SPARQL, have been reported as overly complex and containing too steep a learning curve (Gislason). Such technologies seem to be over-designed and overly complicated—suitable only for use by specialists.

Overall, there exist some very mature technologies for big data analytics, but these technologies need to be industrialized and made accessible to everyone (Benjamins). People outside of the core big data community should become aware of the possibilities of big data, to obtain wider support (Das). Big data is moving beyond the Internet industry and into other non-technical industries. An easy-to-use big data platform will help in the adoption of big data technologies by non-technical industries.

5.4.1.3   Data

An obvious key ingredient to big data solutions is the data itself. Our interviewees identified several issues that need to be addressed.

Large companies such as Google and Facebook are working on big data and they will focus their energies on certain areas and not on others. EU involvement could

support a big data ecosystem that encourages a variety of small, medium, and large players, where regulation is effective and data is open (Thompson).

In doing so, it is important to realize that there is far more data out there than most people realize and this data could help us to make better decisions to identify threats and see opportunities. A lot of the data needed already exists, but it is not easy to find and use this data. Solving this issue will help businesses, policy makers, and end users in decision-making. Just making more of the world's data available at people's fingertips will have a substantial effect overall. There will be a significant impact for this item in emergency situations such as earthquakes and other natural disasters (Halevy) (Gislason).

However, making data available in pre-Internet companies and organizations is difficult. In Internet companies, there was a focus on using collected data for analytic purposes from the very beginning. Pre-Internet companies face issues with privacy, legal as well as technical, and process restrictions in repurposing the data. This holds even for data that is already available in digital form, such as call detail records for telephone companies. The processes around storing and using such data were never set up with the intention of using the data for analytics (Benjamins).

Open data initiatives can play an important role in helping companies and organizations get the most out of data. Once a dataset has gone through the necessary validations with regard to privacy and other restrictions, it can be reused for multiple purposes by different companies and organizations and can serve as a platform for new business (Hendler). It is therefore important to invest in processes and legislation that support open data initiatives. Achieving an acceptable policy seems challenging. As one of our interviewees' notes, there is an inherent tension between open data and privacy—it may not be possible to truly have both (Tori). But also closed datasets should be addressed. A lot of valuable information, such as cell phone data, is currently closed and owned by the telecom industry. The EU should look into ways to make such data available to the big data community, while taking into account the associated cost of making the data open. Also, how the telecom industry can benefit from making data open while taking into account any privacy concerns (Das). The web can also serve as an important data source. Companies such as Data Publica rely on snapshots of the web (which are 60–70 terabytes) to support online services. Freely available versions of web snapshots are available, but more up-to-date versions are preferred. These do not necessarily have to be free, but cheap. The big web players such as Google and Facebook have access to data related to searches and social networks that have important societal benefit. For example, dynamic social processes such as the spread of disease or rates of employment are often most accurately tracked by Google searches. The EU may want to prioritize the European equivalent of these analogous to the way the Chinese have cloned Google and Twitter (Bancilhon).

As open datasets become more common, it becomes increasingly challenging to discover the dataset needed. One prediction estimates that by 2015 there will be over 10 million datasets available on the web (Hendler). Valuable lessons can be learnt from how document discovery evolved on the web. Early on there was a

registry—all of the web could be listed on a single web page; then users and organizations had their own lists; then lists of lists. Later Google came to dominate by providing metrics on how documents link to other documents. If an analogy is drawn to the data area, it is currently in the registry era. It needs crawlers to find big datasets, good dataset metadata on contents, links between related datasets, and a relevant dataset ranking mechanism (analogous to page rank). A discovery mechanism that can only work with good quality data will drive data owners to publish their data in a better way, analogous to the way that search engine optimization (SEO) drives the quality of the current web (Tennison).

5.4.1.4   Languages

Most of the big data technologies originated in the United States and therefore have primarily been created with the English language in mind. The majority of the Internet companies serve an international audience and many of their services are eventually translated into other languages. Most services are initially launched in English though and are only translated once they gain popularity. Furthermore, certain language-related technology optimizations (e.g. search engine optimizations) might work well for English, but not for other languages. In any case, languages need to be taken into account at the very beginning, especially in Europe, and should play an import role in creating big data architectures (Halevy).

5.4.2   Emerging Paradigms for Big Data Analysis

5.4.2.1   Communities

The rise of the Internet makes it possible to quickly reach a large audience and grow communities around topics of interest. Big data is starting to play an increasingly important role in that development. Our interviewees have mentioned this emerging paradigm on a number of occasions.

- **Rise of data journalists:** Who are able to write interesting articles based on data uploaded by the public to infrastructure such as the Google Fusion Tables. The Guardian journalist Simon Rogers won the Best UK Internet Journalist award for his work[10] based on this platform. A feature of journalistic take-up is that data blogs typically have a high dissemination impact (Halevy).
- **Community engagement in local political issues:** Two months after the school massacre in Connecticut[11] local citizens started looking at data related to gun

---

[10] http://www.oii.ox.ac.uk/news/?id=576

[11] http://en.wikipedia.org/wiki/Sandy_Hook_Elementary_School_shooting

permit applications in two locations and exposed this on a map.[12] This led to a huge discussion on the related issues (Halevy).

- **Engagement through community data collection and analysis:** The company COSM (formerly Pachube) has been driving a number of community-led efforts. The main idea behind these is that the way data is collected introduces specific slants on how the data can be interpreted and used. Getting communities involved has various benefits: the number of data collection points can be dramatically increased; communities will often create bespoke tools for the particular situation and to handle any problems in data collection; and citizen engagement is increased significantly.

  In one example, the company crowd sourced real-time radiation monitoring in Japan following the problem with reactors in Fukushima. There are now hundreds of radiation-related feeds from Japan on Pachube, monitoring conditions in real time and underpinning more than half a dozen incredibly valuable applications built by people around the world. These combine "official" data, "unofficial" data, and also real-time networked Geiger counter measurements contributed by concerned citizens (Haque).

- **Community engagement to educate and improve scientific involvement:** Communities can be very useful in collecting data. Participation in such projects allows the public to obtain a better understanding of certain scientific activities and therefore helps to educate people in these topics. That increase in understanding will further stimulate the development and appreciation of upcoming technologies and therefore result in a positive self-reinforcing cycle (Thompson).

- **Crowdsourcing to improve data accuracy:** Through crowdsourcing the precision of released UK Government data on the location of bus stops was dramatically increased (Hendler).

These efforts play well into the future requirements section on data. A community-driven approach to creating datasets will stimulate data quality and lead to even more datasets becoming publicly available.

5.4.2.2   Academic Impact

The availability of large datasets will impact academia (Tori) for two reasons. First, public datasets can be used by researchers from disciplines such as social science and economics to support their research activities. Second, a platform for sharing academic dataset will stimulate reuse and improve the quality of studied datasets. Sharing datasets also allows others to add additional annotations to the data, which is generally an expensive task.

---

[12] http://tinyurl.com/kvlv64l

Next to seeing big data technologies affecting other scientific disciplines, other scientific disciplines are being brought into computer science. Big Internet companies like Yahoo are hiring social scientists, including psychologists and economists, to increase the effectiveness of analysis tools (Mika). More generally speaking, as the analysis of data in various domains continues an increasing need for domain experts arises.

5.5   Sectors Case Studies for Big Data Analysis

This section describes several big data case studies outlining the stakeholders involved, where applicable, and the relationship between technology and the overall sector context. In particular, it covers the following sectors: the public sector, health sector, retail sector, logistics, and finally the financial sector. In many cases the descriptions are supported by the interviews that were conducted, and add further evidence of the enormous potential for big data.

5.5.1   Public Sector

Smart cities generate data from sensors, social media, citizen mobile reports, and municipality data such as tax data. Big data technologies are used to process the large datasets that cities generate to impact society and businesses (Baeza-Yates). This section discusses how big data technologies utilize smart city data to provide applications in traffic and emergency response.

5.5.1.1   Traffic

Smart city sensors that can be used for applications in traffic include induction loop detection, traffic cameras, and license plate recognition cameras (LPR). Induction loops can be used for counting traffic volume at a particular point. Traffic cameras can be combined with video analytic solutions to automatically extract statistics such as the number of cars passing and average speed of traffic. License plate recognition is a camera-based technology that can track license plates throughout the city using multiple cameras. All these forms of sensing help in estimating traffic statistics, although they vary in degree of accuracy and reliability.

Deploying such technology on a city-wide level results in large datasets that can be used for day-to-day operations, as well as applications such as anomaly detection and support in planning operations. In terms of big data analysis, the most interesting application is anomaly detection. The system can learn from historical data what is considered to be normal traffic behaviour for the time of the day and the day of the week and detect deviations from the norm to inform operators in a command

and control centre of possible incidents that require attention (Thajchayapong and Barria 2010). Such an approach becomes even more powerful when combining the data from multiple locations using data fusion to get more accurate estimates of the traffic statistics that allow the detection of more complex scenarios.

5.5.1.2 Emergency Response

Cities equipped with sensors can benefit during emergencies by obtaining actionable information that can aid in decision-making. Of particular interest is the possibility to use social media analytics during emergency response. Social media networks provide a constant flow of information that can be used as a low-cost global sensing network for gathering near real-time information about an emergency. Although people post a lot of unrelated information on social media networks, any information about the emergency can be very valuable to emergency response teams. Accurate data can help in obtaining the correct situational awareness picture of the emergency, consequently enabling a more efficient and faster response that can reduce casualties and overall damage (Van Kasteren et al. 2014).

Social media analytics is used to process large volumes of social media posts, such as tweets, to identify clusters of posts centred around the same topic (high content overlap), same area (for posts that contain GPS tags), and around the same time. Clusters of posts are the result of high social network activity in an area. This can be an indication of a landmark (e.g. the Eiffel tower), a planned event (e.g. a sports match), or an unplanned event (e.g. an accident). Landmark sites have high tweet volumes throughout the year and can therefore be easily filtered out. For the remaining events machine-learning classifiers are used to automatically recognize which clusters are of interest for an emergency response operator (Walther and Kaisser 2013).

Using social media data for purposes that it was not originally intended for is just a single example of the significant impact that can occur when the right data is presented to the right people at the right time. Some of our interviewees explained that there is far more data out there than most people realize and this data could help us to make better decisions to identify threats and see opportunities. A lot of the data needed already exists, but it is not always easy to find and use this data (Gislason) (Halevy).

5.5.2 *Health*

The previous section spoke of the data that is repurposed in applications that differ strongly from the original application that generated the data. Such cases also exist in the healthcare sector. For example, dynamic social processes such as the spread of disease can be accurately tracked by Google searches (Bancilhon) and call detail

records from Telefonica have been used to measure the impact of epidemic alerts on human mobility (Frias-Martinez et al. 2012).

Big data analytics can be used to solve significant problems globally. The EU is therefore advised to produce solutions that solve global problems rather than focus solely on problems that affect the EU (Thompson). An example is the construction of clean water wells in Africa. The decision on where to locate wells is based on spreadsheets that may contain data that has not been updated for 2 years. Given that new wells can stop working after 6 months this causes unnecessary hardship and more (Halevy). Technology might offer a solution, either by allowing citizen reports or by inferring the use of wells from other data sources.

The impact in local healthcare is expected to be enormous. Various technological projects are aimed at realizing home healthcare, where at the very least people are able to record health-related measurements in their own homes. When combined with projects such as smart home solutions, it is possible to create rich datasets consisting of both health data and all kinds of behavioural data that can help tremendously in establishing a diagnosis, as well as getting a better understanding of disease onset and development.

There are, however, very strong privacy concerns in the healthcare sector that are likely to block many of these developments until they are resolved. Professor Marco Viceconti from the University of Sheffield outlined in his interview how certain recent developments such as k-anonymity can help protect privacy. A dataset has k-anonymity protection if the information for each individual in the dataset cannot be distinguished from at least $k - 1$ individuals whose information also appears in the dataset (Sweeney 2002). Professor Viceconti envisions a future system that can automatically protect privacy by serving as a membrane between a patient and an institute using the data, where data can flow both ways and all the necessary privacy policies and anonymization processes are executed automatically in between. Such a system would benefit both the patient, by providing a more accurate diagnosis, and the institute, by allowing research using real-world data.

5.5.3 Retail

O2 UK together with Telefónica Digital recently launched a service called Telefónica Dynamic Insights. This service takes all UK mobile data, including location, timing of calls and texts, and also when customers move from one mast to another. This data is mapped and repurposed for the retail industry. The data is first anonymized, aggregated, and placed in the cloud. Then analytics are run which calculate where people live, where they work, and where they are in transit. If this data is then combined with anonymized customer relationship management (CRM) data, it can determine the type of people who pass by a particular shop at a specific time-point. It can also calculate the type of people who visit a shop, where they live, and where else they shop (termed catchment).

This service supports real-estate management for retailers and contrasts well with present practice. What retailers do today is that they hire students with clickers just to count the number of people who walk past the shop, leading to data that is far less detailed. The service is thus solving an existing problem in a new way. The service can be run on a weekly or daily basis and provides completely new business opportunities. In addition to retail the service could be run in other sectors, for example, within the public sector it could analyse who walks past an underground station. Combining mobile data with preference data could open up new propositions for existing and new industries. This example is a taste of what is to come, the sum of which will definitely improve the competitiveness of European industry (Benjamins).

5.5.4   Logistics

In the United States, 45 % of fruits and vegetables reach the plate of the consumer and in Europe 55 % reaches the plate. Close to half of what is produced is lost. This is a big data problem: collecting data over the overall supply chain, analysing systems related to the distributed food, and identifying leaks and bottlenecks in the process would have an enormous impact. If implemented there would be a better handle on prices and a fairer distribution of wealth among all the agents in the food supply chain. Big data technology is important and so is access to the right data and data sources (Bancilhon).

5.5.5   Finance

The World Bank is an organization that aims to end extreme poverty and promote shared prosperity. Their operations strongly rely on accurate information and they are using big data analytics to support their activities. They plan to organize competitions to drive the analytic capabilities to obtain an alternative measure for poverty and to detect financial corruption and fraud at an early stage.

In terms of poverty, an important driver is to get more real-time estimates of poverty, which make it possible to make better short-term decisions. Three examples of information sources that are currently being explored to obtain the information needed are: (1) Twitter data can be used to look for indicators of social and economic well-being; (2) poverty maps can be merged with alternative data sources such as satellite imagery to identify paved roads and support decisions in micro financing; and (3) web data can be scraped to get pricing data from supermarkets that help in poverty estimation.

Corruption is currently dealt with reactively, meaning actions are only taken once corruption has been reported to the Worldbank. On average only 30 % of the money is retrieved in corruption cases when dealt with reactively. Big data

analytics will make more proactive approaches feasible, resulting in higher returns. This requires creating richer profiles of the companies and the partners that they work with. Data mining this in-depth profile data together with other data sources would make it possible to identify risk-related patterns.

Overall, it is important for the Worldbank to be able to make decisions, move resources, and make investment options available as fast as possible through the right people at the right time. Doing this based on limited sets of old data is not sustainable in the medium to long term. Accurate and real-time information is critical during the decision-making process. For example, if there is a recession looming, one needs to respond before it happens. If a natural disaster occurs, making decisions based on data available directly from the field rather than a 3-year-old dataset is highly desirable (Das).

5.6   Conclusions

Big data analysis is a fundamental part of the big data value chain. We can caricature this process using an old English saying that what this component achieves is to "turn lead into gold". Large volumes of data which may be heterogeneous with respect to encoding mechanism, format, structure, underlying semantics, provenance, reliability, and quality is turned into data which is usable.

As such big data analysis comprises a collection of techniques and tools some of which are old mechanisms recast to face the challenges raised by the three Vs (e.g. large-scale reasoning) and some of which are new (e.g. stream reasoning).

The insights gathered on big data analysis presented here are based upon 19 interviews with leading players in large and small industries and visionaries from Europe and the United States. The choice was taken to interview senior staff members who have a leadership role in large multinationals, technologists who work at the coalface with big data, founders and CEOs of the new breed of SMEs that are already producing value from big data, and academic leaders in the field.

From our analysis it is clear that delivering highly scalable data analysis and reasoning mechanisms that are associated with an ecosystem of accessible and usable tools will produce significant benefits for Europe. The impact will be both economic and social. Current business models and process will be radically transformed for economic and social benefit. The case study of reducing the amount of food wasted within the global food production life cycle is a prime example of this type of potential for big data.

To summarize, big data analysis is an essential part of the overall big data value chain which promises to have significant economic and social impact in the European Union in the near to medium term. Without big data analysis the rest of the chain does not function. As one of our interviewees stated in a recent discussion on the relationship between data analysis and data analytics:

Analytics without data is worthless. Analytics with bad data is dangerous. Analytics with
good data is the objective.[13]

We wholeheartedly agree.

References

Anicic, D., Fodor, P., Rudolph, S., Stühmer, R., Stojanovic, N., & Studer, R. (2011). ETALIS:
    Rule-based reasoning in event processing. In: *Reasoning in event-based distributed systems*
    (pp. 99–124). Studies in Computational Intelligence, vol. 347, Springer.
Baeza-Yates, R. (2013). Yahoo. *BIG Project Interviews Series.*
Bancilhon, F. (2013). Data Publica. *BIG Project Interviews Series.*
Barbieri, D. F., Braga, D., Ceri, S., Della Valle, E., & Grossniklaus, M. (2010). C-SPARQL: A
    continuous query language for RDF data streams. *International Journal of Semantic Comput-
    ing, 4*(1), 3–125.
Benjamins, R. (2013). Telefonica. *BIG Project Interviews Series.*
Bhaduri, K., Das, K., Liu, K., Kargupta, H., & Ryan, J. (2011). *Distributed data mining bibliog-
    raphy.* http://www.cs.uinbc.edu/-hillol/DDM-BIB.
Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International Journal
    on Semantic Web and Information Systems, 5*(3), 1–22.
Carbonell, J., & Jade, G. (1998). The use of MMR, diversity-based reranking for reodering
    documents and producing summaries. *SIGIR* (pp. 335–336). Melbourne, Australia: ACM.
Cheng, G. T. (2011). RELIN: Relatedness and informativeness-based centrality for entity sum-
    marization. In L. W. In: Aroyo, *ISCW, Part I. LNCS vol. 7031* (pp. 114–129). Heidelberg:
    Springer.
Cheng, G., Ge, W., & Qu, Y. (2008). Falcons: Searching and browsing entities on the semantic
    web. In *Proceedings of the 17th International Conference on World Wide Web*
    (pp. 1101–1102). Beijing, China: ACM.
Chu, C., Kim, S. K., Lin, Y.-A., Yu, Y., Bradski, G., Ng, A. Y., et al. (2007). Map-reduce for
    machine learning on multicore. *Advances in Neural Information Processing Systems, 19*, 281.
Das, P. L. (2013).Worldbank. *BIG Project Interviews Series.*
Dorn, C., Taylor, R., & Dustdar, S. (2012). Flexible social workflows: Collaborations as human
    architecture. *IEEE Internet Computing, 16*(2), 72–77.
Fensel, D. (2007). Unifying reasoning and search to web scale. *IEEE Internet Computing, 11*(2),
    94–96.

---

[13] Richard Benjamins in a personal communication.

Fensel, A., Fensel, D., Leiter, B., Thalhammer, A. (2012). Effective and efficient online commu-
     nication: The channel model. In *Proceedings of International Conference on Data Technolo-*
     gies and Applications (DATA'12) (pp. 209–215), SciTePress, Rome, Italy, 25–27 July, 2012.
Fensel, D., van Harmelen, F., Andersson, B., Brennan, P., Cunningham, H., Della Valle, E.,
     et al. (2008). *Towards LarKC: A platform for web-scale reasoning*. Los Alamitos, CA: IEEE
     Computer Society Press.
Freitas, A., & Curry, E. (2014). Natural language queries over heterogeneous linked data graphs: A
     distributional-compositional semantics approach. In *Proceedings of the 19th International*
     Conference on Intelligent User Interfaces (IUI), Haifa.
Freitas, A., da Silva, J. C. P, O'Riain, S., & Curry, E. (2013). Distributional relational networks. In
     Proceedings AAAI Fall Symposium, Arlington.
Frias-Martinez, V., Rubio, A., & Frias-Martinez, E. (2012). Measuring the impact of epidemic
     alerts on human mobility (Vol. 12). Pervasive Urban Applications – PURBA.
Fuentes-Fernandez, R., Gomez-Sanz, J. J., & Pavon, J. (2012). User-oriented analysis of interac-
     tions in online social networks. *IEEE Intelligent Systems, 27*, 18–25.
Gislason, H. (2013). Datamarket.com. *BIG Project Interviews Series*.
Halevy, A. (2013). Google. *BIG Project Interviews Series*.
Haque, U. (2013). Cosm. *BIG Project Interviews Series*.
Hasan, S., & Curry, E. (2014). Approximate semantic matching of events for the internet of things.
     ACM Trans Internet Technology, 14, 1–23. doi:10.1145/2633684.
Heath, T., & Bizer, C. (2011) Linked data: Evolving the web into a global data space (1st edn). In
     Synthesis Lectures on the Semantic Web: Theory and Technology, 1(1): 1–136. Morgan &
     Claypool.
Hendler, J. (2012). RPI. *BIG Project Interviews Series*.
Komazec, S., Cerri, D., & Fensel, D. (2012). Sparkwave: Continuous schema-enhanced pattern
     matching over RDF data streams. In *Proceedings of the 6th ACM International Conference on*
     Distributed Event-Based Systems (DEBS '12) (pp. 58–68). New York, NY: ACM. doi:10.1145/
     2335484.2335491.
Le, Q. V, Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., et al. (2011). Building high-
     level features using large-scale unsupervised learning. *International Conference on Machine*
     Learning.
Liptchinsky, V., Khazankin, R., Truong, H., & Dustdar, S. (2012). A novel approach to mod eling
     context-aware and social collaboration processes. In: *Advanced Information Systems Engi-*
     neering, (pp. 565–580), Springer.
Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., & Hellerstein, J. M. (2010).
     GraphLab: A new framework for parallel machine learning. *The 26th Conference on Uncer-*
     tainty in Artificial Intelligence (UAI 2010), Catalina Island, California, July 8–11.
Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics.
     (pp. 522–536). *The Semantic Web - ISWC*.
Mika, P. (2013). Yahoo. *BIG Project Interviews Series*.
Novacek, V., Handschuh, S., & Decker S., (2011). Getting the meaning right: A complementary
     distributional layer for the web semantics. *International Semantic Web Conference*
     (1):504–519.
Rowe, M., Angeletou, S., & Alani, H. (2011). Predicting discussions on the social semantic web
     (pp. 405–420). *The Semantic Web: Research and Applications*.
Singhal, A. (2012). Introducing the knowledge graph. Retrieved from googleblog: http://
     googleblog.blogspot.com/2012/05/introducing-knowledge-graphthings
Stavrakantonakis, I. (2013a). Personal data and user modelling in tourism. In: *Information and*
     Communication Technologies in Tourism 2013 (pp 507–518).
Stavrakantonakis, I. (2013b). Semantically assisted Workflow Patterns for the Social Web. In
     Proceedings of the 10th Extended Semantic Web Conference ESWC 2013 PhD Symposium
     track. (pp. 692–696).

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557–570.

Tennison, J. (2013). Open data institute. *BIG Project Interviews Series*.

Thajchayapong, S., & Barria, J. A. (2010). Anomaly detection using microscopic traffic variables on freeway segments. *Transportation Research Board of the National Academies*, 10-2393.

Thalhammer, A., Knuth, M., & Sack, H. (2012a). Evaluating entity summarization using a game-based ground truth. *International Semantic Web Conference (2)* (pp. 350–361). Boston: Springer.

Thalhammer, A., Toma, I., Roa-Valverde, A. J., Fensel, D. (2012b). Leveraging usage data for linked data movie entity summarization. In *Proceedings of the 2nd International Workshop on Usage Analysis and the Web of Data*. Lyon, France: USEWOD co-located with WWW.

Thompson, B. (2013). BBC. *BIG Project Interviews Series*.

Tori, A. (2013). *BIG Project Interviews Series*.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research, 37*(1), 141–188.

Van Der Aalst, W., Ter Hofstede, A., Kiepuszewski, B., & Barros, A. (2003). Workflow patterns. *Distributed and Parallel Databases, 14*(1), 5–51.

Van Kasteren, T., Ulrich, B., Srinivasan, V., & Niessen, M. (2014). Analyzing tweets to aid situational awareness. In *36th European Conference on Information Retrieval*.

Voisard, A., & Ziekow, H. (2011). ARCHITECT: A layered framework for classifying technologies of event-based systems. *Information Systems, 36*(6), 937–957. doi:10.1016/j.is.2011.03.006.

Walther, M., & Kaisser, M. (2013). Geo-spatial event detection in the twitter stream. *Advances in Information Retrieval*, 356–367.

Wieland, M., Kopp, O., Nicklas, D., & Leymann, F (2007). Towards context-aware workflows. *In: CAiSE*. pp. 11–15.

Zillner, S., Rusitschka, S., Munne, R., Lippell, H., Vilela, F. L., Hussain, K., et al. (2014). D2.3.2. Final version of the sectorial requisites. Public deliverable of the EU-Project BIG (318062; ICT-2011.4.4).