Chapter 15
Cross-sectorial Requirements Analysis for Big Data Research

Tilman Becker, Edward Curry, Anja Jentzsch, and Walter Palmetshofer

15.1 Introduction

This chapter identifies the cross-sectorial requirements for big data research necessary to define a prioritized research roadmap based on expected impact. The aim of the roadmaps is to maximize and sustain the impact of big data technologies and applications in different industrial sectors by identifying and driving opportunities in Europe. The target audiences for the roadmaps are the different stakeholders involved in the big data ecosystem including industrial users of big data applications, technical providers of big data solutions, regulators, policy makers, researchers, and end users.

The first step toward the roadmap was to establish a list of cross-sectorial business requirements and goals from each of the industrial sectors covered in part of this book and in Zillner et al. (2014). The consolidated results comprise a prioritized set of cross-sector requirements that were used to define the technology, business, policy, and society roadmaps with action recommendations. This chapter presents a condensed version of the cross-sectorial consolidated requirements. It discusses each of the high-level and sub-level requirements together with the associated challenges that need to be tackled. Finally the chapter concludes with

T. Becker (✉)
German Research Centre for Artificial Intelligence (DFKI), Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
e-mail: tilman.becker@dfki.de

E. Curry
Insight Centre for Data Analytics, National University of Ireland Galway, Lower Dangan, Galway, Ireland
e-mail: edward.curry@insight-centre.org

A. Jentzsch • W. Palmetshofer
Open Knowledge Foundation (OKF), Singerstr. 109, 10179 Berlin, Germany
e-mail: anja.jentzsch@okfn.org; walter.palmetshofer@okfn.org

J.M. Cavanillas et al. (eds.), *New Horizons for a Data-Driven Economy*,
DOI 10.1007/978-3-319-21569-3_15

a prioritization of the cross-sectorial requirements. As far as possible, the roadmaps have been quantified to allow for a well-founded prioritization and action plans (e.g. policies).

15.2 Cross-sectorial Consolidated Requirements

In order to establish a common understanding of requirements as well as technology descriptions across domains, the sector-specific requirement labels were aligned. Each sector provided their requirements with the associated user needs, and similar and related requirements were merged, aligned, or restructured to create a homogenous set.

While most of the requirements exist within each of the sectors, the level of importance for the requirement in each sector varies. For the cross-sector analysis, any requirements that were identified by at least two sectors as being a significant requirement for that sector were included into the cross-sector roadmap definition. Thus, the initial list of 13 high-level requirements and 28 sub-level requirements was reduced to 5 high-level requirements and 12 sub-level requirements (see Table 15.1). Within this chapter, the discussion on each cross-sectorial requirement has been condensed and minor updates applied. Full details are available in Becker et al. (2014).

15.2.1 Data Management Engineering

The high-level requirement *data management engineering* aims at efficient strategies to manage heterogeneous data sources and technologies. Data management engineering has four sub-requirements:

- *Data enrichment*
- *Data integration*
- *Data sharing*
- *Real-time data transmission*

15.2.1.1 Data Enrichment

The sub-requirement *data enrichment* aims to make unstructured data understandable across domains, application, and value chains.

In the *health sector*, data enrichment is of high relevance, since 90 % of health data is only available in unstructured formats without semantic labels informing applications on the content of the data. In particular, approaches for the semantic annotation of medical images and medical text are needed.

Table 15.1 Consolidated cross-sectorial requirements (and demanding sectors)

Technological Requirement	Number of Demanding Sectors	Health	Public	Finance & Insurance	Energy & Transport	Telecom &Media	Retail	Manufacturing
Data Management Engineering	3		X			X	X	
Data Enrichment	2	X				X		
Data Integration	5	X	X	X		X	X	
Data Sharing	4	X	X	X	X			
Real-Time Data Transmission	3		X				X	X
Data Quality	3	X		X			X	
Data Improvement	2					X	X	
Data Security and Privacy	7	X	X	X	X	X	X	X
Data Visualization and User Experience	2						X	X
Deep Data Analytics	3		X	X			X	
Modelling Simulation	3		X				X	X
Natural Language Analytics	3		X				X	X
Pattern Discovery	3		X	X		X		
Predictive Analytics	2		X			X		
Prescriptive Analytics	3				X	X	X	
Real-Time Insights	5		X		X	X	X	X
Usage Analytics	2			X		X		

In the *telecom and media sector*, data enrichment includes ontologies (e.g. eTOM SID), data transformation, addition of metadata, formats, etc., taking into account that the data sources are heterogeneous (including social media information, audio, customer data, and traffic data, for example). Data coming from different sources and in different formats, produced by heterogeneous systems, have to be processed together. In order to address these requirements, the following challenges need to be tackled:

- Information extraction from text
- Image understanding algorithms
- Standardized annotation framework

15.2.1.2 Data Sharing and Integration

The sub-requirement *data sharing and integration* aims to establish a basis for the seamless integration of multiple and diverse data sources into a big data platform. The lack of standardized data schemas, semantic data models, as well as the fragmentation of data ownership are important aspects that need to be tackled.

As of today, less than 30 % of *health data* is shared between healthcare providers (Accenture 2012). In order to enable seamless data sharing in the health and other domains, a standardized coding system and terminologies as well as data models are needed.

In the *telecom* sector, data has been collected for years and classified according to business standards based on eTOM (2014), but the data reference model does not yet contemplate the inclusion of social media data. A unified information system is required that includes data from both the telecom operator and the customer. Once this information model is available, it should be incorporated in the eTOM SID reference model and taken into account in big data telecom-specific solutions for all data (social and non-social) to be integrated.

In the *retail sector*, standardized product ontologies are needed to enable sharing of data between product manufacturers and retailers. Services to optimize operational decisions in retail are only possible with semantically annotated product data.

In the *public sector*, data sharing and integration are important to overcome the lack of standardization of data schemas and fragmentation of data ownership, to achieve the integration of multiple and diverse data sources into a big data platform. This is required in cases where data analysis has to be performed from data belonging to different domains and owners (e.g. different agencies in the public sector) or integrating heterogeneous external data (from open data, social networks, sensors, etc.).

In the *financial sector*, several factors have put organizations in a situation where a large number of different datasets lack interconnection and integration. Financial organizations recognize the potential value of interlinking such datasets to extract information that would be of value either to optimize operations, improve services to customers, or even create new business models. Existing technology can cover most of the requirements of the financial services industry, but the technology is still not widely implemented.

In order to address these requirements, the following challenges need to be tackled:

- Semantic data and knowledge models
- Context information
- Entity matching
- Scalable triple stores, key/value stores
- Facilitate core integration at data acquisition
- Best practice for sharing high-velocity and high-variety data
- Usability of semantic systems
- Metadata and data provenance frameworks

- Scalable automatic data/schema mapping mechanisms

15.2.1.3 Real-Time Data Transmission

The sub-requirement *real-time data transmission* aims at acquiring (sensor and event) information in real time.

In the *public sector*, this is closely related with the increasing capability of deploying sensors and Internet of Things scenarios, like in public safety and smart cities. Image sensors have followed Moore's Law, doubling megapixel density per dollar every 2 years (PWC 2014). Distributed processing and cleaning capabilities are required for image sensors in order to avoid overloading the transmission channels (Jobling 2013) and provide the required real-time analysis to feed situational awareness systems for decision-makers.

In the *manufacturing sector*, sensor data must be acquired at high sample rates and needs to be transmitted close to real time in order to be used effectively. Decisions can be made at central planning, command, and control points, or can be made at a local level in a distributed fashion. Data transmission must be sufficiently close to real time, greatly improving on the currently long intervals (hourly or greater) in which inventory data is sampled. The hostile working environment in manufacturing may hamper data transmission.

For the *retail sector*, it is important that the data from sensors inside the store are acquired in real time. This includes visual data from cameras and customer locations from positioning sensors.

In order to address these requirements, the following challenges need to be tackled:

- Distributed data processing and cleaning
- Read/write optimized storage solutions for high velocity data
- Near real-time processing of data streams

15.2.2 Data Quality

The high-level requirement, *data quality*, describes the need to capture and store high-quality data so that analytic applications can use the data as reliable input to produce valuable insights. Data quality has one sub-requirement:

- *Data improvement*

Big data applications in the *health sector* need to fulfil high data quality standards in order to derive reliable insights for health-related decisions. For instance, the features and parameter list used for describing patient health status needs to be standardized in order to enable the reliable comparison of patient (population) datasets.

In the *telecom and media sectors*, despite the fact that data has been collected already for years, there are still data quality issues that make the information un-exploitable without pre-processing.

In the *financial sector*, data quality is not a major issue in internally generated datasets, but information collected from external sources may not be fully reliable.

In order to address these requirements, the following challenges need to be tackled:

- Provenance management
- Human data interaction
- Unstructured data integration

15.2.2.1 Data Improvement

The sub-requirement *data improvement* aims at removing noise/redundant data, checking for trustworthiness, and adding missing data.

In the *telecom and media sectors*, this relates to the ability to improve the commercial offering of the service provider based on the available information in traditional systems, as well as advanced techniques such as predictive, speech, or prescriptive analytics.

In the *retail sector*, both sensor data and data extracted from web sources (i.e. product data and customer data) are error prone and need to be checked for trustworthiness. Therefore data improvement procedures are required that help to remove incorrect/redundant data and noise.

- Human validation via curation
- Automatic removal of large amounts of noise at scale
- Scalable semantic validation

15.2.3 Data Security and Privacy

The high-level requirement *data security and privacy* describes the need to protect highly sensitive business and personal data from unauthorized access. Thus, it addresses the availability of legal procedures and the technical means that allow the secure sharing of data.

In *healthcare applications*, a strong emphasis has to be put on data privacy and security since some of the usual privacy protection approaches could be bypassed by the nature of big data. For instance, in terms of health-related data, anonymization is a well-established approach to de-identify personal data. Nevertheless, the anonymized data could be re-identified (El Emam et al. 2014) when aggregating big data from different data sources.

Big data applications in *retail* require the storage of personal information of customers in order for the retailer to be able to provide tailored services. It is very

important that this data is stored securely to ensure the protection of customer privacy.

In the *manufacturing sector*, there are conflicting interests in storing data on products for easy retrieval and protection of data from unauthorized retrieval. Data collected during production and use may well contain proprietary information concerning internal business processes. Intellectual property needs to be protected as far as it is encoded in product and production data. Regulations for data ownership need to be established, e.g., what access may the manufacturer of a production machine have to its usage data.

Privacy protection for workers interacting in an Industry 4.0 environment needs to be established. Data encryption and access control into object memories needs to be integrated. European and worldwide regulations need to be harmonized. There is a need for data privacy regulations and transparent privacy protection.

In the *telecom and media sector*, one of the main concerns is that big data policies apply to personal data, i.e., to data relating to an identified or identifiable person. However, it is not clear whether the core privacy principles of the regulation apply to newly discovered knowledge or information derived from personal data, especially when the data has been anonymized or generalized by being transformed into group profiles. Privacy is a major concern which can compromise the end users' trust, which is essential for big data to be exploited by service providers. An Ovum (2013) Consumer Insights Survey revealed that 68 % of Internet users across 11 countries around the world would select a "Do-Not-Track" feature if it was easily available. This clearly highlights some amount of end users' antipathy towards online tracking. Privacy and trust is an important barrier since data must be rich in order for businesses to use it.

Finding solutions to ensure data security and privacy may unlock the massive potential of big data in the *public sector*. Advances in the protection and privacy of data are key for the public sector, as it may allow the analysis of huge amounts of data owned by the public sector without disclosing sensitive information. In many cases, the public sector regulations restrict the use of data for different purposes for which it was collected. Privacy and security issues are also preventing the use of cloud infrastructures (e.g. processing, storage) by many public agencies that deal with sensitive data. A new approach to security in cloud infrastructure may eliminate this barrier.

Data security and privacy requirements appear in the *financial sector* in the context of building new business models based on data collected by financial services institutions from their customers (individuals). Innovative services could be created with technologies that reconcile the use of data and privacy requirements. In order to address these requirements, the following challenges need to be tackled:

- Hash algorithms
- Secure data exchange
- De-identification and anonymization algorithms
- Data storage technologies to encrypted storage and DBs; proxy re-encryption between domains; automatic privacy-protection

- Advances in "privacy by design" to link analytics needs with protective controls in processing and storage
- Data provenance to enable usage transparency and metadata for privacy information

15.2.4 Data Visualization and User Experience

The high-level requirement *data visualization and user experience* describes the need to adapt the visualization to the user. This is possible by reducing the complexity of data, data inter-relations, and the results of data analysis.

In *retail* it will be very important to adapt the information visualization to the specific customer. An example of this would be tailored advertisements, which fit the profile of the customer.

In *manufacturing* human decision-making and guidance need to be supported on all levels: from the production floor to high-level management. Appropriate data visualization tools must be available and integrated to support browsing, controlling, and decision-making in the planning and execution process. This applies primarily to general big data but extends to and includes special visualization of spatiotemporal aspects of the manufacturing process for spatial and temporal analytics.

In order to address these requirements, the following challenges need to be tackled:

- Apply user modelling techniques to visual analytics
- High performance visualizations
- Large-scale visualization based on adaptive semantic frameworks
- Multimodal interfaces in hostile working environments
- Natural language processing for highly variable contexts
- Interactive visualization and visual queries

15.2.5 Deep Data Analytics

The high-level requirement *deep data analytics* is the application of sophisticated data processing techniques to yield information from multiple, typically large datasets comprised of both unstructured and semi-structured data. Deep data analysis has seven sub-requirements:

- *Modelling and simulation* covers domain-specific tools for modelling and simulation of events according to changes from past events.
- *Natural language analytics* aims at extracting information from unstructured sources (e.g. social media) to enable further analysis (for instance sentiment mining).

- *Pattern discovery* aims at identifying patterns and similarities.
- *Real-time insights* enable the analysis of real-time data for instant decision-making.
- *Usage analytics* provide analysis of the usage of product, service, resources, process, etc.
- *Predictive analytics* utilize a variety of statistical, modelling, data mining, and machine learning techniques to study recent and historical data to make predictions about the future.
- *Prescriptive analytics* focus on finding the best course of action for a given situation.

Prescriptive analytics belongs to a portfolio of analytic capabilities that include descriptive and predictive analytics. While descriptive analytics aims to provide insight into what has happened, and predictive analytics helps model and forecast what might happen, prescriptive analytics seeks to determine the best solution or outcome among various choices, given the known parameters.

In the *public sector*, deep data analytics can help in several scenarios where information should be extracted from data. In the scenario of monitoring and supervision of online gambling operators, the challenge is to detect specific criminal or illegal behaviours using pattern discovery to deliver real-time insights. Similar insights are needed in the supervision of markets regulated by the public sector (energy, telecommunications, stock markets, etc.).

Other application scenarios also need deep data analytics, as in the case of public safety in smart cities, where real-time insights can enable the analysis of fresh/real-time data for instant decision-making. In these scenarios, situational awareness systems can be built using real-time data provided by networks of sensors and near real-time data captured from social networks through natural language analytics. Smart cities situation awareness can also apply modelling and simulation tools for managing events (e.g. managing large crowds of people in public events) to anticipate the results from decisions taken to influence the current conditions in real-time.

Other application scenarios like predictive policing may require the use of predictive analytics to provide insights based on the learning from previous situations. This would allow for optimal security resources allocation, according to the prediction of incidents, which may be based on temporal patterns or related to specific events of any kind (sport events, weather conditions, or any other variable).

For the *telecom and media sectors*, deep data analytics are required in order to improve customer experience, either by tailoring the offerings, by improving customer care, or by proactively adapting resources (e.g. network) to meet the customer expectations in terms of service delivery. This can be achieved by obtaining a 360° customer view, which allows a better understanding of the customer and predicts their needs or demands. Advanced and flexible customer segmentation, knowing customer likes and dislikes, deeply analysing user habits,

customer interactions, etc., help communication and content service providers to find patterns and sentiment out of the data, allowing cross selling based on multiple factors. Since Quality of Experience (QoE) and customer satisfaction can differ very quickly (as mood does), analytics should ideally provide the means to calculate and automate the best next action in real time.

Historical and online analytical processing of big data will be adopted as the insights gained will make planning and operations more precise. Real-time analytics on the other hand still faces some technological challenges, which may well be the reason for the lack of adoption of real-time analytics in energy and transportation. Manual steps in typical data analytics processes, such as data wrangling, for example, do not scale for the speed and volume of data to be analysed in operational efficiency scenarios in energy and transportation optimization.

In the *retail sector*, operational decisions can be optimized by analysing unstructured data from the web. This can be information about upcoming regional events, weather data, or even potential natural disasters that can be extracted from social networks using natural language analytics. Data, like visual data from cameras, acquired from sensors inside the store needs to be analysed to extract specific patterns, such as patterns of customer movement. Customer segmentation is possible by analysing customer–product and customer–staff interactions. This information can also be used to run prescriptive analytics. These are required to allow intelligent inventory, intelligent staff scheduling, and floor plan/ product location optimization.

In order to address these requirements, the following challenges need to be tackled:

- Data integration, linking, and semantics
- Sentiment analysis
- Machine learning
- Integrating semantics into large-scale modelling and simulation environments
- Increasing scalability and robustness of information extraction, named entity recognition, machine learning, linked data, entity linking, and co-reference resolution
- Validation of pattern analytics outputs and natural language analytics outputs with humans via curation
- Integration of natural language analytics into data usage scenarios
- Semantic pattern technologies including stream pattern matching and scalable complex pattern matching
- Analytical databases to efficiently support predictive analytics
- Combining large-scale reasoning with statistical approaches
- Predictive maintenance: predict failures, determine maintenance intervals Support for failure analysis
- Extend predictive analytics to prescriptive analytics
- Complex event processing applies business rules (or other frameworks) continuously on defined (short) interval of real-time data stream with low latency

- In-memory technology, new visualization and interaction techniques, automatic system reactions to enable ad hoc queries on large datasets to be executed with minimal latencies
- Real-time and in-stream analytical processing

15.3 Prioritization of Cross-sectorial Requirements

An actionable roadmap should have clear selection criteria regarding the priority of all actions. In contrast to a technology roadmap for the context of a single company, a European technology roadmap needs to cover developments across different sectors. The process of defining the roadmap included an analysis of the big data market and feedback received from stakeholders. Through this analysis, a sense of what characteristics indicate higher or lower potential of big data technical requirements was reached.

As the basis for the ranking, a table-based approach was used that evaluated each candidate according to a number of applicable parameters. In each case, the parameters were collected with the goal of being sector independent. Quantitative parameters were used where possible and available.

In consultation with stakeholders, the following parameters were used to rank the various technical requirements. The ranking parameters included:

- Number of affected sectors
- Size of affected sector(s) in terms of % of GDP
- Estimated growth rate of the sector(s)
- Possible prognosticated estimated growth rate by the sector due to big data technologies
- Estimated export potential of the sector(s)
- Estimated cross-sectorial benefits
- Short-term low-hanging fruit

Using these insights, a prioritization composed of multiple parameters was created, which give a relative sense of which technological requirements might be poised for greater gains and which would face the lowest barriers. The ranking of cross-sectorial technical requirements is presented in Table 15.2 and is illustrated in Fig. 15.1, where colour indicates the level of estimated importance, and the size of the bubble the estimated affected sectors of the industries. It is important to note that these indices do not offer a full picture, but they do offer a reasonable sense of both potential availability and capture across sectors. There are certain limitations to this approach. Not all relevant numbers and inputs were available as the speed of technology development and adoption relies on several factors. The ranking relies on forecasts and estimates from third parties and the project team. As a consequence, it is not always possible to determine precise numbers for timelines and

Table 15.2 Prioritization of technical cross-sectorial requirements

Prioritization	Technological requirements	Score
Level 1: Urgent		
	Data security and privacy	78
	Data management engineering—data integration	69.25
	Deep data analytics—real-time insights	61.5
	Data management engineering—data sharing	48.5
Level 2: Very important		
	Data quality	40.5
	Data management engineering—real-time data transmission	37
	Deep data analytics—modelling simulation	37
	Deep data analytics—natural language analytics	37
	Deep data analytics—pattern discovery	34.25
	Deep data analytics	31.75
	Data management engineering	31.5
Level 3: Important		
	Data management engineering—data enrichment	29.5
	Data visualization and user experience	29.5
	Deep data analytics—prescriptive analytics	29.5
	Deep data analytics—usage analytics	26.75
	Data quality—data improvement	24
	Deep data analytics—predictive analytics	20.75

specific impacts. Further investigation into these questions would be desirable for future research. Full details of the ranking process are available in (Becker, T., Jentzsch, A., & Palmetshofer, W. 2014).

15.4 Summary

The aim of the cross-sectorial roadmap is to maximize and sustain the impact of big data technologies and applications in the different industrial sectors by identifying and driving opportunities in Europe. While most of the requirements identified exist in some form within each sector, the level of importance of the requirements between specific sectors varies. For the cross-sector requirements, any requirements that were identified by at least two sectors as being a significant requirement for the sector were included into the cross-sector roadmap definition. This led to the

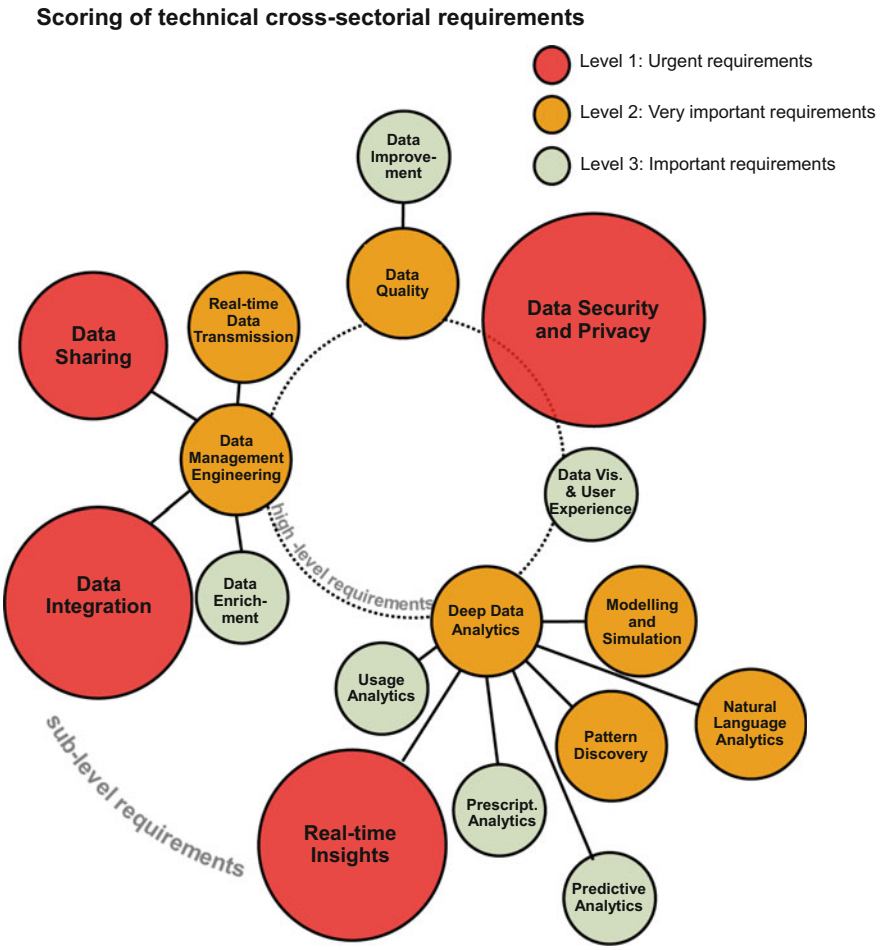

Fig. 15.1 Cross-sectorial requirements prioritized

identification of 5 high-level requirements and 12 sub-level requirements with associated challenges that need to be tackled.

Each cross-sectorial requirement was prioritized based on their expected impact. The consolidated results comprise a prioritized set of cross-sector requirements that were used to define the cross-sectorial roadmaps with associated action recommendations.

References

Accenture. (2012). *Connected health: The drive to integrated healthcare delivery*. Online: www.acccenture.com/connectedhealthstudy

Becker, T., Jentzsch, A., & Palmetshofer, W. (2014). *D2.5 Cross-sectorial roadmap consolidation*. Public deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

El Emam, K., Arbuckle, L., Koru, G., Eze, B., Gaudette, L., Neri, E., et al. (2014). De-identification methods for open health data: The case of the Heritage Health Prize claims dataset. *Journal of Medical Internet Research, 14*(1), e33.

eTOM. (2014). *TM forum*. Retrieved from Business Process Framework: http://www.tmforum.org/BestPracticesStandards/BusinessProcessFramework/1647/Home.html

Jobling, C. (2013, July 31). *Capturing, processing, and transmitting video: Opportunities and challenges*. Retrieved from Military embedded systems: http://mil-embedded.com/articles/capturing-processing-transmitting-video-opportunities-challenges/

Ovum. (2013). Retrieved from http://ovum.com/press_releases/ovum-predicts-turbulence-for-the-internet-economy-as-more-than-two-thirds-of-consumers-say-no-to-internet-tracking/

PWC. (2014). *Image sensor: Steady growth for new capabilities*. Retrieved from PWC: http://www.pwc.com/gx/en/technology/mobile-innovation/image-sensor-steady-growth-new-capabilities.jhtml

Zillner, S., Bretschneider, C., Oberkampf, H., Neurerer, S., Munné, R., Lippell, H., et al. (2014). *D2.4.2 Final version of sectors roadmap*. Public deliverable of the EU-Project BIG (318062; ICT-2011.4.4).