**Project Acronym:** BIG  
**Project Title:** Big Data Public Private Forum (BIG)  
**Project Number:** 318062  
**Instrument:** CSA  
**Thematic Priority:** ICT-2011.4.4

### D2.4.2 Final version of Sector’s Roadmap

<table>
<thead>
<tr>
<th>Work Package:</th>
<th>WP2 Strategy &amp; Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due Date:</td>
<td>31/07/2014</td>
</tr>
<tr>
<td>Submission Date:</td>
<td>07/11/2014</td>
</tr>
<tr>
<td>Start Date of Project:</td>
<td>01/09/2012</td>
</tr>
<tr>
<td>Duration of Project:</td>
<td>26 Months</td>
</tr>
<tr>
<td>Organisation Responsible of Deliverable:</td>
<td>ATOS</td>
</tr>
<tr>
<td>Version:</td>
<td>0.851</td>
</tr>
<tr>
<td>Status:</td>
<td>Final</td>
</tr>
</tbody>
</table>

#### Author name(s):  
Sonja Zillner, Claudia Bretschneider, Heiner Oberkampf (Methodology, Health chapter)  
Sabrina Neurerer (Health chapter)  
Ricard Munné (Public chapter)  
Helen Lippell (Telco & Media chapter)  
Felicia Lobillo Vilela (Telco & Media chapter)  
Elsa Prieto, Pedro Soria (Finance & Insurance chapter)  
Sebnem Rusitschka (Energy & Transport chapter)  
Tilman Becker (Manufacturing chapter)  
Denise Paradowski (Retail chapter)  
Siemens AG  
UIBK  
ATOS  
PA  
ATOS  
ATOS  
Siemens AG  
DFKI  
DFKI

#### Reviewer(s):  
Tilman Becker (Health and Telco & Media chapters)  
John Domingue (Public chapter)  
Helen Lippell (Finance & Insurance chapter)  
Nelia Lasierra (Energy, Transport, Manufacturing, Retail chapters)  
Walter Palmetshofer (Final Review)  
DFKI  
STI  
PA  
UIBK  
OKFN
| Dissemination level: | ☑️ PU - Public ☐ CO - Confidential, only for members of the consortium (including the Commission) ☐ RE - Restricted to a group specified by the consortium (including the Commission Services) |

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)
## Revision history

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Modified by</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>21/05/2014</td>
<td>Tilman Becker (DFKI)</td>
<td>TOC provided,</td>
</tr>
<tr>
<td>0.2</td>
<td>30/06/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Document template and first version of Scope and Methodology Section</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Contributions from Sonja Zillner (SIEMENS) to methodology (organizing the</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>workshop, developing templates and providing a description of the</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>methodology)</td>
</tr>
<tr>
<td>0.3</td>
<td>06/08/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Added Health and Retail chapters</td>
</tr>
<tr>
<td>0.4</td>
<td>20/08/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Added Telco, Media and Entertainment chapters</td>
</tr>
<tr>
<td>0.5</td>
<td>27/08/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Added Finance and Insurance chapters</td>
</tr>
<tr>
<td>0.6</td>
<td>12/09/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Added Public, Energy, Transport and Manufacturing Chapters</td>
</tr>
<tr>
<td>0.7</td>
<td>17/9/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Integrated new versions of Telco &amp; Media and Retail</td>
</tr>
<tr>
<td>0.8</td>
<td>13/10/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Integrated final versions of Public, Telco &amp; Media and Retail chapters</td>
</tr>
<tr>
<td>0.81</td>
<td>20/10/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Integrated final versions of Health, Finance and Energy &amp; Transport chapters</td>
</tr>
<tr>
<td>0.82</td>
<td>21/10/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Integrated final version of Manufacturing chapter</td>
</tr>
<tr>
<td>0.83</td>
<td>24/10/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Added text to Executive summary and review of manufacturing chapter</td>
</tr>
<tr>
<td>0.84</td>
<td>28/10/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Additional review changes to manufacturing chapter</td>
</tr>
<tr>
<td>0.85</td>
<td>07/11/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Final version according to reviewer comments</td>
</tr>
<tr>
<td>0.851</td>
<td>07/11/2014</td>
<td>Ricard Munné (ATOS)</td>
<td>Added Sonja Zillner in the 0.2 revision as contributor for methodology</td>
</tr>
</tbody>
</table>
The BIG Consortium (http://www.big-project.eu/) grants third parties the right to use and distribute all or parts of this document, provided that the BIG project and the document are properly referenced.

THIS DOCUMENT IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS DOCUMENT, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
Executive Summary

The snapshot of the current key-findings of the various sectors can preliminarily be summarized as follows:

Within the healthcare sector, we could identify many very promising big data applications within our sector analysis. However, in order to realise those applications, one needs to enable the seamless access to the various health data sets. As of today, the access to health data is only possible in a very constrained and limited manner. In order to improve this situation and for establishing the basis for the wide-spread implementation of big data applications in the healthcare sector, several technical requirements need to be addressed: 1) the content of unstructured health data (such as images or reports) is enriched by semantic annotation 2) data silos are conquered by means of efficient technologies for semantic data storage and exchange 3) technical means backed by legal frameworks to ensure the transparent sharing and exchange of health data are in place and finally 4) means for improving and assessing the data quality are available. We will detail the needed steps and discuss open R&D topics leading towards an efficient data management in the healthcare domain. By fostering the aspect of seamless health data access, one establishes the needed grounding for the implementation of promising value-creating big data-based healthcare applications.

The public sector needs to tackle the requirements derived from the effective application of big data to take a step ahead in terms of improving its effectiveness through the use of big data technologies. The sector needs to exploit the full potential of large scale data analysis and the integration and sharing of existing information to address the society demands for: (I) providing more transparency and fighting against fraud without spending additional public resources; (II) improving its operative efficiency through the analysis of the historical operational data through predictive, modelling and simulation tools providing better services to citizens. Another factor in which the development of big data can provide significant improvements is the area of safety applications, where the deployment of advanced natural language analytics, predictive and simulation tools can improve significantly the way to prevent and respond to emergencies.

The finance and insurance sector is particularly faced with the challenges of extremely large, dynamic, and heterogeneous sources of information. The everyday work and the business success of all decision makers in this industry depend on the availability of highly trustable, easily acquirable information, but also on analytics. Visualization has become an important tool to support decision support by providing new insights.

The telecom, media and entertainment industries are relatively well-advanced in utilising digital technologies and extremely interested in the potential of Big Data. However, the requirements identified by the BIG project indicate that there is much work to do at research and policy levels to support the growing ecosystem of diverse businesses engaged in analysing, enhancing and delivering content and data.

One of the main challenges within the telecom sector as far as Big Data technologies are concerned is related to the fact that, in spite of being a highly technological sector which has been collecting data for years already, there is still some strong inertia among decision makers to think in traditional DWH parameters, relational databases and practices. Besides, the telecom sector not only involves huge amounts of data but also complex data coming from the network, which is its main distinctive element, and also data generated from various sources – social networks, consumer behaviour, mobility and mobile/wireless communications; even sensor driven networks in machine-to-machine or device-to-device applications. The availability of data
scientists, who are familiar with data and machine learning (horizontal Big Data features), and at the same time are aware of the sector needs and specificity is also a problem for the generalised uptake of Big Data in telecom.

In line with historical practices in telecom, where proprietary platforms and solutions have always been preferred to open source ones, there are a number of emerging telecom-specific Big Data commercial platforms available in the market. It can be observed that they cover Big Data areas only partially (e.g. speech analytics or social media analytics seldom included among the list of features). This implies that a telecom player adopting Big Data will require several of these or other generic platforms integrated together involving multidisciplinary teams.

Whilst media organisations don’t handle the extreme volumes of data that say, scientific companies might do, the sector-specific challenges relate to ingesting, transforming, and joining up heterogeneous sources of data. This is necessary both for batch analysis for business intelligence, customer experience management, and other needs, but also real-time analysis particularly for news and press applications.

Another challenge that telecom, media and entertainment share with many other sectors is to find technical solutions which help businesses exploit Big Data but maintain compliance with all regulatory frameworks for data protection and privacy. This problem is exacerbated by uncertainty about the direction of travel for such regulation. Europe in general has more appetite for personal data protection than the US, meaning that the needs of companies to innovate have to be balanced against the rights of citizens to have their data protected.

Energy, and transportation sectors are very close regarding their big data value potential, but they are also at very different stages, regarding digitization, automation, and liberalization. Along with telecommunication, these sectors deliver the resource and infrastructure backbone for our society regarding energy, mobility, and information. Technically, they are also comparable because they can be modelled and represented as multimodal flow networks. Many of the efficient data structures and algorithms to tackle big data computing and analytics will be usable for both sectors, at the same time these two sectors will have many cross-optimization points.

The technology roadmap for energy and transportation concentrates on four prioritized areas: Scalable data access and sharing, whilst enhancing data interpretability as well as preserving privacy and confidentiality – which poses a typical trade-off in big data settings. Technologies such as encrypted storage, data-level encryption will be enablers. At the same time more investments will be needed to adapt promising research in linked data, data provenance, and differential privacy for the applications of big data in energy and mobility. Deep analytics, especially in the field of machine learning, is indispensable to extract value from the potentially massive amounts of data coming from intelligent infrastructures and their always-on end users. Whilst predictive analytics techniques have reached a maturity and scalability, which energy and mobility players could and should adapt immediately – the innovation trigger of the coming years will be real-time prescriptive analytics, which allows for online decision support in dynamic flow systems, as energy and transportation networks have become. Ultimately, the application of big data technologies within these sectors will lead to full automation. Clearly, the culture, existing work force and organizations, as well as the regulatory environment are not equipped to deal with such a transformation. These are not considered in the technology roadmap, but will be input into the EU BYTE project, which aims at analysing the technological paradigm shifts including the socio-economic and legal perspectives to refine the research roadmap as well as define and communicate a sound policy framework for the European Digital Single Market.

In the manufacturing Sector, there are common requirements across a number of typical application scenarios that are analysed here. The main theme from this analysis for the manufacturing sector is data integration, in particular the integration of sensor data into a
A comprehensive data model that needs to be extended to eventually cover the complete manufacturing chain from raw materials to post-consumer. Also, manufacturing data must be integrated with business software, starting with ERP and business process systems, and also additional tasks such as production plant planning. A second theme is a set of research questions grouped around the hostile working environments of manufacturing. Sensors, data transmission, and user interaction must function robustly under such hostile environments, calling for redundancy and new technologies where applicable.

Another theme, although non-technical, is grouped around data protection and data privacy. Company IPR is potentially encoded in product (or part, or object) memories and must be protected from unauthorised access. Workers’ data privacy issues must be addressed by the appropriate stakeholders.

In a changing market, the optimal strategy will be a dual strategy, combining the creation of market leaders with the establishment of leading markets. Market leaders build on top of developed and developing base technologies, i.e., Big Data technologies and other, cyber-physical systems-related manufacturing technologies. In the manufacturing sector, market leaders are machine and plant engineering and construction companies, manufacturers of automation technology and corresponding integrators and service companies in the ICT sector. The leading markets in Europe and the world are the production facilities, including their network of suppliers and services, i.e., their entire value chain. Both perspectives apply to the European market and should thus be combined in a strategy for adapting the manufacturing sector to the Big Data and Industry 4.0 related changes.

The retail sector is obliged to address the requirements of a new generation of customers who expect information to be available at any time. These customers want to be treated as individuals what makes new kinds of services necessary. Retailers have to collect personal data from the customers to be able to address their needs. This information can be acquired by sensors that are installed inside the store and by extracting information from the internet. In addition, analysing sensor data in the store gives the retailers insights that are necessary to optimize operational decisions.
Table of Contents

Executive Summary ............................................................................................................. 5
1. Scope and Methodology .................................................................................................. 12
   1.1. Objectives and Scope .............................................................................................. 12
   1.2. Roadmap Methodology ............................................................................................ 12
   1.3. References ............................................................................................................... 15
2. Sector Health ............................................................................................................... 16
   2.1. Introduction .............................................................................................................. 16
   2.2. Application scenarios .............................................................................................. 16
   2.3. Health Sector requirements .................................................................................... 18
      2.3.1 Data Enrichment ............................................................................................... 18
      2.3.2 Data Integration and Sharing ............................................................................ 19
      2.3.3 Data Privacy and Security ................................................................................ 21
      2.3.4 Data Quality ...................................................................................................... 22
   2.4. Mapping Requirements to Technology ................................................................... 23
      2.4.1 Technology Roadmap ......................................................................................... 23
      2.4.2 Timeline ............................................................................................................ 28
      2.4.3 Related EU Projects ......................................................................................... 30
   2.5. Conclusion and Recommendations ....................................................................... 31
   2.6. Abbreviations and acronyms .................................................................................. 32
   2.7. References .............................................................................................................. 33
3. Public Sector ................................................................................................................. 36
   3.1. Introduction .............................................................................................................. 36
   3.2. Application scenarios .............................................................................................. 36
      3.2.1 Monitoring and supervision of regulated activities for on-line gambling operators .................................................................................................................. 36
      3.2.2 Operative efficiency in Labour Agency ............................................................... 37
      3.2.3 Public Safety in Smart Cities ............................................................................ 38
      3.2.4 Predictive policing using open data ................................................................... 39
      3.2.5 Requirements description ................................................................................ 41
   3.3. Mapping Requirements to Technology ................................................................... 45
      3.3.1 Technology Roadmap ......................................................................................... 52
      3.3.2 Related EU projects ......................................................................................... 54
   3.4. Conclusion and Recommendations ....................................................................... 55
   3.5. Abbreviations and acronyms .................................................................................. 56
   3.6. References .............................................................................................................. 57
4. Finance and Insurance Sector ................................................................................... 59
   4.1. Introduction .............................................................................................................. 59
   4.2. Application scenarios .............................................................................................. 59
      4.2.1 AS1: Market Manipulation ................................................................................. 59
      4.2.2 AS2: Reputational Risk Management ............................................................... 59
      4.2.3 AS3: Retail Brokerage ....................................................................................... 60
4.2.4 AS4: Business development ...........................................60
4.3. Sector requirements ..........................................................60
  4.3.1 Data acquisition ..........................................................60
  4.3.2 Data quality ...............................................................61
  4.3.3 Data extraction and sentiment classification .........................61
  4.3.4 Data integration/sharing ................................................62
  4.3.5 Decision support (DSS) ..................................................62
  4.3.6 Data privacy and security ..............................................63
4.4. Mapping Requirements to Technology ...................................63
  4.4.1 Technology Roadmap ...................................................67
  4.4.2 Related EU projects ......................................................68
4.5. Conclusion and Recommendations ......................................70
4.6. Abbreviations and acronyms ..............................................71
4.7. References ........................................................................72

5. Sector Telecom, Media and Entertainment ..................................73
  5.1. Introduction ......................................................................73
  5.2. Application scenarios .......................................................73
  5.2.1 Telecom ......................................................................73
  5.2.2 Media and Entertainment .............................................75
  5.3. Sector Requirements .........................................................77
  5.3.1 Quantification of data ....................................................80
  5.3.2 Quantification of business priorities ..................................83
  5.4. Mapping Requirements to Technology ................................85
  5.5. Technology Roadmaps .......................................................89
  5.5.1 ACQUISITION ..................................................................92
  5.5.2 CURATION .....................................................................93
  5.5.3 STORAGE ......................................................................95
  5.5.4 ANALYSIS .....................................................................96
  5.5.5 USAGE ..........................................................................97
  5.5.6 Related EU projects .......................................................99
  5.6. Cultural, policy and wider economic aspects for Telco and Media 100
  5.7. Conclusion and Recommendations .....................................101
  5.8. Abbreviations and acronyms ............................................102
  5.9. References ......................................................................103

6. Energy & Transport Sectors .......................................................105
  6.1. Introduction .....................................................................105
  6.2. Application scenarios .......................................................106
  6.2.1 Operational Efficiency ...................................................107
  6.2.2 Customer Experience ....................................................108
  6.2.3 New Business Models ...................................................108
  6.3. Sector Requirements .........................................................109
  6.4. Mapping Requirements to Technology ................................113
  6.4.1 Data Acquisition Requirements ......................................114
  6.4.2 Data Management Requirements .................................114
6.4.3 Data Analytics Requirements .................................................................................115
6.4.4 Data Usage Requirements ...................................................................................116
6.5  Technology Roadmap ..............................................................................................116
  6.5.1 Preliminary Timeline for Technology Roadmap ..................................................123
  6.5.2 Related EU projects ..........................................................................................124
6.6  Conclusion and Recommendations ........................................................................124
6.7  Abbreviations and acronyms ..................................................................................126
6.8  References ..............................................................................................................127

7. Manufacturing Sector ..............................................................................................129
  7.1  Introduction ............................................................................................................129
  7.2  Application scenarios ............................................................................................129
    7.2.1 AS1 Optimisation of production, service, and support ....................................129
    7.2.2 AS2 Optimisation of distribution and logistics .................................................130
    7.2.3 AS3 Production plant planning ........................................................................130
    7.2.4 AS4 Predictive maintenance ...........................................................................130
    7.2.5 AS5 Service-based integrated environment ....................................................131
    7.2.6 AS6 (Big Data) user interfaces ........................................................................131
  7.3  Requirements .........................................................................................................132
  7.4  Mapping Requirements to Technology .................................................................136
    7.4.1 Technology Roadmap ......................................................................................138
    7.4.2 Related EU projects .......................................................................................140
  7.5  Market situation ......................................................................................................142
  7.6  Cultural, policy and wider economic aspects for Manufacturing .........................143
  7.7  Conclusion and Recommendations ....................................................................144
  7.8  References ............................................................................................................147

8. Sector Retail ..............................................................................................................148
  8.1  Introduction ............................................................................................................148
  8.2  Application Scenarios ...........................................................................................148
    8.2.1 AS 1: “In-Store Precision Retailing” ................................................................149
    8.2.2 AS 2: “Operational Decision Management in Retail” .....................................150
  8.3  Sector Requirements .............................................................................................151
  8.4  Mapping Requirements to Technology ................................................................156
  8.5  Conclusion and Recommendations ....................................................................157
  8.6  Abbreviations and Acronyms ..............................................................................158
  8.7  References ............................................................................................................159
Index of Figures

Figure 1: Scope of the Roadmap for Big Data Technologies in Healthcare Sector ................. 23
Figure 2: Roadmap for Public Sector (1/4) ........................................................................... 52
Figure 3: Roadmap for Public Sector (2/4) ........................................................................... 53
Figure 4: Roadmap for Public Sector (3/4) ........................................................................... 53
Figure 5: Roadmap for Public Sector (4/4) ........................................................................... 54
Figure 6: Finance & Insurance. Data Acquisition Roadmap .................................................... 67
Figure 7: Finance & Insurance. Data Quality Roadmap ........................................................... 67
Figure 8: Finance & Insurance. Data Extraction Roadmap ....................................................... 67
Figure 9: Finance & Insurance. Data Integration/Sharing Roadmap ......................................... 68
Figure 10: Finance & Insurance. Decision Support Roadmap .................................................... 68
Figure 11: Finance & Insurance. Data Privacy & Security Roadmap ......................................... 68
Figure 12: Level of data complexity vs. business value ........................................................... 83
Figure 13: Big Data Telco & Media business priorities survey aggregated results 2014 ............ 84
Figure 14: Roadmap for Telecom, Media and Entertainment Sector (1/3) ............................... 89
Figure 15: Roadmap for Telecom, Media and Entertainment Sector (2/3) ............................... 90
Figure 16: Roadmap for Telecom, Media and Entertainment Sector (3/3) ............................... 91
Figure 17: Roadmap for Telecom, Media and Entertainment Sector in a temporal framework ... 92
Figure 18: Excerpt from Gartner’s study on business problems to solve with big data analytics (1= top prio). Heat map also indicates that the highest priority business applications powered by big data analytics can be categorized into Operational Efficiency, Customer Experience, New Business Models ......................................................................................................................... 106
Figure 19: Roadmap for Energy & Transportation ................................................................. 117
Figure 20: Roadmap for the Manufacturing Sector ................................................................. 139
Figure 21: Timeline for the technology roadmap ................................................................. 140
Figure 22: Mapping Application Scenarios to Technical Requirements .................................. 155
Figure 23: Roadmap for Sector Retail .................................................................................. 156

Index of Tables

Table 1. Timeframe of the major expected outcomes (described in Section 2.4.1) .................. 29
Table 2. Ranking of Public Sector requirements among application scenarios ....................... 40
Table 3. Recurrence of Public sector requirements in other sectors ......................................... 41
Table 4. Telecom and media requirements and use cases mapping ........................................ 77
Table 5. Technologies and research questions for Telecom, Media and Entertainment ........... 88
Table 6: Non-technical aspectst for telecom and media ......................................................... 100
Table 7. Timeframe for the major expected outcomes of technology adaptation and innovation ......................................................................................................................... 124
Table 8. Manufacturing requirements and mapping to application scenarios .......................... 135
Table 9. Technologies and research questions for the manufacturing sector ............................ 138
Table 10: Technology needs per industry ............................................................................. 143
Table 11. Required Data Sources in AS1 ............................................................................. 150
Table 12. Required Data Sources in AS2 ............................................................................. 151
1. Scope and Methodology

1.1. Objectives and Scope

“What and when” to maximize and sustain the impact of big data technologies and applications in the various industrial sectors, such as Health, Public, Telco, Media & Entertainment, Finance & Insurances or Retail, Energy & Transport, is the leading question of the Sector’s Roadmap analysis. In order to answer this question, a qualification and prioritization exercise of technologies within business needs has been done.

Each sector has followed the same methodology described in the next section, to produce specific roadmaps. Some cross sector consolidation and homogenization of requirements job was done in order that all sectors using the same descriptions, and also avoiding the overlapping of similar requirements.

1.2. Roadmap Methodology

To create the sector specific roadmaps, we followed a four phase approach, and having as starting point the roadmap process as laid out in section 3.3.2 Sector Roadmaps of BIG deliverable D2.1 Organisation Structure and Operational Procedures (Becker T. et al., 2014).

These phases are:

1) **Investigation phase** to identify sector requirements and big data technologies
2) **Consolidation phase** to establish a common understanding of requirements as well as technology descriptions across domains
3) **Mapping phase** to identify any technologies needed to address the identified cross-sector requirements
4) **Temporal alignment** to highlight which technologies need to be available at what point in time by incorporating the estimated adoption rate by the involved stakeholder.

**Phase one: Investigation**

In the investigation phase, the sector working groups and respectively the technical working groups were investigating parallel in order to identify 1) sector needs and requirements and respectively 2) the state-of-the-art of big data technologies as well as associated research questions:

The sector needs and requirements were identified in three steps as follows:

The aim of the **first step** was to identify both stakeholders and use case applications of Big Data application within the various sectors. Therefore we conducted a literature review including scientific reviews, market studies and other internet sources. This knowledge allowed us to identify and select potential interviews partners and guided us in developing sector specific or technical domain specific questionnaires for the domain expert interviews.

In **second step**, semi-structured interviews were conducted along the developed questionnaires with representatives of all stakeholder groups of the sectors focusing on a) user needs that could be addressed by means of big data or other IT approaches, b) evaluate a list of precompiled big data applications scenarios (which were found in the first step) as well as to describe other promising big data scenarios the interviewees are aware of and c) review with the interviewees a list of possible constraints that are hindering the successful implementation of big data scenarios in the sector.
In the **third step**, by clustering and ranking the discussed use case scenarios, a set of high-level application scenarios were derived, D2.3.2. Final Version of the Sectorial Requisites (Zillner S. et al., 2014). The data collection and analysis strategy was inspired by the triangulation approach (Denzin, 2006). By reviewing and quantitatively assessing the high-level application scenarios, we derived a reliable analysis of user needs and by examining likely constraints of big data applications the relevant requirements that need to be addressed were identified. In order to cross-check and validate the results, dedicated workshops and webinars with industrial stakeholders were conducted.

The state-of-the-art of big data technologies was investigated as follows: The objective of the BIG Technical Whitepapers is to "Determine the latest Big Data technologies, and specify its level of maturity, clarity, understandability and suitability for implementation." To allow for an extensive investigation and detailed mapping of developments, the technical working groups deploy a combination of both a top-down and bottom-up approach, with a focus on the latter.

The approach of the working groups is based on a 4-step approach:

1. Literature research
2. Subject matter expert interviews
3. Stakeholder workshops
4. Technical survey

The literature survey was complemented with a series of interview with subject-matter experts for relevant topics areas. Interviews followed a template which was adapted to the specific context and interview partner. Together with the sectorial forums the technical working groups have held stakeholder workshops to engage the community within the project. The first workshop was held at EDF 2013 to announce the project to the community and gather participants. The second workshop took place at EDF 2014 to present the results of the project for feedback and further validation.

Alignment among the TWGs and between the TWGs and the SFs was important and facilitated through early exchange of drafts, individual meetings and the collection of consolidated requirements through the SFs.

**Phase two: Consolidation**

In order to align the sector-specific labelling of, a consolidation of requirement descriptions was accomplished. For doing so, each sector provided his or her requirements with the associated addressed user needs within a dedicated excel document. The work was initiated at the fourth BIG plenary meeting that took place in Darmstadt during the 21st and 22nd May 2014. In dedicated meetings, similar and related requirements were merged, aligned or restructured. Thus, the initial list of 13 high-level requirements and 28 sub-level requirements could be reduced to 8 high-level requirements and 25 sub-level requirements.

**Phase three: Mapping**

For mapping the technology to requirements, two steps were needed. In a first step, the technical working groups indicated which technology (the technology groups have investigated) can be used to address the consolidated requirements. Besides providing a mapping between requirements and technologies, the technical working groups also indicated the associated research challenge.

Within a one-day workshop, held in Munich on the 25th of June 2014, the initial mapping of technologies and requirements could be consolidated in two steps. First, the indicated technological capabilities were analysed in further detail by describing how the sector-specific aspects of each cross-sector requirement can be handled. Second, for each cross-sector requirement it was investigated whether the technologies from various technical working groups need to be combined in order to address full scope of the requirement.
At the end of the discussion, any technologies that were requested by two or more sectors were included into the cross-sector roadmap definition.

**Phase four: Temporal Alignment**

After having identified the needed technologies, their temporal alignment needed to be accomplished. For doing so, we needed to answer two questions:

- How long is the development time of the technology?
- When will the technology be aligned with the business requirements?

The development time indicates for each technology how many month or years are needed to solve the associated research challenges. This time frame basically relies on the technical complexity of the challenge as well as on the current R&D status. For determining the adoption rate of the big data technology, or respectively the associated big data use case, the non-technical requirements (such availability of business cases, suitable inventive structures, legal frameworks, etc.) as well as the relative benefits as well as total cost for all involved stakeholder (Adner, 2012) needed to be considered.
1.3. References


Becker T. et al. (2014). *D2.1 Organisation Structure and Operational Procedures. Confidential Deliverable of the EU-Project BIG (318062; ICT-2011.4.4)*.

Curry E. et al. (2014). *D2.2.2. Final version of Technical white paper. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4)*.


Zillner S. et al. (2014). *D2.3.2. Final Version of the Sectorial Requisites. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4)*.
2. Sector Health

2.1. Introduction

Several developments in the healthcare sector, such as escalating healthcare cost, increased need for healthcare coverage and shifts in provider reimbursement trends, trigger the demand for big data technologies in order to improve the overall efficiency and quality of care delivery. For instance, the McKinsey Study (McKinsey & Company, 2011) indicates a high financial impact of big data applications in the healthcare domain, on the order of a $300 billion value per year only for the US healthcare. Similarly impressive numbers are provided by IBM: within the Executive Report of IBM Global Business Services (Korster and Seider, 2010), the authors describe the healthcare system as highly inefficient, i.e. approximately US$ 2.5 trillion are wasted annually and efficiency can be improved by 35% which is in comparison to other industries the largest opportunity for efficiency improvements. Moreover, major players are investing in the growth market of medicine for an aging population, for instance Google founded the new company Calico to tackle age-related health problems. In sum, big data applications in healthcare indicate high future potential and opportunities.

However, to the best of our knowledge, only a limited number of already implemented big data based application scenarios can be found already today. Although non-advanced healthcare analytics applications - such as analytics for improved accounting, quality control or clinical research - are available in a wide-spread manner, those applications do not yet make use of the potential of big data technologies. This is mainly due to the fact that health data cannot be easily accessed. High investment and efforts are needed to enable efficient health data management and seamless health data access as the foundation for big data applications. As a consequence, convincing business cases are difficult to identify as the burden of initially required investments strongly reduces any profit expectations. In other words, one of the biggest challenges in the health care domain for the realisation of big data applications is the fact that high investments, standards and frameworks as well as new supporting technologies are needed in order to make health data available for subsequent big data analytics applications. Thus, the efficient management and integration of health data is a key requirement for big data applications in the healthcare domain that needs to be addressed.

Within our sector investigations (Zillner et al., 2014a), (Zillner et al., 2014b), we found out that the highest impact of big data applications in the healthcare domain is expected when it becomes possible to not only rely on one single but various data sources such that different aspects from the various domains can be related. Therefore, the availability and integration of all related health data sources, such as clinical data, claims, cost and administrative data, pharmaceutical and R&D data, patient behaviour and sentiment data as well as the health data on the web, is of high relevance.

2.2. Application scenarios

Our analysis of the health sector showed (Zillner et al., 2014b) that several big data based application scenarios exist that aim towards aligning the need of improved quality -- which in general implies increased cost of care -- with the need of improved efficiency of care. Common to all identified big data applications is the fact that they all require means to semantically describe and align various heterogeneous health data sources, means to ensure high data
quality, means that address health data privacy and security, as well as means for data analytics on integrated health data sets.

For example, **Public Health Analytics applications** demonstrate the potential opportunities as well as associated technical requirements that are associated with big data technologies. Public health applications rely on the management of comprehensive and longitudinal health data from chronic (e.g. diabetes, congestive heart failure) or severe (e.g. cancer) diseases from the specific patient population in order to aggregate and analyse treatment and outcome data. Gained insights are very valuable as they help to reduce complications, slow diseases’ progression as well as improve treatment outcome. For instance, Sweden is since 1970 continuously investing in public health analytics initiatives leading to 90 registries that cover today 90% of all Swedish patient data with selected characteristics (some cover even longitudinal data) (Soderland, 2012). A related study (Health Consumer Powerhouse, 2009) showed that Sweden has the best health-care outcomes in Europe by average healthcare costs (9% of GPD). In order to achieve this, the health data (which is stored in structured (e.g. lab reports) as well as unstructured (e.g. medical reports, medical images) need to be semantically enriched (**Data Enrichment**) in order to make the implicit semantics of the health data understandable across the involved organisations and stakeholders. In addition, a common infrastructure with common standards allowing for seamless data sharing (**Data Sharing**) as well as for the physical integration of multiple data sources into one platform (**Data Integration**) are needed. In order to be compliant to high data security and privacy requirements that are needed to protect the sensitive nature of longitudinal health data, common legal frameworks as well as technical means for data anonymisation need to be in place (**Data Security and Privacy**). Moreover, in order to ensure the comparability of health data sets, processes ensuring high data quality through the standardised documentation as well as systematic analysis of health and outcome data of the specific patient population are required (**Data Quality**).

In terms of data handling, the other identified application scenarios yield very similar technical requirements. For instance, **Comparative Effectiveness Research** applications aim to compare the clinical and financial effectiveness of interventions in order to increase the efficiency and quality of clinical care services. For doing so, large datasets encompassing clinical data (information about patient characteristics), financial data (cost data) and administrative data (treatments and services accomplished) are critically analysed in order to identify the clinically most effective as well as most cost-effective treatments that work best for particular patients. Or **Clinical Operation Intelligence** applications aim to identify waste in clinical processes in order to optimize them accordingly. By analysing medical procedures, performance opportunities, such as improved clinical processes, fine-tuning and adaptation of clinical guidelines, can be realised. Other examples are **Clinical Decision Support (CDS) applications** seeking to enhance the efficiency and quality of care operations by assisting clinicians and healthcare professionals in their decision making process by enabling context-dependent information access, by providing pre-diagnostic information or by validating and correcting of data provided. A further category of scenarios are applications addressing the **Secondary Usage of health data** that rely on the aggregation, analysis, and concise presentation of clinical, financial, administrative, as well as other related health data in order to discover new valuable knowledge, for instance to identify trends, predict outcomes, or influence patient care, drug development, or therapy choices. Or finally, **Patient Engagement applications** that focus on establishing a platform/patient portal that fosters the active patient engagement in the context of patients’ health care processes. Any health apps that run on top of the patient platform rely on the integration of episodic health data from clinical settings as well as non-episodic data captured by devices monitor health-related parameters, such as activity, diet, sleep or weight.

Incorporating the **differences in terms of importance and relevance**, one can state that all big data-based application scenarios described above need to address the **following four technical requirements**: 1) **Data Enrichment**, i.e. the goal is to establish automated means for semantically describing the content of unstructured health data. 2) **Data Sharing and Integration**, i.e. the goal is to store health data in a more or less standard form that can be
shared efficiently as well as move easily and fast from one location to another location. 3) Data Security and Privacy, i.e. the goal is to establish legal procedures and technical means that allow the sharing and communication of data and findings. 4) Data Quality, i.e. the goal is to capture and store health data in high quality such that analytics applications can use the data as reliable input to produce valuable insights.

2.3. Health Sector requirements

For analysing the challenges regarding the large volume, variety and velocity of health data, one needs to distinguish between structured and unstructured data. As of today, large volume of structured is already present. For instance, if one integrates all related data sources of a classical US hospital setting (such as IDNs), one easily exceed one petabyte of data. This data is not only of high volume but also – due to the diverse information sources – of high variety (Wiggins and Otterbach, 2013). However, for the coming years, it is estimated that up to 90% of health data will be provided in various forms of unstructured formats, such as images, video, social web, reports, etc. (Lünendonk, 2013). This high increase of unstructured health data will eclipse the whole scenario in terms of volume and variety (Wiggins and Otterbach, 2013).

Our investigation showed that the mentioned requirements cannot be solved by technological capabilities only, but also rely on the implementation or availability of dedicated processes, standards or frameworks. In order to reflect this, we decided to incorporate those non-technical requirements within our analysis.

2.3.1 Data Enrichment

**Description:** The requirement Data Enrichment addresses the need to make the whole volume and variety of medical data available. Medical data is rich in variety on several levels such as formats, data structure, content, and vocabulary. The market research institute IDC estimates that in the coming year 90% of health data will be provided in unstructured format, such as medical reports, medical images, videos, mp3-files or communications on the social web (Lünendonk, 2013). This implies that the majority of information relies on dedicated pre-processing steps ensuring its “enrichment” by so-called semantic labels before it can be used as input for data analytics applications. For instance, by enhancing medical reports with semantic labels, this information is machine analysable and can be integrated with additional (structured) demographic, diagnosis and treatment information into a holistic view on the patient.

**Underlying Aspects:** For enriching medical data, data enrichment technologies need to tackle requirements on three different levels:

a) **Semantic enrichment of unstructured data:** Accessing the whole volume of medical big data, such as medical text or medical images, requires the additional enrichment of unstructured data with structured, semantic labels that represent the content. This explicit content representation adds data that has been hidden at first and makes the original content semantically accessible and processable. In this context, data enrichment does not only include the extraction of relevant information (IE) from the original data but also making their semantics explicit. E.g. anatomical entities mentioned in the texts are recognised as those
and the describing passages can be linked using sophisticated text analysis techniques (Bretschneider et al., 2013).

b) Medical Data Enrichment Framework: While the preceding aspects addressed the enrichment on conceptual level, a standardised enrichment framework supporting technical integration of the underlying technologies is needed. Besides enabling the standardised integration of software provided for data enrichment, such a framework also supports clinical IT departments in their data and system integration tasks. Available annotation frameworks such as UIMA\(^1\) can be used as their basis.

**Required functionalities:** Data enrichment technologies are required for all health care application scenarios that rely on the integration of heterogeneous unstructured data sources. For instance, public health analytics applications rely on the explicit description of the content captured within medical reports and medical images. Or, comparative effectiveness research applications need to extract the semantics of discharge and treatment reports in order to establish the basis for comparing treatment sequences in relation to their outcome quality.

### 2.3.2 Data Integration and Sharing

**Description:** The requirement Data Integration and Sharing describes the need to efficiently use, integrate and share data originating from various resources. As of today, the adoption of seamless healthcare information exchange technology in Europe is still lacking behind (Accenture, 2012). In average for instance, in Germany less than 26 %, in UK less than 46% and in France less than 27% of the healthcare provider use healthcare information exchange technology (for more detailed numbers see the above mentioned study). In US, the adaptation of EHR technology has due the Meaningful Use Program of the HITECH act in 2009 (i.e. incentive payments of up to $63.750 for eligible professionals and eligible hospitals who can demonstrate implementation and use of certified EHR technology) reached a tipping point. In March 2014, 86% of healthcare professionals and 94% of hospitals have registered for the program and respectively, 68% of healthcare professionals and 90% of hospital demonstrated EHR and IHE capabilities needed to receive incentive payments (CMS, 2014)

The underlying goal is to store health data in a more or less standard form as well as to establish technology standards to make the data easily accessible. To obtain comprehensive sets of health data, we need to be able to integrate data from various resources and different organisations in a longitudinal manner. For instance, analytics on rare diseases and their treatment highly rely on integrated data from many different hospitals over a period of time.

**Underlying Aspects:** Efficient data integration and sharing depends on standardisation of coding systems and terminologies as well as of data models:

a) Use of standardised coding systems/vocabularies: Standardised coding systems are currently used only for high-level information such as diseases (ICD), lab-values (LOINC), problems (SNOMED CT), procedures (HCPCS, CPT-4), and medications (RxNORM). However, not all of them are internationally used: For instance, there are several national coding systems for procedures. In addition, there is no standard coding system for all clinical data (for example finding descriptions). Even though SNOMED CT is a designated standard for use in U.S. Federal Government systems for the electronic exchange of clinical health

\(^1\) [http://http://uima.apache.org/]
information\(^1\), it is not internationally adopted. Hence, we are facing the situation that most healthcare provider organisations mainly use their own coding systems and data models. We further note that SNOMED CT has several open issues regarding consistency, performance, and unique or normalised representation, see e.g. (Rector et al., 2011), (Cornet, 2009), (Schulz et al., 2011), (Martínez-Costa et al., 2013), (Markwell, 2008). In addition, a high percentage of valuable clinical data, such as findings descriptions, is only available in textual and not coded format.

b) **Use of standardised data models:** While the HL7 Reference Information Model is considered to become the standard data model which should be used in combination with SNOMED CT (Benson, 2009) in today's implemented EHR systems, the majority of technologies rely on their own data model. Further, the various clinical departments often use their own reporting templates without specifying the common, cross-departmental semantics. This leads to huge integration efforts even within one organisation. As of today, those integration challenges are addressed by using either adaptations of standard data warehouse solutions from horizontal IT providers like Oracle (Oracle Healthcare Data Model\(^2\)), Teradata (Teradata Healthcare Logical Data Model\(^3\)), or IBM (IBM Healthcare Provider Data Model\(^4\)) or new solutions like the i2b2 platform\(^5\). However, those integration efforts are still limited to one single provider or, respectively to one integrated delivery networks (IDNs). In order to foster cross-provider data integration, several initiatives can be observed. For instance, OpenEHR\(^6\) envisions to set an open standard for EHR data and the IMI EHR4CR project\(^7\) aims to establish a platform to enable integration of EHR data from different hospitals for clinical studies. Within the imaging domain, the DICOM (Digital Imaging and Communications in Medicine\(^8\)) standard, for specifying image metadata is available to enable data sharing across provider settings.

c) **Longitudinal representation of patient data:** In addition to the case centric data representation of today's clinical information systems, a longitudinal data representation is needed. Only if data is analysed over a long time period, disease progression or the success of treatments can be evaluated correctly.

**Required functionalities:** Data integration and sharing is needed for all application scenarios since big data is obtained only if one can easily combine data different types and from various organisations. All described applications scenarios require that

- data is captured using standardised data models. These coding systems and models need to be mature in the sense that they can be used by different groups or organisations for different purposes in an unambiguous way. On the one hand this requires that the semantics need to be precisely defined and on the other hand the models need to be general enough to capture a large variety of data.

- data is captured using standardised coding systems or reference terminologies. Coding systems need to enforce a unique or normalised representation and prevent misinterpretation.

- context information (meta data of meta data) is available to inform the application about the origin of the data source, such as the provenance or legal status information

---

\(^1\) http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html  
\(^2\) https://www.db.bme.hu/files/Manuals/Oracle/Oracle11gR2/doc.112/e18026/toc.htm  
\(^3\) www.teradata.de/logical-data-models/healthcare/  
\(^5\) https://www.i2b2.org/  
\(^6\) http://www.openehr.org/  
\(^7\) http://www.ehr4cr.eu/  
\(^8\) http://dicom.nema.org/
- a *longitudinal data representation* that enables to describe relations of findings over time is ensured.

### 2.3.3 Data Privacy and Security

**Description:** The requirement *Data Privacy and Security* describes the need to protect highly sensitive health related and personal data from unauthorised access and damage. Especially in terms of big healthcare data applications, an even stronger 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, anonymisation is a well established approach to de-identify personal data. Nevertheless, the anonymised data could be re-identified (El Emam, et al., 2012) when aggregating big data from various different data sources.

**Underlying Aspects:** Each big data application needs to put a strong focus on data privacy and security. To ensure an appropriate degree of data protection, the following aspects have to be taken into account:

a) **Legal framework:** The implementation of big healthcare data applications highly depends on the legal framework that needs to be taken into account. Since there is no internationally valid legal instalment, many different – sometimes even conflicting – national data protection laws need to be considered (Göbel, 2013).

b) **Longitudinal reusability:** Within the healthcare domain and in medical research, it is sometimes necessary to longitudinally assess a patient's health status. For instance in prospective medical studies, where the same patient is followed during some period of time, de-identified personal data needs to be re-identified because of important clinical findings.

c) **System heterogeneity:** When aggregating various different data sources in healthcare, one has to deal with many heterogeneous systems and data privacy and security enhancing methods. The proper alignment of a specific patient in the information system of one healthcare provider to the same patient in another's healthcare provider information system needs to be guaranteed (Cloud Security Alliance, 2012).

d) **Scalability of established approaches:** Certain data security and privacy approaches may be bypassed by big data, i.e. when aggregating various different data sources (Cloud Security Alliance, 2012).

**Required Functionalities:** Even though data privacy and security has no direct impact on the quality of big healthcare data applications, it is an important requirement for the majority of the described application scenarios. The strong protection of individual data allows to establish a strong backbone for comparing clinical treatments, outcomes and processes. In all application scenarios,

- the need to protect individual, personalised health-related data from unauthorised access, modification, manipulation, or deletion requires to be assured. I.e. cryptographically enforced access control and secure communication, data provenance, scalable and composable privacy-perservering data managing and analytics are required.
the need to protect health-related data from destructive forces or acts of nature is essential. Here security best practice for non-relational data stores requires to be taken into account.

As the health-related data might be exchanged on a regional, national, and (in the future) international level, data privacy and security approaches need to follow commonly agreed standards and guidelines.

2.3.4 Data Quality

Description: The requirement Data Quality describes the need to capture and store health data in high quality such that analytics applications can use the data as reliable input to produce valuable insights. In contrast to many well-known big data applications in other sectors that allow to generate valuable insights by mainly looking for patterns in data, big data applications in the healthcare domain need to fulfil high data quality standards in order to derive reliable insights for health-related decision. For instance, the features and parameter list used for describing the patient health status need to be standardised in order to enable the reliable comparison of patient (population) data sets.

Underlying Aspects: Each big data application relies on the quality of the data sets used. In the healthcare domain the data quality depends on the efficient handling of four aspects:

e) Data quality of the original data sources: The data quality in big health data application depends on the quality of data of the original data sources. Usually medical product providers are not responsible for the quality (e.g. completeness, accuracy) of the data collected and documented in hospitals. Nevertheless, they provide tools for data collection, documentation and analysis. These tools can help to support hospitals and healthcare providers to ensure that data is complete. Additionally, the tools can help to improve data quality (e.g. plausibility checks, mandatory items).

f) Coverage and level of detail of the collected data items: i.e. to which extent a comparable set of features and parameters across patients from different healthcare delivery settings is captured (i.e. standardised processes of capturing data).

g) Common semantics: the approach for describing the meaning of data items, i.e. to which extent the involved parties and data sources rely on the same or aligned standards for describing the semantics of collected data items (semantic labelling of health data).

h) Handling of media disruption: A major issue regarding data quality, which directly links it to data digitalisation, are media disruptions. The more media disruptions exist in a process chain, the worse the data quality gets.

Therefore, before analyzing health data, two assessment steps have to be carried out: a) the quality of the data sets used for Big Data applications has to be evaluated and b) quality has to be improved by using appropriate approaches (e.g. multiple data imputation).
**Required Functionalities:** High data quality is needed for the majority of the described application scenarios. For instance, the three application scenarios *Comparative Effectiveness Research*, *Public Health Analytics* and *Clinical Operation Intelligence* rely on the semantic integration and alignment of very different health data sources in order to establish the sound basis for comparing clinical processes and treatments. In all three application scenarios,

- the health data used for big data analytics might originate from different data sources and/or different (type of) organisations and is available to a large extent only in unstructured format,

- the data analytics envisioned relies on the same coverage of data items, e.g. the systematic collection of data sources ensuring the same coverage of variables and attributes,

- the comparison of data items requires their semantic labelling as well as the availability of context information in an agreed manner.

As the input data sets encompass clinical data, administrative, financial data, and outcome data, the data capturing process as well as the semantic labelling of data units in the various areas need to follow commonly agreed standards and guidelines.

### 2.4. Mapping Requirements to Technology

#### 2.4.1 Technology Roadmap

By analyzing the underlying research question of the required technologies, we could derive a big data technology roadmap for the healthcare sector (see Figure 1).

<table>
<thead>
<tr>
<th>Technical Requirement</th>
<th>Technology</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Enrichment</strong></td>
<td>Medical IE Algorithm</td>
<td>Identification of Relevant Information Entities</td>
</tr>
<tr>
<td></td>
<td>Medical Image Understanding</td>
<td>Automated detection of abnormal structures</td>
</tr>
<tr>
<td></td>
<td>Medical Annotation Framework</td>
<td>Standards fostering of IE algorithm integration</td>
</tr>
<tr>
<td><strong>Data Sharing and Integration</strong></td>
<td>Semantic Data Representation</td>
<td>Creation of mature data models</td>
</tr>
<tr>
<td></td>
<td>Semantic Knowledge Models</td>
<td>Improvement of existing biomedical ontologies</td>
</tr>
<tr>
<td></td>
<td>Context Representation</td>
<td>Provenance, data usage, licence</td>
</tr>
<tr>
<td><strong>Data Privacy and Security</strong></td>
<td>Hash algorithms</td>
<td>Hash algorithms</td>
</tr>
<tr>
<td></td>
<td>Secure Data Exchange</td>
<td>IHE profiles</td>
</tr>
<tr>
<td></td>
<td>De-identification Algorithms</td>
<td>Anonymization, Pseudonymization, k-Anonymity</td>
</tr>
<tr>
<td><strong>Data Quality</strong></td>
<td>Provenance Management</td>
<td>Trust &amp; permission management mechanism</td>
</tr>
<tr>
<td></td>
<td>Human-Data Interaction</td>
<td>Natural language UI &amp; schema agnostic queries</td>
</tr>
<tr>
<td></td>
<td>Unstructured Data Integration</td>
<td>Unstructured Data Integration</td>
</tr>
</tbody>
</table>

*Figure 1: Scope of the Roadmap for Big Data Technologies in Healthcare Sector*
Data Enrichment

In Scope / Out-of scope: In order to enrich the data in health care environments, different aspects are of importance: 1) Unification of structured information content and information extraction from unstructured content for information access and availability has been adapted in existing information architectures, such as GATE or cTAKES (Savova et al., 2010). How these first approaches and incorporated algorithms can be adapted to the specific needs of the health care domain has to be analysed. 2) With the availability of big data from larger numbers of health institutions, the algorithms and frameworks must satisfy the arising scalability requirements. In particular the performances of routines for information extraction from unstructured, non-standardised reports have to be addressed.

Technologies and Research Questions: In order to improve data enrichment for health care data, several technological advances are needed:

1. Medical Information Extraction Algorithms that are tailored for medical texts and establish the basis for incorporating available, external domain knowledge are needed. As of today, existing approaches focus mainly on pattern-based extraction. In order to recognize relevant information entities, which exist beyond the well-known ones (such as diseases, organs, etc.), further research is necessary. First outcomes should focus on Knowledge-Based Information Extraction Algorithms that make use of medical knowledge from existing ontologies (such as disease information, demographic information about the patient, research outcome from related clinical trials, and other LOD resources) within the information extraction process is a very young research area. By relying on Gartner's evaluation of the Text Analytics technologies (Gartner, 2013), we estimate a timeframe of up to 2-5 years for the development of semantic IE algorithms in the medical domain. As we aim to benefit from as well as reuse available research results, we expect the first baseline algorithm to operate in English language and on clinical reports only. Further outcomes need to foster advances of IE approaches to a larger variety of text types as well as multiple languages. As of today, research focuses mainly on clinical documents with well-known and described vocabulary, syntactical and semantic properties. However, the advent of Social Media technologies fosters physician-patient or inter-patient communication that differs from the known structures (Chretien and Kind, 2013). The IE research outcome for English language (which is currently the baseline language for IE research) cannot be transferred without additional effort other European languages. Numerous publications show that the language support of algorithms either required adaptation or redesign: French (Abacha, 2013), German (Krieger et al., 2014), Dutch (Spyny et al., 1998), Swedish (Kokkinakis, 2010)). Thus, we estimate that the consideration of the multiple different text types and other languages in the information extraction algorithms requires detailed review and further research efforts.

2. Image understanding algorithm: Imaging processing algorithms for Automated Detection of anatomical structures and Abnormal Structures (including automated measuring), such as (Seifert et al., 2009), (Feulner et al., 2009), or (Seifert et al., 2010). Segmentation algorithms have gained intense research in the last decade. Wang and Summers (2012) give a detailed overview of the field and show that Machine Learning has been well established for this task. Algorithms for abnormality detection will be based on this research and enables (partial) automated and more effective diagnosis processes. Derived from the experience in
the THESEUS MEDICO\(^1\) project and the implementation of an automated recognition algorithms for anatomical entities in CT images (Seifert et al., 2010), development time of algorithms for abnormality detection is estimated to three to four years.

3. **Standardised Medical Annotation Framework**: A standardised medical text processing and understanding framework supports technical integration of annotation technologies; this incorporates the definition of data formats (output and exchange formats) and information delivered from semantic annotation systems.

*First outcomes are supposed to cover a Standardized Formats and Interfaces for Annotation Modules*: UIMA has been established as de-facto standard framework for annotation of unstructured content in the medical domain (used in systems such as IBM's Watson (Giles and Wilcox, 2011), cTAKES (Savova et al., 2010), GATE and many others (Roeder, 2010), (Kang, 2010)). However, the implementing systems are focused, stand-alone systems without integration into the heterogeneous information management architecture of health care providers. For unified integration and exchange of the unstructured content modules and their annotation results, formats and interfaces for communication between software modules, such as annotators, have to be defined. By building on the preceding prototypical implementation of information extraction algorithms, the *implementation of a mature standardised medical annotation framework* will become possible. As the development of a standardized annotation framework relies on the involvement of various stakeholders in medical information integration, we estimate the this research question to be addressed in five year time.

### Data Integration and Sharing

**In Scope / Out-of scope**: Within clinical settings data integration and sharing strongly depends on (1) semantically precise terminologies and (2) data models, (3) common exchange formats and (4) their adoption as a standard by different groups and organisations. The creation of semantically sound data models and vocabularies is clearly in scope of the technology roadmap. However the adoption of standards and increasing the incentives to share data are not part of the roadmap since these requirements are non-technical.

**Technologies and Research Questions**: In order to improve data integration and sharing available data models and coding systems have to be improved and standardised.

1. *Semantic data models* are key enablers of data integration and sharing. Only semantically sound data models enforce *unambiguous representation* of data and thus allow efficient data integration and reuse. Existing models like the HL7 Reference Information Model version 3 however lack coherence and sound definitions of the classes which make it difficult to use in implementations.

One important outcome are *common semantic data models for structured patient data that aligns* existing terminologies suitable for annotating relevant data units. In a first step, such semantic data models will be capable to represent the comprehensive patient data, such as disease information, clinical findings, procedures, lab values, etc., that is available in structured format, and - in a second step – also capable to represent the content of unstructured data sources, such as medical images or radiology reports.

\(^1\) [http://www.joint-research.org/das-theseus-forschungsprogramm/medico/](http://www.joint-research.org/das-theseus-forschungsprogramm/medico/)
2. **Semantic Knowledge Models**: In the biomedical domain ontologies and terminologies are used to represent the knowledge for different domains. These ontologies and terminologies are used in combination with data models to express clinical data such as findings and observations. Existing ontologies such as SNOMED CT need to be improved to prevent ambiguous or even false representations. These are structural adaptations to make the ontology consistent and also consolidation of the terms such that unique encoding is enforced.

*In order to align semantic knowledge models and terminologies*, existing biomedical ontologies and terminologies will provide an important backbone for data integration. However these terminologies themselves need to be aligned and standardized to support interoperability. The necessary efforts can be seen by a relatively 'simple' alignment/harmonization of ICD-11 with SNOMED as shown in (Rodrigues 2013) and (Rodrigues, 2014).

3. **Context information**: Additionally to common exchange formats, context related data, such as information about data usage, ownership of data, or data provenance, have to be provided and managed in a standardised manner to allow more efficient sharing of data at large scale, where information from various resources is analysed. This requires standards for describing context information, i.e. meta-data about meta-data.

Important outcomes are means for *Context representation for data repositories*. In order to achieve interoperability of the increasing number of open data repositories, such as linkedct.org, openFDA¹, bio2rdf.org etc., which often originate from different healthcare and life science disciplines, the corresponding repositories need to provide context information of their data in a reusable manner. As of today, first initiatives, for instance the W3C Semantic Web Health Care and Life Sciences Interest Group (HCLS IG), are targeting the described challenge. However, future work is needed to align and standardize the different evolving approaches.

In addition, means for the *Context representation for all patient data* are preceding important results: By extending the common semantic data models (see above) suitable for clinical data representation with the corresponding context data, such as provenance and licensing information on the level of single data elements, the transparent sharing and exchange across organisations becomes possible.

As mentioned in the research questions, the time needed for implementing means enabling successful data integration and sharing highly depends on the availability of agreed standards. This fact is reflected by the federal regulation US (CFR, 2014) in the context of the meaningful use initiative in the US.

**Data Privacy and Security**

**In Scope / Out-of scope:**

Within the healthcare domain, different aspects of health data privacy and security need to be taken into account: 1) Existing data privacy enhancing methods need to be assessed for whether they satisfy all requirements within the scope of big data. 2) Existing privacy enhancing methods need to be adapted and enhanced to meet the requirements demanded for big data applications or new approaches need to be developed. 3) A common legal framework and international guidelines need to be established and followed (Göbel, 2013).

---

¹ https://open.fda.gov/
The assessment and enhancement of existing privacy and security approaches as well as the development of new approaches are aspects that are considered by the technology roadmap development. However, case (3) is out of scope of the technology roadmap.

Technologies and Research Questions: In order to improve data privacy and security, several technology advances are needed:

1. **Hash algorithms** are mainly used as an encryption technique to facilitate integrity checking, cryptography, indexing, or digital signatures. A hash algorithm is a one-way function that substitutes a string with arbitrary length by a hash value with a specific length. Such one-way hash functions make it computationally impossible to reconstruct the input string out of the resulting hash value. They could also be used to generate pseudo-identifiers within the scope of longitudinal studies. For hash algorithms, it is essential to be robust and collision resistant.

   Ever since storing private data in the cloud evolved in the recent years, **privacy enhancing through hash algorithms** become an interesting option for preserving data privacy. There are already techniques available to preserve privacy of personal data in the cloud, but it has to be evaluated, whether these approaches can also handle Big Data or if adaption or innovations are needed. Efficient privacy-preserving Big Data Processing (Karvelas et al., 2014) is enabled by several techniques that are still in research and need to be evaluated, after successful evaluation disseminated and finally adopted.

2. **Secure data exchange** that enables health-related data to be shared across institutional or country boundaries, is essential, for instance for electronic health records. The IHE\(^1\) (Integrating the Healthcare Enterprise) has made efforts in secure plug-and-play access. **IHE profiles** (e.g. IHE Cross-enterprise document sharing) are widely used. Nevertheless, they are still the focus of research activities.

   To integrate healthcare and to enable secure data exchange, approaches like the **IHE profiles for secure data exchange** are necessary. Therefore IHE offers the Cross Enterprise Document Sharing (XDS). As of today, this approach has “few implementations due to the difficulty in getting from today’s reality to the desired future state that is defined by the XDS integration framework” (Kaschinske, 2014). Nevertheless IHE is working to solve the big data problem in the healthcare domain and provides potential solutions that are tested in so-called IHE Connecathons\(^2\).

3. **De-identification algorithms** are key enablers of big data applications within the healthcare domain (Neubauer and Kolb, 2009). This can be achieved by **anonymisation** or **pseudonymisation** (FP7,1995)(FP7,2008). In contrast to anonymisation, which is the complete removal of all identifiers, pseudonymisation allows to link data associated with pseudo-identifiers independent of time and place of data collection. **K-anonymity** (El Emam and Dankar, 2008), which ensures the anonymity of data when aggregating various different data sources, is a promising research field.

   One important outcome in this context are **anonymization, pseudonymization and k-anonymity approaches for big data**. Anonymization and pseudonymization of data were handled as very good approaches to preserve the privacy of a person’s identity in the healthcare domain (e.g. for clinical studies). However, when talking about big data, these approaches are not sufficient anymore (Zang and Bolot, 2011), since the integration of several different big data source may allow unintentional re-identification of data. Therefore new anonymization or pseudonymization techniques are required, that eliminate the possibility of unintended re-identification. As of today k-anonymity was thought to be a possible solution to this problem. Nevertheless, recent studies (Shokri et al., 2010) have

---

1. http://www.ihe.net/
shown that k-anonymity fails to provide strong privacy protection. Therefore enhancements are needed.

Data Quality

In Scope / Out-of scope: Within clinical settings, there are two aspects of data quality of patient data that are of relevance: 1) Data quality is usually measured in terms of frequency of missing and incorrect values in the data. 2) Data quality and information density is also strongly correlated to the degree of automation of the process and quality of the tools and IT infrastructure. That is, low level of automation, low-quality tools, and poor architecture will inevitably lead to loss of data over time and will not permit efficient data utilisation. Fixing the data produced in such environments is extremely difficult.

By using samples of data and knowledge of the ‘truth’, it is possible to estimate data quality in case (1); however case (2) requires a systematic approach to improving the underlying processes and, therefore, is mainly an implementation challenge (which is not focus of the roadmap development).

Technologies and Research Questions: In order to improve health data quality, several technology advances are needed:

1. **Provenance management** is a key enabler of trust for health data curation. Providing curators the context to the particular data sets that are considered as trustworthy allows them to capture their data curation decisions. For improving the provenance management in the healthcare domain, *data-level trust and permission management mechanisms* that are fundamental to supporting data management infrastructures for data curation need to be developed.

2. **Human-data interaction technologies** that foster improved health data quality (e.g. data coverage) in particular of relevance in healthcare settings to enable ease-of-use user interaction (e.g. data assessment processes) that fits to the particular workflows of healthcare professionals. In particular, natural language interfaces or schema-agnostic query formulation are a promising research direction to support healthcare professionals in data quality assessment processes.

3. Reliable approaches for the high percentage of unstructured data sources in healthcare settings are needed to ensure high data quality for, in particular for medical images or medical reports. Future research needs to extend and adapt existing NLP pipelines, entity and relation algorithms (Savova et al., 2010), and image segmentation algorithm (Seifert et al., 2010) to the address the specific characteristics of health text and image data (see also Section Data Enrichment).

**2.4.2 Timeline**

In order to find out which technologies are needed at what point in time as well as how technologies interrelate with each other, a systematic approach for predicting technology developments is needed. The developed technology roadmap establishes such a framework by
aligning user needs and associated requirements with technological advances and the related research questions.

In contrast to technology roadmap developments accomplished in the context of a single company, our approach covers the development of a technology roadmap for the European market. As a consequence, it was not possible to come up with a precise timeline of technology milestones, as the speed of technology development and its adoption on the market relies a) on the degree to which the identified non-technical requirements will be addressed and b) on the extent to which European organizations are willing to invest in big data developments and use case implementations.

In order to implement the scope of the technical goals laid out in the section before, Table 1 depicts the estimated timeline it may take until the different research challenges are solved by indicating expected outcomes (described in the section before) that build on top of each other.

<table>
<thead>
<tr>
<th>Technical Requirement</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Enrichment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Standardized Formats and interfaces for Annotation modules</td>
<td>Knowledge-Based Information Extraction Algorithm</td>
<td>Algorithm for Anomaly Detection in Images</td>
<td>Definition and Implementation of Medical Annotation Framework</td>
<td></td>
</tr>
<tr>
<td>Data Integration</td>
<td>Context representation for data repositories</td>
<td>Aligned semantic knowledge models and terminologies</td>
<td>Common semantic data model for structured patient data</td>
<td>Common semantic data model for unstructured patient data</td>
<td>Context representation for all patient data</td>
</tr>
<tr>
<td>Data Security &amp; Privacy</td>
<td>IHE profiles for secure data exchange</td>
<td>Privacy enhancing through hash algorithms</td>
<td>Anonymisation, pseudonymisation and k-anonymity approaches for Big Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Quality</td>
<td>Methods for Trust and Permission management</td>
<td>Natural Language UI &amp; Schema agnostic Queries</td>
<td>Integrated workflows for Trust and Permission Management</td>
<td>Context-aware integration of unstructured data</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Timeframe of the major expected outcomes (described in Section 2.4.1)
2.4.3 Related EU Projects

Related projects founded by the European Union are conducted in the context of four frameworks:

- The most related projects have been conducted within the 7th framework under the ICT theme. Out of the 2262 projects, we identified seven to be of relevance, because they relate directly or indirectly to the requirements identified in the BIG Health Care project. E.g., within the Linked2Safety\(^1\) project, Dr. Athos Antoniades shows that techniques such as aggregated data in form of data cubes can be used to extract data from anonymised EHRs for usage within clinical trials with consideration of patients’ safety (related to Data Privacy and Security requirement).

- A larger number of projects conducted within the 7th framework Health theme (22 out of 1001) can provide valuable insight on how data-driven tools are developed today for Data Enrichment and Integration: Most projects aim to integrate the so far distributed data and expertise on OMICS data (projects such as GEUVADIS\(^2\) or PROTEOMEXCHANGE\(^3\)). Still, their results have to be considered on a more abstract level to be input for the development of the technical roadmap.

- Within the 2nd Health Program projects, potentially 36 projects out of 308 assigned to the Health Information category can be attributed to the Public Health Analytics application scenario. Most projects (such as the BIOEF\(^4\)) aim to create a common network and infrastructure for a group of experts or information on a specific disease. The number of projects funded show the need for integrated information in the health care domain. Most information resides in the institutions and cannot be accessed for integrated clinical trials. That is why currently each of the single projects needs to investigate how to share, integrate, and publish data (Data Sharing and Integration).

- Projects conducted in the Competitiveness and Innovation Framework Programme (CIP) take a more general view on research on data-driven research for society. The EPSOS project\(^5\) takes a patient-centric view and stresses the Data Security issue in Data Integration. Next to the data security requirements, the restriction of data access on the ‘need to know’ principle is stressed by technologies of authorisation and authentication.

Today, most related projects deal with the defined requirements only 1) in parts and 2) always in the context of a specific disease or (public) health related research question, or even have to deal with the question of 3) how to setup a privacy-considering, common infrastructure for data sharing. The technical requirements arising with the availability of big data are often only considered as a side question. Therefore, questions of Data Enrichment, Data Quality and Data Integration are of minor research focus. However, most project frameworks identified the questions and will address them in their forthcoming calls (such as Horizon2020, CIP or the

\(^1\) http://www.linked2safety-project.eu
\(^2\) http://www.geuvadis.org
\(^3\) http://www.proteomexchange.org/
\(^4\) http://www.bioef.org/
\(^5\) http://www.epsos.eu/
3rd Health Program). Therefore, an observation of the projects to be started and their outcomes will enrich the development of the technology roadmap with hands-on questions from clinical research.

2.5. Conclusion and Recommendations

The success of big data application relies on the availability and seamless access of health data. This is very challenging due to several reasons: a) health data is not only very heterogeneous but a large percentage is only available in unstructured format b) health data is stored in distributed locations across providers and stakeholders c) health data is only of value if high data quality standards are met and d) due to its sensitive character, data privacy and security requirements for health data are rather high.

Within our investigation, we identified the various technologies and associated research questions that need to be available to address the mentioned challenges. In addition, we highlighted the critical role of agreed and adopted standards that need to be in place.
### 2.6. Abbreviations and acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cTAKES</td>
<td>Clinical Text Analysis and Knowledge Extraction System</td>
</tr>
<tr>
<td>DICOM</td>
<td>Digital Imaging and Communication in Medicine</td>
</tr>
<tr>
<td>EDF</td>
<td>Engineering Development Forum</td>
</tr>
<tr>
<td>EHR</td>
<td>Electronic Health Record</td>
</tr>
<tr>
<td>GATE</td>
<td>General Architecture for Text Engineering</td>
</tr>
<tr>
<td>HL7</td>
<td>Health Level 7</td>
</tr>
<tr>
<td>ICD</td>
<td>International Classification of Diseases</td>
</tr>
<tr>
<td>IHE</td>
<td>Integrating the Healthcare Enterprise</td>
</tr>
<tr>
<td>IE</td>
<td>Information Extraction</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>TWG</td>
<td>Technical Working Group</td>
</tr>
<tr>
<td>UIMA</td>
<td>Unstructured Information Management Architecture</td>
</tr>
<tr>
<td>UMLS</td>
<td>Unified Medical Language System</td>
</tr>
</tbody>
</table>
2.7. References


Chretien K., Kind T. (2013): *Social Media and Clinical Care*. In Journal of Curculation


El Emam et al. (2014): *De-identification methods for open health data: the case of the Heritage Health Prize claims dataset*. In Journal of Med Internet Res. 2012 Feb 27;14(1)


3. Public Sector

3.1. Introduction

The Public sector has a great potential to benefit from big data technologies, but for accomplishing this it needs to overcome some barriers like resistance to change, data silos and limitations imposed by legislations for the reuse of information for other purposes that the original for which it was collected.

In this roadmap for public sector, four application scenarios are reviewed to look for the requirements and research questions for the development of big data. Some of the innovative applications may come from the security area which is mainly responsibility of the sector. However, great improvements can be also gained in the operative efficiency of standard administrative processes, and in the effective application of government policies.

3.2. Application scenarios

Four representative application scenarios (AS) where described in the Public Sector (PS) chapter of the requirements document (Zillner S. et al., 2014). These scenarios represent some of the most significant improvements big data can provide to this sector. Other application scenarios that share part of the requirements are enumerated along with each scenario, to highlight its representativeness for the sector.

For each of these scenarios we describe their applicability and representativeness in the sector, as well as their most important requirements from the big data point of view.

3.2.1 Monitoring and supervision of regulated activities for on-line gambling operators

In the EU, revenues from on-line gambling were €10.54 billion in 2012, with an estimated 6.84 million consumers,

On-line gambling regulation across countries in Europe is very different in terms of which type of games can be licensed, whether other operators licensed in other EU member states can offer or advertise their online services (Olswang). Recently the European Commission has published a recommendation on consumers protection in online gambling in terms of minors protection, players accounts, registration and their activity (Davis, 2014), but currently only a few countries have successfully regulated internet gaming (France, Spain and Italy being the most prominent) that require some type of monitoring of the online games (Newell, 2014).

The risks faced by online game market, according to the Spanish regulator (DGOJ - Dirección General de Ordenación del Juego), and that the regulation tries to avoid through the monitoring of on-line game are:

- Playing by non-authorised people (minors, self-excluded, people related to operators)
- Operators not offering clean and honest game
- Not assuring privacy and personal data protection
The goal of this application scenario is to monitor the on-line gambling operators for the control of regulated activities and detection of fraud. The amount of data received in real-time, on daily, and monthly basis cannot be processed with standard database tools. Nowadays the data is received and stored, but no active analysis is performed, only on-demand. The user of this application is the public body in charge of the supervisory activity. This procedure is a regulatory obligation from the public administration; the on-line gambling operators must provide the information to the regulatory public through a specific communications channel. Real-time data is received from gambling operators every five minutes (Zillner S. et al., 2014).

This AS can benefit from new disruptive applications thanks to the power of big data analytics. Controlling and monitoring the regulated activity of on-line gambling operators through the analysis of real time data from games, and daily data from users, games, and jackpots, to protect users and prevent and prosecute fraud and other criminal operations related to the on-line gambling. This is accomplished through deep data analytics for pattern discovery, and even the use of real time insights that enable the analysis of fresh / real-time data. In this scenario data security and privacy are extremely important, moreover when the gaming information is analysed with additional data coming from other domains.

This on-line gambling scenario is similar to others in the public sector where the control and monitoring of regulated activities is required by law by the competent public sector bodies, specifically in areas like energy, telecommunications, and stock markets. Likewise, the same techniques for pattern discovery can be applied for:

a) fraud detection from tax evasion and money laundering activities, with information partially stored in public sector files and partially in financial institutions

b) fraud detection from pensions, unemployment benefits and public subsidies to businesses by analysing information already available in public sector files

In these to last examples the on-line analysis of information is not required as in the main example for the on-line gambling, but the analysis pattern discovery process is quite similar, looking for abnormal behaviour patterns.

### 3.2.2 Operative efficiency in Labour Agency

The German Federal Employment Agency is an organization with a network of more than 700 agencies and branch offices nationwide with more than 100,000 employees. Its tasks are job and training placement, career counselling and providing benefits replacing employment income such as unemployment benefit and insolvency payments. The agency has to manage six different main data sources with users’ related information, from Employment History, Participation in Measures History, Jobseeker History and Employment Biographies to name the most important ones to achieve an optimization of the job placement programs.

The goal of this AS is to enable a new range of personalized services for its users, more than six million, improve customer services and cut operations costs in German Federal Labour agency. Providing more effective job placement services, as previously all unemployed workers were receiving the same standard services despite having different profiles. This is accomplished through the integration of all the diverse data sources related to the scenario analysis into a big data platform. The segmentation of the users allowed providing more targeted interventions for the unemployed based on the analysis of the historical data and the
success of the job placement programs to each segment of unemployed. Discovering the
relationship among the different initial conditions of the users and the effectiveness of the
programs is a key factor. This allows tailoring the new placement programs to each user's
needs (Zillner S. et al., 2014).

This scenario is similar to others in the public sector where improvements in efficiency though
the analysis of existing operational data is required like:

a) Improvement of public services through internal analytics, based on performance
indicators. Exploiting information already available from current processes can help to
improve performance and compare it across different geographical units, or even
provide information on vendors and service providers, allowing better procurement
decisions, and therefore, saving money.

b) Personalization of public services to adapt to citizen needs. Segmentation and tailoring
of public services to individuals, and by increasing efficiency and citizen satisfaction

3.2.3 Public Safety in Smart Cities

Smart cities equipped with sensors and elaborated linked data infrastructures help the public
sector keep cities and their citizens safe. Many safety scenarios require quick decision making
based on the current situational awareness picture. Having accurate and up-to-date information
allows better and faster responses during emergencies and results in less damage and
casualties. Typical sources for obtaining such information can come from emergency response
calls, surveillance cameras and mobile forces (such as a police patrol car) that arrived at a site.
In recent years social media have shown interesting potential for gathering information that aids
obtaining an accurate situational awareness picture (van Kasteren, Ulrich, Srinivasan, &

This AS provides improvements in terms of effectiveness, as all the information related to public
safety is brought to a single place. Improving security in cities by exploiting all available data
provided by the smart cities infrastructure and by emergency services, and then providing quick
decision making based on situational awareness pictures. Additional environmental information
can be extracted through the analysis of social media networks. All the information collected is
combined to provide a hypothesis of the situation. The requirements for these types of
applications comprise:

- use of real time data transmission techniques for sensor data collection,
- data integration of multiple data sources into the big data analytical platform,
- and natural language analytics to extract information from sources like social networks,
speech records from emergency services and CCTV cameras
- real-time insight to enable real-time data processing to search for patterns and
relationships
- modelling and simulation techniques to anticipate the results from decisions taken to
influence the current conditions
- data security and privacy, as most data related to incidents can be personal
Other ASs in the security domain share part of the requirements for this scenario, as those described in the Public Sector chapter of the first version of the roadmap (Zillner S. et al., 2013), like:

a) Management of emergencies and disasters. This is the generalization of the Public Safety in Smart Cities application scenario where the smart city sensors are replaced by other generic sensors like radars, satellite and environment sensors, but the analysis requirements are nearly the same that in the Public Safety in Smart Cities scenario.

b) Crime detection and prevention. This scenario requires the analysis and recognition of suspicious behaviour patterns from social media networks and advanced image recognition and matching from surveillance cameras and other video sources, as well as integration with other related existing sources like police and criminal records.

c) Security threat detection. For the prediction and prevention of security threats in the domain of homeland security, including cyber security and critical infrastructures. This scenario requires the analysis of some specific information, like reports from network vulnerabilities analysis and machine data traffic, but also from other common sources to the security scenarios, like media content, social media networks, and other generic web content like blogs, news and RSS feeds).

3.2.4 Predictive policing using open data

Open data initiatives make datasets freely available to the public, so data can be used without restrictions from copyright or patents. Governments around the world have started open data initiatives to make public sector data available to the public for sake of transparency and to allow third parties to offer services based on the data. One such service can be described as predictive policing where historical crime data is used to automatically discover trends and patterns in the data. Such patterns help in gaining insights into crime related problems a city is facing and allow a more effective and efficient deployment of mobile forces (Wang, Rudin, Wagner, & Sevieri, 2013) (Zillner S. et al., 2014).

This AS provides improvements in terms of effectiveness, as making use of the information stored in historical records; it allows a better use of the public security resources. This application scenario extracts information from available historical crime open data sets to automatically provide predictive information on crimes based on the discovery of trends and patterns in data, and therefore allowing the efficient deployment of security forces. This application requires the use of pattern discovery analytics, to discover similar behaviours in criminal activities, and the use of predictive analytics that through machine learning can automatically discover trends in such criminal activities.

This scenario is a specific application of the availability of Open Government Data as described in the Public Sector chapter of the first version of the roadmap (Zillner S. et al., 2013). Predictive analytics can be also used in other scenarios for the planning of public services (like education, social services for elderly, public transport, etc.) or to perform analysis and forecasts on fundamental areas of economic activity (e.g. financial, food and raw materials markets), based on existing series of statistical and/or open data, as in the scenario described in the PS chapter of the first version of the roadmap (Zillner S. et al., 2013). Sector requirements.
The following eight requirements from three different domains are derived from the previous Public Sector application scenarios:

Data management engineering domain:
1. Real Time Data Transmission
2. Data Sharing/Data Integration

Deep data analytics domain
3. Natural Language Analytics
4. Real Time Insights
5. Predictive Analytics
6. Pattern discovery
7. Modelling and simulation

Data security and privacy
8. Data Security and Privacy

The ranking of these requirements according to their recurrence among application scenarios is detailed in Table 2. Four of eight requirements – Pattern discovery, Data Sharing/Data Integration, Real Time Insights, Data Security and Privacy – are listed in more than one AS, and therefore are considered basic to the development of big data in the sector. Whereas the other four only appear in one scenario, so they are required to the development of that specific AS.

<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Pattern discovery</th>
<th>Data Sharing/Data Integration</th>
<th>Real Time Insights</th>
<th>Data Security and Privacy</th>
<th>Real Time Data Transmission</th>
<th>Natural Language Analytics</th>
<th>Predictive Analytics</th>
<th>Modelling and simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Monitoring and supervision of on-line gambling operators</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Operative efficiency in Labour Agency</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Public Safety in Smart Cities</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>4. Predictive policing using open data</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. Ranking of Public Sector requirements among application scenarios
In Table 3 it is shown the recurrence of the public sector requirements in other sectors. This was the result from the Consolidation phase, as described in section 1.2 Roadmap Methodology. Only one of the requirements - Pattern discovery - does not appear in the ASs from other sectors, all the others appear at least once in the application scenarios from the other sectors, so most of them have some cross sectorial relevance.

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Data Sharing / Data Integration</th>
<th>Real Time Insights</th>
<th>Data Security and Privacy</th>
<th>Real Time Data Transmission</th>
<th>Natural Language Analytics</th>
<th>Predictive Analytics</th>
<th>Modelling and simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern discovery</td>
<td>Telco &amp; Media Transport &amp; Energy Health Manufacturing</td>
<td>Telco &amp; Media Retail Manufacturing</td>
<td>Transport &amp; Energy Health Retail Manufacturing</td>
<td>Retail</td>
<td>Retail</td>
<td>Health Manufacturing Retail Manufacturing</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3. Recurrence of Public sector requirements in other sectors**

### 3.2.5 Requirements description

Information to address the technical questions from Public sector application scenarios requirements

**AS 1. Monitoring and supervision of on-line gambling operators**

Requirements: Pattern discovery, Real Time Insights, Data Security and Privacy

- **Variety of data:** In this scenario the types of data to be analysed are basically structured files with information from the users (registry and account status) and games (game description, status, amounts, players). Additionally, the analysis process can look into other structured data sets with bank account operations and tax information.

- **QoS requirements:** For the data on games, which is real-time, there is some level of failure tolerance, but this is acceptable as far as data can be processed later and do not compromise the service level agreements. The admitted level of wrong pattern discoveries will depend on the service level agreement.

- **Acquisition:** Data transferred from on-line game operators is encrypted as it contains sensitive data

- **Characteristics of data provided**
  - **Volume of data:** In Spain there were 276,958 monthly active users in 2013 with a total of 1.5 million of registered users. It is a growing market with a turnover of 5.438 million euros in 2013. There were a total of 51 operators registered in Spain in 2013. In terms of data to be provided from operators to the controlling agency, every month it has to be provided User registry information and Account status per user (one record each). This means 3 million records every month. The operators must provide on daily basis information on operator account, jackpots and living games, and every
5 minutes every operator must send a file with the status and players of current games. This means that every five minutes the sum of all games managed by all operators must be in the order of magnitude of thousands records (DGOJ - Dirección General de Ordenación del Juego).

- Data format: XML
- Data sources are legally approved operators who must fulfil the technical specifications for the data required. (Ministerio de Economía y Hacienda, 2011)
- Data contains information on Users registry (daily and monthly); Game account (daily and monthly), Operator account (daily and monthly), Jackpots and living games (daily and monthly), Games (real time).
- Information/knowledge to be extracted from the data: Patterns of abnormal activities
- Data is used by domain specialist at the supervisory agency.
- Data to be stored in-house, because the information is sensitive and cannot use cloud storage.

**AS 2. Operative efficiency in Labour Agency**

Requirements: Pattern discovery, Data Sharing / Data Integration

- Variety of data: Historical data on interventions of job and training placement, and career counselling programs for users as well as its profile information, as work history, skills, education, training. Typically, the analysis process will look into those structured data sets with historical data and electronic records of the agency programs.

- QoS requirements: Predictive policing using open data: The system is failure tolerant for failures, as no real-time data is processed. The admitted level of wrong pattern discoveries will depend on the service level agreement.

- Acquisition: Data is structured in relational databases

- Characteristics of available data
  - The main entity, users of Labour agency are more than six million. The main databases that set up the system are (I) Employment history; (II) Benefit Recipient History; (III) Participation in Measures History File; (IV) Unemployment Benefit II Recipient History; (V) Jobseeker History; (VI) Integrated Employment Biographies (Research Data Centre (FDZ)).

  - The information is available in structured formats from databases.

  - Data source are the applications managed by the own Labour Agency, so its quality can be considered high.

  - Data contains historical information on users in relation with their employment (job positions) and unemployment (periods of unemployment, and interventions done during these periods).

  - Information/knowledge to be extracted from the data: Results of the different placement programs handled by the employment agency, spotting those ineffective ones.

  - Data is used by domain specialist at the labour agency.

  - Data is stored in-house, due to privacy requirements. Information can be anonymised before it can be stored and processed outside the agency (if this option is considered).
AS3. Public Safety in Smart Cities

Requirements: Data Sharing / Data Integration, Real Time Insights, Data Security and Privacy, Real Time Data Transmission, Natural Language Analytics, Modelling and simulation.

- Variety of data: Nature of data to be handled is of different types: Sensor data (speed radars, meteorological, environmental quality, surveillance cameras, and license plate recognition cameras); social media networks, information from emergency services (streams of text and speech).

- QoS requirements: The robustness and correctness of the system are very important, as data needs to be acquired and analysed in real-time to provide a situational awareness picture to decision makers.

- Characteristics of sensor data:
  - Thousands to ten thousands of cameras generating petabytes of data depending on the density of cameras and the size of the city.
  - Cameras provide unstructured video and images. Audio from sensor for gunshot detection. Other sensors like environmental and meteorological sensors provide structured data.
  - Data sources depend on the different organizations responsible for the sensors networks. They can provide raw sensor data to semi-processed data.
  - The data collected from sensors has a temporal component (timestamp).
  - Sensor data is real-time or near real-time.
  - The information to be extracted from the sensors networks is useful environmental information that can help in the decision making process according to the current situation.
  - The information is to be used by feeding the appropriate situational awareness systems used by decision makers.

- Characteristics of social media networks information:
  - Estimated volumes are 500 million tweets a day (Internet Live Stats, 2014) and 30 thousand million pieces of content shared on Facebook each month (Branckaute, 2010).
  - Unstructured data coming from social media networks (text and pictures). And social media provide unstructured text, video and images.
  - Data provided by social media networks users (not 100% reliable on individual basis).
  - Data may have temporal and spatial components, (time stamp and location).
  - Velocity, the volume of social media messages related to a local emergency may be very low compared to the total number of messages managed by social media networks, but they will be produced in a relative short period of time, this means a high volume of messages to be managed in a short period of time.
  - Information to be extracted from social media data is all the information that can be related to emergency events, already known by the emergency management or not.
  - The information extracted is to be used as an alternate detection network of emergency events by the decision makers managing the situational awareness system.

- Characteristics of emergency services information:
  - Comparatively small sizes of operational data and existing databases.
Unstructured data in text and speech formats, and semi structured data like geolocation of emergency equipment.

- Data is provided by the emergency services in agreement with the control centre.
- Some of the data may have associated temporal components (text, speech), and some of the data may have additional spatial components (emergency equipment locations).
- The information from the emergency services will be transferred in real time, so there will not be a continuous data stream as it will have data peaks coinciding with emergencies.
- The information to be extracted is basically the description and location of emergencies.
- The information extracted is to be used as additional information on emergencies events by the decision makers managing the situational awareness system.

- Visualization: situational awareness system with location of emergencies and equipment on maps, with additional information like social media networks topics and sensor information located on the map as well (including images from surveillance cameras).

### AS 4. Predictive policing using open data

**Requirements:** Pattern discovery, Predictive Analytics

- **Variety of data:** Structured open data with historical crime data.
- **QoS requirements:** Predictive policing using open data: The system is failure tolerant for failures, as no real-time data is processed. The admitted level of wrong pattern discoveries will depend on the service level agreement.
- **Acquisition:** It depends on the city which provides the data. In the case of Chicago crime incident data is available in different structured and semi structured formats as CSV, JSON, PDF, RDF, RSS, XLS, XLSX or it can be programmatically accessed using OData or the SODA API. In the case of London it is provided in CSV.

- **Characteristics of data provided**
  - **Volume of data:** Depending on the area covered and the level of detail of the information, the volume of data will be quite different. As an example, for the city of London crime data is 135 Mbytes for the whole 2011 year in CSV format.
  - The information is available as structured and semi structured formats (RDF, XML and CSV/spreadsheet formats as the most usual)
  - The most reliable information can be provided by police and/or security offices. The most detailed can be provided at local level, while information provided at regional or national level will be grouped at some level and therefore unusable for the purpose of predictive policing.
  - The information used will be basically of historical type, so there are no requirements for data acquisition speed (Velocity)
  - The knowledge to be extracted from the data is crime patterns in terms of spatial and temporal dimensions to provide an optimization of resources in those dimensions as well.
  - Data is to be used by domain specialists at the law enforcement agencies.

- **Visualization:** Geospatial information (2D and 3D maps) providing concentration of crimes by time period, with the possibility to filter by crime type.
3.3. Mapping Requirements to Technology

For each requirement in the sector, technical working groups have identified and mapped the applicable technologies and the research questions to be developed, in order to cope with the sector requirements. All references presented here are from report D2.2.2 Final version of Technical white paper, (Curry E. et al., 2014).

Pattern discovery: Identifying patterns and similarities to detect specific criminal or illegal behaviours in the application scenario of monitoring and supervision of on-line gambling operators (and also for similar monitoring scenarios from public sector). Also in the scenario to improve operative efficiency in the Labour Agency, discovering patterns of success and failure in the application of employment programs to specific unemployed profiles. In the predictive policing scenario, pattern discovery is useful to detect similarities in criminal acts that can be predicted with the analysis of the series of such criminal patterns.

Data Analysis
- Technology: Semantic pattern technologies including stream pattern matching. As described in D2.2.2, sections 3.5.1.1 Large-scale Reasoning and 3.3.2.1 RDF data stream pattern matching.
- Research Question: Scalable complex pattern matching.
  - Pattern matching is something that semantics can help with tremendously. Already SPARQL can be used to find very basic patterns. This means we can do pattern matching over data sets in the billions. Reaching trillions will take 5 years.

Data Curation
- Technology: Validation of pattern analytics outputs with humans via curation, described in D2.2.2, section 4.5.6 Current Approaches for Data Curation.
- Research Question: Curation at scale depends on the interplay between automated curation platforms and collaborative approaches leveraging large pools of data curators, see D2.2.2, section 4.3.2 Emerging Trends: Scaling-up Data Curation.
  - The application of machine learning approaches to the discovery of data curation patterns is still at early research stage. The area is critical for scaling and automating data curation activities and has a very high impact in the data analysis pipeline. More fundamental research needs to be developed alongside basic software infrastructure platforms. Given the focused investment on basic research and software infrastructure development, commercial application results could be reached in six to ten years.

Data Storage
- Technology: Analytical Databases, Hadoop, Spark, Mahout. See D2.2.2, section 5.4.2 Data Storage Technologies.
- Research Question: Standard Array Query Language
  - From a storage point of view, the technologies are analytical databases that address the problem of better performance as the DB itself would be able to execute analytical code. Currently there is a lack of standardized query languages but efforts such as ArrayQL are on their way. Currently there is no wide spread adoption yet
and existing DBs (SciDB, Rasdaman) are more used in the scientific community. This may change in 3-5 years from now.

**Data Sharing / Data Integration:** Required to overcome lack of standardization of data schemas and fragmentation of data ownership. Integration of multiple and diverse data sources into a big data platform. As in the cases where it is required to perform data analysis from data belonging to different domains and owners (e.g. different agencies in the public sector) or integrating heterogeneous external data (from social networks, sensors)

**Data Acquisition**
- Technology: To facilitate the integration as well as analysis. See D2.2.2, section 2.6.3 Government, Public, Non-profit.
- Research Question: Data Fragment Selection? Sampling? Scalability? See D2.2.2, section 2.5.1 Future Requirements/Challenges.
  - Existing distributed processing approaches can be used to address this problem. One would simply have to write a small piece of code within say a Storm bolt that enables it to simply reject some data (sampling) or only use certain data (fragment selection). The scalability is ensured by the distribution of bolts across more hardware. Fragment selection can be difficult in current architectures when the data to be processed has a graph-like topology. Here, optimal fragment selection algorithm can be NP-hard. Implementing even approximations of such algorithms on current architectures is a difficult endeavour. Solutions will be brought about by quantum computers (to be around in 5-10 years according to Morello and Dzurak from the University of New South Wales. 15-20 years seems more realistic.)

**Data Analysis**
- Technology: Linked Data provides the best technology set for sharing data on the Web. Linked Data and ontologies provide mechanisms for integrating data (map to same ontology; map between ontologies/schemas/instances). See D2.2.2 section, 3.5.3.2 Data abstraction based on ontologies and communication workflow patterns, details how the use of ontologies can help the data integration.
- Research Question: Scalability, dealing with high speed of data and high variety.
  - Technology is already available and semantic models can be developed using existing languages. However to come up with a common model for integration requires understanding of all proprietary models and thus be able to cover with the suggested model all the individual cases. Efforts should be devoted first to achieve an understanding of interest of the involved parties and then work towards the definition of the common framework. Dealing with billions of nodes is feasible now. Dealing with Trillions will take 3-5 years.
- Research Question: Making semantic systems easy to use by non-semantic (logic) experts.
  - These systems were not designed to be used by non-computer scientists so for general users specific interfaces will always be created. IT professionals can learn the basics associated with Linked Data in a short amount of time. Improvements could be made to the whole tooling infrastructure and for integrating Linked Data tools with other Big Data tools. There now exists very good open educational resources on Linked Data – see [http://euclid-project.eu/](http://euclid-project.eu/). It will take 5 years at least to have a comprehensive tooling support.
- Research Question: Robust scalable mechanisms for provenance.
The type of data will determine the accessibility and required analysis mechanism (simple data formats vs images) although the same technologies can be used to provide with an integration framework. Technologies are already here and efforts should be dedicated in the particularities of each type of data to be integrated and thus to be considered in the model.

- Research Question: Scalable automatic data and schema mapping mechanisms. These can build upon background knowledge supplied by the existing Web of Data as well as any private semantic resources.
  - As with the scalability question of dealing with high speed and high variety of data, technologies are already available. Challenges stated more at the "understanding" level among stakeholders.

Data Curation / Storage
- Technology: metadata and data provenance frameworks (HadoopProv [Sherif, Sohan und Hopper 2013]), [Ikeda, Park und Widom 2011]. See D2.2.2, section 5.4.3.5 Challenges in Data Provenance.
- Research Question: What are standards for common data tracing formats? Provenance on certain storage types, e.g., graph databases, is still computationally expensive. See D2.2.2, sections 4.7.6 Standardization & Interoperability and 5.5.1.2 Security and Privacy.
  - The ground data model and conceptual model standards for representing provenance data is in place (e.g. W3C PROV-O, COGS Vocabulary). Most of the challenge concentrates on the integration of the provenance representation standards and provenance capture mechanisms into existing data curation, ETL and data management tools. The integration of provenance-awareness into existing tools can be achieved in the short-term (2-3 years) once this becomes a critical market demand.

Real Time Insights: Enable analysis of fresh /real-time data for instant decision making, for obtaining real time information from the data. As it is, in the case of Public Safety in Smart Cities to build situational awareness systems form real time data provided from networks of sensors and near real-time data captured from social networks. This is applicable alto to monitoring and supervision of regulated activities with real-time data, as in the on-line gambling operators’ application scenario, or for the supervision of any real time activity which is regulated by the public sector, like energy, telecommunications, stock markets, etc.

Data Analysis
- Technology: Linked Data and machine learning technologies can support analysis. See D2.2.2, section 3.5.2 Stream data processing.
- Research Question: High performance while coping with the 3 Vs.
  - A lot of work is required to enable real time insights at large scale with semantic/Linked Data technologies. Current approaches rely on materialisation to create all triples which can be inferred. This bloats the data size but increases performance. Real time deep analytics is 5+ years away.

Data Storage
- Technology: Drill, Impala, In-memory databases. See D2.2.2, section 5.4.2.3 Big Data Query Platforms.
• Research Question: How can ad-hoc queries on large data sets be executed with minimal latencies? See D2.2.2, section 5.7. Conclusions.
  o Limitations are in supporting unknown ad-hoc queries with seconds or sub-second latencies for which the system has not been specifically optimized. Current approaches include schema-on read technologies such as Impala, Drill or Spark SQL. Continuous performance improvements can be expected on this front. Other approaches in the field of approximate query processing (AQP) trade off accuracy for response time. This is an active research field and may reach further maturity in a few years’ time.

Data Usage
• Technology: Complex Event Processing applies Business Rules (or other frameworks) continuously on defined (short) interval of real time data stream with low latency. See D2.2.2, section 6.5.2 Decision Support.
• Research Question: In-memory technology, new visualization and interaction techniques, automatic system reactions.
  o Big data sets must be transformed to high-level information before meaningful visualisation. High-level information can also relate to trends analysis and clustering analysis and abstraction and dimensionality reduction (four trends) is mainly dependent on the quality of such analysis algorithms more than corresponding visualisations. Visualisation is also supported by relating graphs to existing images, e.g., maps (GIS). With regard to speed: GPU-based visualisation systems can easily display tens of millions of data sets on a map in interactive speed (30 refreshes per second) and systems with multiple GPUs scale to the order of a billion data points. Give map-d.com a try.

Data Security and Privacy: Legal procedures and technical means that allow the secure and privacy respecting sharing of data.
The solutions to this requirement may unlock the massive use of big data in public sector; even it is not directly related with the properties of big data. Advances in the protection and privacy of data are key for public sector, as it may allow the analysis of huge amounts of data owned by the public sector without disclosing sensitive information, as in many cases the public sector regulations restrict the use of data for different purposes for which it was collected. These privacy and security issues are also preventing the use of cloud infrastructures (processing, storage) by many public agencies that deal with sensitive data. A new approach of security in cloud infrastructure may unlock this barrier.

Data Storage
• Technology: Encrypted storage and DBs; proxy re-encryption between domains; automatic privacy-protection (e.g., differential privacy). See D2.2.2, section 5.4.3 Security and Privacy, which discusses security and privacy related to big data.
• Research Question: Advances in "privacy by design" to link analytics needs with protective controls in processing and storage. A legal framework, e.g., the General Data Protection Regulation (GDPR), has to be harmonized among EU member states. Beyond legislation, data and social commons are required. See D2.2.2, section 5.5.1.2 Security and Privacy.
  o Queries on encrypted storage are still and active research field. The main challenge is the extremely high computational load. Various researches have addressed this problem, but applicability to real use cases still needs to be demonstrated. This will require at least further three years of research.
Real Time Data Transmission: Acquire (sensor) information in real time.

The capability of placing sensors is increasing in Public Safety Smart Cities application scenarios, specifically; the image sensors have followed Moore’s Law, doubling megapixel density per dollar every two years (PWC, 2013-2014). It will be required to provide distributed processing and cleaning capabilities for image sensors in order to do not collapse the transmission channels (Jobling, 2013) and provide just the required information to the real-time analysis systems that will be feeding the situational awareness picture systems for decision makers.

Data Acquisition
- Technology: Storm. See D2.2.2, section 2.4.2.1 Storm.
- Research Question: Distributed processing and cleaning
  - Current broadband architectures can pass on gigabits of data per second. The main bottleneck lies in the pre-processing of the data streams before they go into storage. Here, solutions such as Storm work well as they allow easy distributed processing and pre-analysis before storage. The approach scales quasi-linearly with the number of nodes used for processing and thus with the processing speed of hardware. Current approaches should be able to let the user know the type of resources that they will require to perform tasks specified by the user (e.g., process 10GB/s). First approaches towards these ends are emerging and they should be available on the market within the next 5 years.

Data Storage
- Technology: Current best practice: write optimized storage solution (e.g. HDFS), Columnar stores
- Research Question: What are best practices to persistently store a high velocity data? How to improve random read/write performance of database technologies.
  - The Lambda architecture described by Marz and Warren reflects the current best practice standard for persisting high velocity data. Effectively it addresses the shortcoming of insufficient random/read write performances of existing DB technologies. Performance increases will be continuous and incremental and simplify overall development of Big Data technology stacks. Technologies could reach a level of maturity that leads to simplified architectural blueprints in three to four years.

Natural Language Analytics: Extract information from unstructured online sources (e.g. social media) to enable for instance sentiment mining. Recognition of data from natural language inputs like text and sound and objects and behaviours from image and video. The expected growth of Twitter users’ in 2018 is 70% from the users’ volume in 2014 (Beck, 2014), this means, assuming the same growth in tweets, an increase from 500 million tweets a day to 850 million tweets, so more and more natural language information will be available from social networks.

Data Analysis
- Technology: Information extraction, named entity recognition, machine learning, linked data. (AN) Entity linking and co-reference resolution.
- Research Question: Increasing scalability and robustness.
There are research efforts on scalable information extraction. Entity linking and coreference is harder. Robust scalable solutions are 3-5 years away at least.

**Data Curation**
- Technology: Validation of NLA outputs with humans via curation.
- Research Question: Curation at scale depends on the interplay between automated curation platforms and collaborative approaches leveraging large pools of data curators. See D2.2.2, Natural Language Processing pipelines, at section 4.7.8 Unstructured & Structured Data Integration.
  - There is a major demand for software infrastructures which can integrate information extraction to data curation platforms, in which human computation can refine and correct extraction errors and provide feedback which can be used for training information extraction algorithms. Most of the effort consists on the development of software infrastructures to integrate NLP pipelines to data curation pipelines. Technically, this integration can be achieved in the short-term (2-3 years).

**Data Storage**
- Technology: Distributed File systems (HDFS), noSQL Databases (especially key/value and document stores).
- Research Question: How can unstructured data be stored in large scale for later analysis? What are best practices to store and manage unstructured data?
  - There are no real technical limitations for storing unstructured data. Various technologies such as HDFS, Object Storage and key/value stores are able to handle unstructured data at considerable scale. Cost of storage is much more a limiting factor that forces to implement retention policies that eventually discard old data.

**Data Usage**
- Research Question: How can Natural Language Analytics be integrated into Data Usage Scenarios?
  - Sentiment Analysis and Opinion Mining are well-established research areas with existing and (computationally) scalable products on the market. Initial set-up requires manual intervention for customisation. Establishing large-size knowledge bases (see IBM Watson) is the current research and development question. Other applications of NLP and NLU (Understanding) include relation-extraction, co-reference resolution, event recognition, and terminology extraction (from large sets of unstructured documents in not-yet-modelled domains).

**Predictive Analytics**: Provide predictions learning from previous situations, in order to provide optimal resource allocation for public services. As in the application scenario for predictive policing, where the goal is to distribute security forces and resources according to the prediction of incidents, based on temporal patterns or specific events related to any kind of cause (sports events, weather conditions, or any other variable).

**Data Analysis**
- Technology: Linked Data and machine learning technologies can support analysis. Large scale reasoning. See D2.2.2, section 3.5.1.3 Large-scale Machine Learning.
• Research Question: High performance while coping with the 3 Vs. Combining large scale reasoning with statistical approaches.

Data Storage
• Technology: Analytical Databases.
• Research Question: How can databases efficiently support predictive analytics?
  o From a storage point of view, the technologies are analytical databases that address the problem of better performance as the DB itself would be able to execute analytical code. Currently there is a lack of standardized query languages but efforts such as ArrayQL are on their way. Currently there is no wide spread adoption yet and existing DBs (SciDB, Rasdaman) are more used in the scientific community. This may change in 3-5 years from now.

Data Usage
• Technology: Predictive Maintenance: predict failures, determine maintenance intervals. Support for failure analysis. See D2.2.2, section 6.5.3 Predictive Analysis.
• Research Question: Extend predictive analytics to prescriptive analytics.
  o Predictive analytics (existing technology) is applicable to any problem (in the public sector) with regular needs for action/intervention as, e.g., maintenance. Savings result from deviations from scheduled actions (early action to prevent malfunction, delayed action to save unnecessary action). Prescriptive analytics is a research question and aims at (automatically) deriving procedures. In the public sector, many such procedures are given by regulation/legislation and cannot be changed easily.

Modelling and simulation: Domain specific tools for modelling and simulation of events according to changes from past events to anticipate the results from decisions taken to influence the current conditions in real-time in scenarios of public safety. Reactions can be tested, for example, in the case of managing large crowds of people in public events and also, to anticipate the responses to actuations during the management of emergencies and disasters.

Data Analysis
• Technology: Linked Data/semantics.
• Research Question: Integrating semantics into large scale modelling and simulation environments.
  o The combination of ontological models and rules can be used to trigger alarms according to known knowledge in a flexible manner. Its combination with data mining techniques can be used to anticipate results according to past events.

Data Storage
• Technology: Best practice. Lambda architecture, temporal databases.
• Research Question: how can time-series data be managed in a general way for effective analysis?
  o On one hand there are temporal database that include temporal concepts as first class citizens, but query interface do typically not support the expressiveness of analytical databases. Another challenge is that in the context of IoT devices time-series data is often spatio-temporal data, but there are virtually no spatio-temporal
databases on the market. As a consequence spatio-temporal modelling and optimizations have to be done on application level. But spatio-temporal databases are an active research field and results may beyond a five year time scale.

**Data Usage**
- Technology: Standards in (semantic) modelling; application of simulation in planning (e.g., plant planning).
- Research Question: making models explicit and/or transparent. See D2.2.2, section 6.6.1 Future Requirements.
  - Map-based event and crowd visualisation products exist, but do not make underlying models explicit/transparent. This is a research question with a long timeline (beyond 2020), depending on the complexity of the underlying models. Simple models, e.g. for crowd movement, employ simulations that (i) can be visualised and (ii) their parameters can be visualised too.

### 3.3.1 Technology Roadmap

The following figures illustrate the main requirements and appropriate technologies in Public Sector. The open research challenges are also mapped to the corresponding technologies.

**Technology Roadmap**

Public Sector (1/4)

<table>
<thead>
<tr>
<th>Technical Requirement</th>
<th>Technology</th>
<th>Research Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern discovery</td>
<td>Scalable complex pattern matching for trillion records</td>
<td>High performance coping with the 3 Vs</td>
</tr>
<tr>
<td>- Identifying patterns and similarities</td>
<td>Semantic pattern technologies</td>
<td>Linked Data and machine learning technologies supporting analysis</td>
</tr>
<tr>
<td></td>
<td>Machine learning approaches to the discovery of data curation patterns (6 to 10 years of R&amp;D)</td>
<td>Ad-hoc queries with minimal latencies</td>
</tr>
<tr>
<td></td>
<td>Validation of pattern analytics outputs with humans via curation</td>
<td>'This active research field may reach further maturity in a few years'</td>
</tr>
<tr>
<td></td>
<td>Standard Array Query Language</td>
<td>Analytical Databases</td>
</tr>
<tr>
<td></td>
<td>Analytical Databases</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2: Roadmap for Public Sector (1/4)**
# Technology Roadmap

## Public Sector (2/4)

<table>
<thead>
<tr>
<th>Technical Requirement</th>
<th>Technology</th>
<th>Research Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Sharing / Data Integration</td>
<td>Fragment selection for graph-like data (GIS data) through Quantum computing (timeline beyond 2024)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Facilitate the integration as well as analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scalability, high speed and data variety for trillion records</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Semantic systems ease of use</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linked Data for sharing and ontologies for integrating data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Integration of provenance-awareness into existing tools</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Metadata and data provenance frameworks</td>
<td></td>
</tr>
<tr>
<td>Data Security and Privacy</td>
<td>Privacy by design – Queries on encrypted storage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Encrypted storage and DBs</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: Roadmap for Public Sector (2/4)

## Technology Roadmap

## Public Sector (3/4)

<table>
<thead>
<tr>
<th>Technical Requirement</th>
<th>Technology</th>
<th>Research Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Time Data Transmission</td>
<td>Distributed processing and cleaning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data acquisition: Storm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Improving random read/write performance of DB technologies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Write optimized storage solution</td>
<td></td>
</tr>
<tr>
<td>Predictive Analytics</td>
<td>Efficient support of predictive analytics in DBs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Analytical DBs</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Roadmap for Public Sector (3/4)
### Technology Roadmap

**Public Sector (4/4)**

<table>
<thead>
<tr>
<th>Technical Requirement</th>
<th>Technology</th>
<th>Research Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Language Analytics</td>
<td>Increase scalability and robustness</td>
<td></td>
</tr>
<tr>
<td>Extract information from unstructured online sources</td>
<td>Information extraction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Entity linking and co-reference resolution</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sw infrastructures integrating NLP pipelines into data curation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Validation of NLA outputs with humans via curation</td>
<td></td>
</tr>
<tr>
<td>Modelling and simulation</td>
<td>Management of time-series data for effective analysis</td>
<td></td>
</tr>
<tr>
<td>Modelling and simulation of events according to changes from past events</td>
<td>Temporal databases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Making models explicit and/or transparent (timeline beyond 2020)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Application of simulation in planning</td>
<td></td>
</tr>
</tbody>
</table>

![Figure 5: Roadmap for Public Sector (4/4)](image)

### 3.3.2 Related EU projects

- **BYTE project** (Big data roadmap and cross-disciplinary community for addressing societal externalities) ([http://byte-project.eu/](http://byte-project.eu)).

  The focus of this project is to help European science and industry to capture positive and negative externalities associated with big data to gain a greater share of the big data market. BYTE is using some case studies to collect evidences and questions to highlight opportunities to be investigated in the BYTE roadmap. Case studies scopes are Environmental data, Commercial data, Smart cities, Cultural data, Energy, Health and Transport data.

  The relation with BIG is on the basis that several of the referred use cases on which the research is based are in the scope of public sector or are regulated by it.
3.4. Conclusion and Recommendations

The Public Sector analysis for the roadmap is based on four major application scenarios that are representative of the most important areas where big data can help to improve the sector processes.

- Monitoring and detecting irregular behaviours in regulated activities, as well as in citizens’ and business’ duties, through the use of big data analytics.
- Improving the efficiency of public sector processes. Internally by the optimization of public resources through internal analytics, and externally through the personalization of PS services to adapt them to individual’s needs.
- Provide public safety in smart cities by exploding all the information collected through sensors and integrating multiple data sources to provide a situational awareness picture. This type of applications can be extended to a wide range of security applications, from emergencies management, crime detection and prevention and security threat detection.
- Provide predictive policing using open data, thus allowing a better use of public safety resources. This type of application can be extended to other fields where prediction is useful for the optimization of public assets during the tasks of providing public services to citizens.

The findings, after analysing the requirements and the technologies currently available show that there exist research challenges to develop the technologies to its full potential in order to provide competitive and effective solutions. The main developments are required in the fields of scalability of data analysis, pattern discovery and real time applications. Also required are improvements in provenance for the sharing and integration of data from the public sector.

It is also extremely important to provide integrated security and privacy mechanisms in big data applications, as public sector collects vast amounts of sensitive data, and in many countries legislation limits the use of the data for other purposes for which it was originally obtained, and in any case the respect for the privacy of citizens is a mandatory obligation in the European Union.

Other fields especially interesting for the safety applications in public sector is, the analysis of natural language, which can be useful as a method to gather unstructured feedback from citizens, e.g. from social media networks. The development of effective predictive analytics and modelling and simulation tools for the analysis of historical data are also key challenges to be developed in future research for the aforementioned optimizations of public resources.

In addition to these technology related requirements, in the project report D2.3.2 Final version of Sector’s Requisites (Zillner S. et al., 2014), there were highlighted other types of requirements based on the demands of public sector, more focused on other issues, like the lack of legislative support and political willingness to develop the proper regulation and infrastructures to ease questions like the intellectual property rights and cloud computing. Also the lack of skilled data scientists Big Data skills is a demand. In short, the development of common national or European approaches, which requires a clear leadership to take advantage from the synergies created.
### Abbreviations and acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3Vs</td>
<td>Volume, Velocity, Variety</td>
</tr>
<tr>
<td>AS</td>
<td>Application Scenario</td>
</tr>
<tr>
<td>DB</td>
<td>Database (system)</td>
</tr>
<tr>
<td>HDFS</td>
<td>Hadoop Distributed File System</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>NLA</td>
<td>Natural Language Analytics</td>
</tr>
<tr>
<td>PS</td>
<td>Public Sector</td>
</tr>
</tbody>
</table>
3.6. References


Curry E. et al. (2014): D2.2.2. Final version of Technical white paper. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4).


Zillner S. et al. (2014): D2.3.2. *Final Version of the Sectorial Requisites. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4)*.
4. Finance and Insurance Sector

4.1. Introduction

The Financial Services sector is not often an early adopter of new technologies. In the case of Big Data, there are varying experiences as of today: some organisations are pioneers in the use of Big Data (not only within this sector), while others have not started to consider its use yet. This section presents a roadmap for some of the research challenges in Big Data from the perspective of the Financial Services industry, both to take full advantage of the opportunities brought about by this technology and for the late adopters to get on board and benefit as other organizations are doing already today.

4.2. Application scenarios

Among a wide variety of possibilities, three representative application scenarios (AS) for the Finance Sector (FS) were described in the final version of the sector requirements deliverable D2.3.2 (Zillner, 2014). These scenarios represent examples of what Big Data technologies can provide to this sector, in particular linked to the growing demand for sentiment analysis capabilities; this is, the analysis of electronic text from files, reports, surveys, forms, e-mail and more to identify a user's thinking and priorities. For each of these scenarios it is described they applicability and representativity in the sector, as well as their most important requirements from the big data point of view. A more detailed description of the application scenarios can be found in the final version of the sector requirements deliverable (D.2.3.2).

4.2.1 AS1: Market Manipulation

This application scenario proposes improvements for the identification of false rumours aiming at manipulating and influencing the financial markets in order to take advantage of a situation, such as damaging a company reputation or artificially raising the stock prices.

4.2.2 AS2: Reputational Risk Management

Reputation is the opinion or social evaluation of a group of subjects towards an entity on a certain criterion. Reputational risk is defined by the Basel Committee of the Bank of International Settlements (BIS, 2009) as “the risk arising from negative perception on the part of customers, counterparties\(^1\), shareholders, investors, debt-holders, market analysts, other relevant parties or regulators that can adversely affect a bank’s ability to maintain existing, or establish new, business relationships and continued access to sources of funding (e.g. through

\(^1\) As this term appears along the text, a definition is provided. “Counterpart” is an entity, typically an organisation operating in the financial industry, with which a business relationship is established.
the interbank or securitisation markets). This application scenario focuses on estimating reputational risk through collection and analysis of both unstructured information, belonging to selected external data sources available on the web, and structured data.

4.2.3 AS3: Retail Brokerage

This scenario focuses on supporting private and institutional investors in the difficult process of investment decision-making. Here, collection and analysis of new financial information on markets, companies or financial institutions, is essential to operate and is proven to be a very time consuming and complex activity. The AS lies in the possibility for investors of utilising a sentiment built on contents extracted from textual information belonging to a wide variety of data sources available on the web and using it together with classical quantitative indicators.

4.2.4 AS4: Business development

Tapping into the intelligence that data can potentially provide, Financial Services institutions are experimenting today with using internal datasets that were unused until today to build new business offerings around them. This is already being done by some institutions (BBVA, McGraw Hill), and it is a trend that organizations in many sectors (not only financial services) will probably adopt in the near future.

4.3. Sector requirements

This section presents requirements for the FS sector from the previous application scenarios. As the three scenarios are based on sentiment analysis, they share many of the requirements, which allow them to be grouped into more generic categories.

4.3.1 Data acquisition

For banks and financial services providers, the volume of data they generate, consume, store, and access will increase exponentially year over year.

By 2013, 2.5 quintillion bytes of data were being created daily by business and consumers globally, according to cloud computing group Asigra. And according to Credit Suisse, 23% of the information in the data universe could be useful for analysis if it were properly tagged and analysed and yet only 3% was tagged in 2012. McKinsey & Co. also reports that financial institutions with more than 1000 employees lead all industry sectors in terms of data volume with 5,800 terabytes of data stored on average. The New York Stock Exchange creates 1 terabyte of data per day by itself. (McGrath, 2014)

Moreover, not only is volume increasing, but data complexity is ramping up in parallel, as is the speed at which data flows into the enterprise.
In FS, the applications depend on acquiring and accessing massive amounts of historical heterogeneous information and live feeds of unstructured, semi structured and structured information. A significant amount of data comes from internal structured data, though there is a growing trend towards the external unstructured data (from news, blogs, articles, social networks and websites). Even when there can be a wide variety of data sources to access, the actual ones that are required depend on the design for a specific application.

For example in the considered application scenarios:

- AS1: sentiment data, transaction data (end-user data), market data (e.g. via Interactive Data, Bloomberg, Thomson Reuters...), or instrument reference data
- AS2: builds on external unstructured data, mostly blogs, online newspapers and financial documents available on the web, internal structured data (readily available in a bank), and sentiment assessments, obtained through the analysis of external and unstructured data.
- AS3: applications are evolving to those in which data is generated at very high rates in the form of transient data streams, which can have an impact on the acquisition.
- AS4: the value is in internal data already being collected by FS institutions, so this requirement about data acquisition is not as applicable in AS4.

In addition the storage paradigm needs to provide support to handle persistently the information related to the different unstructured sources: metadata (e.g: XML files), hash, sentiment (indicators, phrases, metadata, corpus data) or ontologies. The knowledge base structure has to be suitable for different purposes querying and have the characteristics required for information integration and normalization.

4.3.2 Data quality

This is probably one of the most important requirements for Financial Services. The more timely, accurate and relevant the data (along with good analytics), the better the assessment of the current financial state is. This requires better processes of identifying and maintaining the data sources of interest, verifying, cleaning, transforming, integrating and deduplicating data, but due to the large amount of available data, there is a need for automation and scalability processes. Language detection methods also need to be refined to improve precision and reliability.

Application scenario 4 (business development) is particularly concerned with this requirement, as the viability of such new business models relies on data quality.

4.3.3 Data extraction and sentiment classification

Though the definition of sentiment is vague, in general, a sentiment on a sentiment object is a positive or negative view, attitude, emotion or appraisal on or from a document author or actor. Sentiment is often expressed in a domain-specific way, and using non-domain-specific vocabulary may lead to misclassifications. The goal is to extract facts and sentiments

---

1 The term "instrument reference data" describes identifier (name, symbol, ISIN), descriptive (industry, trading venue, etc.), and pricing data for financial vehicles (instruments) which is sourced from multiple market data vendors. http://www.definefinance.com/instrument-reference-data
concerning the financial use cases: financial instruments, situations, conditions, indicators, and experts' assessments regarding these instruments, as well as investors' sentiment, etc. The classification of sentiment can be done at several levels: words, phrases, sentences, paragraphs, documents and even multiple documents, and then aggregate.

Data extraction always needs to cope with noise, misinformation, irony, bias or uncertainty. In addition, with sentiment it is important not only determine the sentiment of a piece of information, but how words affect the semantic orientation and how sentiment change.

4.3.4 Data integration/sharing

This describes the task to overcome for disparate data sources the heterogeneity in hardware, software, syntax and/or semantics by providing access tools that enable interoperability.

The data is usually scattered among different heterogeneous sources with differing conceptual representations (different structures and data semantics) but it is encapsulated into a single, homogeneous data source to the end-user.

The motivation for integration may be based on strategic or operational considerations. Regarding strategic considerations and analysis, it may not be required to constantly integrate the data but to integrate data snapshots at a certain point in time. For operational analysis a real-time integration of the most up-to-date information may be required.

Typically data integration is not a once-off conversion but an on-going task, therefore poses the additional constraint that the chosen solution needs to be robust in terms of adaptability, extensibility, and scalability.

This rapid generation of continuous streams of information has challenged the storage, computation and communication capabilities in computing systems, as they impose high resource requirements on data stream processing systems.

4.3.5 Decision support (DSS)

Model-driven DSS emphasise access to and manipulation of statistical, financial, optimisation and/or simulation models. Models use data and parameters provided by decision makers to aid decision makers in analysing a situation, for instance, assessing and evaluating decision alternatives and examining the effect of changes. This requires integrating information from the knowledge base into financial event detection models, visualization models, decision models, and for scalable execution of these models.

For some application scenarios, the response of the system should support real time or near-real-time insights. The velocity of the response is subject to the end-user requirements.

In DSS, visualisation is an extremely useful tool for providing overviews and insights into overwhelming amounts of data to support the decision support. The visualization is tied to the specific application or use case, offering different perspectives of the sentiment data. For example in the considered application scenarios:

- AS1: instrument-based alerts within the market for the specific instruments
AS2: (aggregated) counterparts average sentiment over time (All, per year, 30 days, etc.), counterparts’ sentiment volume over time, reputational topics for counterparts (30 days and 1 day), (aggregated) sentiment charts, or topic charts.

AS3: most integrating information for a stock in combination with related data market. For example: stock data overview, sentiment data overview over time, historical stock prices charts, tag cloud charts around a stock, content map charts, canyon flow charts around a stock over time.

Visualization should provide a user-friendly specification of entities of interest, with visualization of several time series simultaneously, able to handle multiple-request at the time. On top of that, more that the different perspectives, the visualization usability lies in the dashboard associated functionalities, such as search engines, configurable views, drop-down, colour maps, and indicators derived from the aggregated data. In any case, the visualization has to be adapted to the customer using the service.

4.3.6 Data privacy and security

Top priorities for the FS sector today include on-going regulatory compliance (e.g: Sarbanes-Oxley (SOX) (U.S.Government, 2002) Act, EU data protection directive (Parliament, 1995), cyber security directive (Parliament, Proposal for a Directive of the European Parliament and of the Council concerning measures to ensure a high common level of network and information security across the Union, 2013)) and risk mitigation, continued adaptation to the expectations of consumers for anywhere/anytime service, reducing operational costs and increasing efficiencies through use of cloud-based services.

Banking and financial institutions need to secure the storage, transit and use of corporate and personal data across business applications, including online banking and electronic communications of sensitive information and documents.

The increasingly global nature and high-interconnectivity of the industry makes it necessary to comprehensively address international data security and privacy regulations, from the front to the back-end, and along the full supply chain, including third parties. Data is not always stored in-house but at third parties’. As some information used in formulating the responses may come from private data sources, the access to information must be tailored to the different roles or categories of end-users. Using commercial “cloud” services as data storage locations poses potential privacy and security problems since the terms of service for these products are often poorly understood.

4.4. Mapping Requirements to Technology

For each requirement in the sector, this section presents applicable technologies and the research questions to be developed, in order to cope with the sector requirements. All references presented here are from report D2.2.2 Final version of Technical white paper, (Curry E. et al., 2014).

Data Acquisition

- Technology:
Three current-day technologies are identified here (Acquisition pipeline, proprietary APIs, data crawler engines), which help the data acquisition process nowadays.

- Research Question: Data stream management
  - Current data analysis in the stored-data domain shall need to move to management of data in the data stream itself.

- Research Question: Privacy and anonymization at collection time.
  - The data collection process shall require to have an intrinsic aspect of data anonymization and/or decoupling of personal data from data emanating from business processes or otherwise.

- Social APIs:
  - Moving ahead of existing proprietary (or even open) APIs, social APIs into financial services data sets need to be investigated. Some institutions are already doing this, but for the vast majority of the business, this is a research question, on how such social APIs will unleash the value of big data for new business models.

**Data quality**

- Scalable data curation and validation.
  - The central challenge of data curation models in the Big Data era is to deal with the long tail of data and to improve data curation scalability, by reducing the cost of data curation and increasing the number of data curators, allowing data curation task to be addressed under limited time constraint (D2.2.2, section 4.3.2)

- New methods to improve precision and reliability.
  - Precision of data should improve as a function of highly scalable data curation techniques. Reliability will result from data that is known not to have been tampered with. Techniques abound today to ensure this, but for large datasets it will be necessary to have new methods that can scale with the size of datasets.

**Data extraction**

- Language Modelling
  - Possible directions of work in this sense could be about obtaining keywords and key-phrases by using statistical language models. One proposal is to use pointwise KL-divergence between multiple language models to score both phraseness and informativeness, to later be combined into a single score to rank extracted phrases (Tomokiyo and Hurst. 2003. *A language model approach to key-phrase extraction*).

- Machine Learning
  - The size of datasets in financial services makes it necessary for new machine learning techniques to satisfy the newly required inference functionality. This is also highlighted in D2.2.2, section 3.5.1.

- Scalability in real-time
  - Real-time information is of interest in some application scenarios of Financial Services. As in previous requirements, the challenge of processing large datasets...
represents a requirement for research in the scalability of data processing in real-time as datasets grow in size and number.

Data integration/sharing

- Wrappers/mediators to encapsulate distributed data & Automatic data and schema mapping
  o Sources of data in the Financial Services industry can be distributed across organizations, or across time and space. It may be beneficial to encapsulate data from multiple sources in an automatic manner to provide the view of a single source of data, to facilitate its integration in a business process.

- User-specific integration
  o Integration of data for the benefit of specific users (namely, business processes, or target end-user organizations). Shall require agile data integration not only for pre-defined processes, but also for ad-hoc integration of the information.

- Data variety: sentiments, quantitative information
  o Integration of information inferred from open sources with verified (“traditional”) datasets internal to the organization will become the norm in the future. Such integration is being test driven today already, but more research is needed to streamline it.

- Scaling methods for large data volumes and near-real time processing
  o This research challenge is in relation to the “scalability in real-time” described above, under “data extraction”.

Decision support

- Stream-based data mining
  This technique relates to requested 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. (D2.2.2, section 3.5.2)

- Multi-attribute decision models
  o The availability of information from multiple sources will provide multiple attribute types that become available to include in decision models. This is starting to be the case today, but the ubiquity and velocity of large datasets in the near future will make it necessary to research decision models that can adapt to the availability of large numbers of sources with a large volume of data

- Machine learning adaptation to evolving content
  The size of datasets in financial services makes it necessary for new machine learning techniques to satisfy the newly required inference functionality, also in relation to evolution and changes in the content under analysis. This is also highlighted in D2.2.2, section 3.5.1.

- Resource allocation in mining data streams
  Elastic computing today allows for dynamic resource allocation as required. Improvements may be required in resource allocation for near real time support to decision making.

- Improved storage, computation and communication capabilities
Decision making will be based on multiple types of sources with significant amounts of information. The most important aspects in managing such flow of information will be to ensure the necessary communication and computation capacity to take advantage of the available data. Storage shall be relevant when there is a more specific need to maintain part of the information obtained.

Data privacy & security

- Roles-based IdM and access control
  - Access control in the context of large data sets will pose a problem when sensitive data (business process related) beings to be exploited in large data sets and integrated with other data, and accessed by third parties. Research will be necessary to address identity and access management for control of access to sensitive data before and after integration with multiple other sources of data for exploitation in improving Financial Services industry processes, and for creating new business models.

- Privacy by design | Security by design
  - Advances in "privacy by design" to link analytics needs with protective controls in processing and storage. A legal framework, e.g., the General Data Protection Regulation (GDPR), has to be harmonized among EU member states. Beyond legislation, data and social commons are required. See D2.2.2, section 5.5.1.2 Security and Privacy. Queries on encrypted storage are still an active research field. The main challenge is the extremely high computational load. Various researches have addressed this problem, but applicability to real use cases still needs to be demonstrated. This will require at least further three years of research.

- Data Security for public-private hybrid environments
  - The advent of cloud storage and computation services, however, comes at the expense of data security and user privacy. As a matter of fact, customers of cloud services lose control over their data and have no means to control and verify, for example, how data is processed or stored. Customers must trust the cloud providers to not share the outsourced data or code with third parties, and to ensure data integrity. Users also must trust the cloud providers to deploy the necessary security mechanisms in order to counter threats such as hacker attacks, malware infection, etc.

- Enhanced Compliance management (Data protection, others)
  - Research has already been initiated, but needs to continue in providing methodologies and infrastructures that facilitate the monitoring, enforcement, and audit of quantifiable indicators on the security of a business process, and that provide manageable assurance of the security levels, trust levels and regulatory compliance of highly dynamic service-oriented architecture in centralized, distributed (multi-domain), and outsourcing contexts.

- Service Level Agreements
  - SLAs in place today are manageable given the size of datasets today. It will become necessary to produce machine-auditable SLAs and to have SLAs that adapt better to the complex landscape of data sources and their integration into new services.

- Database encryption
  - The security concept of NoSQL databases generally relies on external enforcing mechanisms. Developers or security teams should review the security architecture and policies of the overall system and apply external encryption and authentication controls to safeguard NoSQL databases (D2.2.2, 5.4.3.1)
### 4.4.1 Technology Roadmap

In this section the previously presented main requirements are mapped with appropriate technologies to realize them. This list does not intend to be exhaustive. Furthermore open research challenges are listed, shown in a rough estimated timeline for addressing them and achieving solutions in at least TRL 5/6.

#### Technical Requirement | Technology | Research Challenge
--- | --- | ---
**Data Acquisition** | Acquisition pipeline | Data stream management
| Proprietary APIs | Privacy and anonymization at the collection time
| Social APIs | Data crawler engines

**Figure 6:** Finance & Insurance. Data Acquisition Roadmap

#### Technical Requirement | Technology | Research Challenge
--- | --- | ---
**Data Quality** | Manual processing and validation | Automatic, scalable data curation and validation
| | New methods to improve precision and reliability

**Figure 7:** Finance & Insurance. Data Quality Roadmap

#### Technical Requirement | Technology | Research Challenge
--- | --- | ---
**Data extraction** | Language modelling | Scalability in real-time
| Machine learning

**Figure 8:** Finance & Insurance. Data Extraction Roadmap
### 4.4.2 Related EU projects

- **Project FERARI** *(Flexible Event pRocessing for big dAta aRchItectures, FP7).*
FERARI is aimed to define architectures and develop algorithms for efficient, real-time Big Data technologies of the future. The project intends to exploit the structured nature of M2M data while retaining the flexibility required for handling unstructured data elements. Taking into account the structured nature of the data will enable business users to express complex tasks, such as efficiently identifying sequences of events over distributed sources with complex relations, or learning and monitoring sophisticated abstract models of the data. http://www.ferari-project.eu/

- **Project FIRST** *(Large scale information extraction and integration infrastructure for supporting financial decision making, FP7-257928).*
  The FIRST project provided an information extraction, information integration and decision making infrastructure for information management in the financial domain. [http://project-first.eu/](http://project-first.eu/)

- **Project HPCFinance** *(High Performance Computing for Finance, FET).*
  A project to promote researchers to develop, integrate, and implement the most realistic and effective financial models on high-performance computing platforms. The ultimate goal is to help to improve the financial strength of banks, pension funds, insurance companies, other financial institutions and households in Europe.

- **Project FOC** *(Forecasting Financial Crises, FET).*
  The goal FOC is to better understand systemic risk and global financial instabilities by means of a novel, integrated and network-oriented approach. FOC has delivered several models of financial networks and indicators of systemic importance such as DebtRank. [http://www.focproject.net/](http://www.focproject.net/)

- **Project DIP** *(Data, Information and Process Integration with Semantic Web Services, FP6 – 507483)*
  A financial ontology was developed for an eBanking case study (mortgage simulator/comparator). The ontology contained financial services, financial products and a stock market ontology describing stock market operations. [http://dip.semanticweb.org/](http://dip.semanticweb.org/)

- **Project PARSIFAL** *(Protection and Trust in Financial Infrastructures, FP7-225344)*
  Was a 1.5 year FP7 project that targeted an objective concerning how to better protect Critical Financial Infrastructures (CFI). In particular PARSIFAL produced an ontology combining the ground ontologies from both the security and the financial sector, focused on three work areas: business continuity, control engineering, and trusted sharing of sensitive/confidential information. [http://www.parsifal-project.eu/](http://www.parsifal-project.eu/)
4.5. Conclusion and Recommendations

The Finance sector analysis for the roadmap is based on four major application scenarios based on areas such as sentiment analysis and exploiting bank’s and insurance companies’ own data to create new business models. The findings of this analysis show that there are still research challenges to develop the technologies to its full potential in order to provide competitive and effective solutions. These challenges appear at all levels of the big data chain and involve a wide set of different technologies, which would make necessary a prioritization of the investments in R&D. In general terms there seems to be a general agreement in the real-time aspects, better data quality techniques, scalability of data management and processing, better sentiment classification methods, and the compliance with security requirements along the supply chain. However it is worth mentioning the importance of the application scenario and the real needs of the end-user in order to determine these priorities. At the same time, apart from the technological aspects, there are other organizational, cultural and legal factors that have not been highlighted, but which will surely play a key role in how the Financial Services market takes on Big Data for its operations and business development.
4.6. Abbreviations and acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>Application scenario</td>
</tr>
<tr>
<td>CFI</td>
<td>Critical Financial Infrastructures</td>
</tr>
<tr>
<td>DSS</td>
<td>Decision Support System</td>
</tr>
<tr>
<td>FP</td>
<td>Framework program</td>
</tr>
<tr>
<td>FS</td>
<td>Financial Sector</td>
</tr>
<tr>
<td>IdM</td>
<td>Identity Management</td>
</tr>
<tr>
<td>REST</td>
<td>Representational State Transfer.</td>
</tr>
<tr>
<td>SOAP</td>
<td>Simple Object Access Protocol</td>
</tr>
<tr>
<td>SOX</td>
<td>Sarbanes-Oxley</td>
</tr>
<tr>
<td>SSL</td>
<td>Secure Sockets Layer</td>
</tr>
</tbody>
</table>
4.7. References


Lombardi, P. (2010): D1.2 Use case requirements specification. FP7-ICT project FIRST.

Lombardi, P. (2010): D2.1 Technical requirements and state-of-the-art. FP7-ICT project FIRST.

Lombardi, P. (2011): D1.1 Definition of market surveillance, risk management and retail brokerage use cases. FP7-ICT project FIRST.


Radzimski, M. (2012): D5.3 Large-scale integrated knowledge base. FP7-ICT project FIRST.

Reinhardt, M. (2013): D8.2 Final demonstration facilities and end-user prototypes. FP7-ICT project FIRST.


5. Sector Telecom, Media and Entertainment

5.1. Introduction

This section builds on previous deliverables of the BIG project to present a final roadmap for Big Data adoption within the telecom, media and entertainment industries. These sectors are advanced in technology usage compared with others such as Healthcare. The vast majority of actionable data in telecom and media is already in digital form (and analogue products such as newspapers have been created through digital technologies for some years now). However, this does not mean that organisations are deriving the fullest possible financial benefit or cost efficiencies from both their existing data and new sources of data.

5.2. Application scenarios

The research performed by the telecom and media sector identified several key application scenarios for Big Data technologies. They are described fully in deliverables (Zillner et al., 2013) and (Zillner et al., 2014). Concise information is provided here for ease of reference. In this deliverable, the telecom sector has condensed some of the use cases and added one specific use case in order to update the state of the art (section 0.)

5.2.1 Telecom

5.2.1.1 [UC1] A call to a telco Customer Care centre

<table>
<thead>
<tr>
<th>Summary</th>
<th>This scenario represents a future daily situation in which Big Data technology can bring telecom operators and customers together. The most interesting side of this story is how far this is from reality today in the Customer Care domain.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synopsis</td>
<td>Customer Care agents have immediate information concerning customer’s habits and incidences and are available to offer the best possible solution to any request.</td>
</tr>
</tbody>
</table>
| Business Objectives | • Improve customer experience  
  • Churn reduction  
  • Customised offering |

5.2.1.2 [UC2] Advanced customer segmentation

| Summary | This use case shows how the combination of data coming from different sources (web, geographical data, network information) can be used as an input for advanced customer segmentation and sales regions definition which can, in turn, help sales departments offering strategies. |
### Synopsis

Identification of user communities by analysing the traffic (calling networks). User communities are considered a segment, which is a group of customers that will react similarly to a message (age, neighbourhood, etc.).

<table>
<thead>
<tr>
<th>Business Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Improve customer experience</td>
</tr>
<tr>
<td>• Advanced segmentation</td>
</tr>
<tr>
<td>• Campaign definition</td>
</tr>
</tbody>
</table>

### 5.2.1.3 [UC3] Dynamic bandwidth increase

**Summary**

This example illustrates how the information retrieved by Call Centres can help identify infrastructure and network problems.

**Synopsis**

A few days after launching of a new gaming product, the operator observes a burst of calls to the call centres and on text mining the transcript data, specialists find a great increase in the keywords alluding to performance. The specific intelligence regarding keyword burst and specific time of day at which this was encountered can derive an automatic actuation on the network in order to dynamically change the provided the bandwidth based on usage.

<table>
<thead>
<tr>
<th>Business Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Improve customer experience</td>
</tr>
<tr>
<td>• Operational excellence</td>
</tr>
</tbody>
</table>

### 5.2.1.4 [UC4] Flexible network

**Summary**

This scenario represents a use case in which Big Data technology can bring important insights concerning the network situation and its adaptation in order to meet customers’ expectations in terms of quality of experience.

**Synopsis**

Prediction of demand and advanced customer profiling are used by the telecom operator to dynamically and automatically modify the network itself.

A Smart Stadium provides different possibilities as far as network access is concerned (WiFi, 4G, etc.). Information gathered from location-based applications and social media point out, thanks to prediction algorithms, that there will be an important concentration of mobile users in the stadium due to an important sports event. Following current technological trends, an operator providing 4G technology has adopted a virtualized P-GW (the PDN Gateway provides connectivity from the user terminal to external packet data networks such as the internet).

As the increase of traffic is forecasted (e.g. massive video uploading to Facebook during a match), the operator can flexibly and automatically duplicate the P-GW and apply load balancing so that customers are served by one P-GW or the other depending on their profile (user usually uploads videos, expected QoE, Facebook user, YouTube consumer,…) or, instead of duplicating instances, the operator may decide to automatically scale up the P-GW capacity in order to cope with the demand.

<table>
<thead>
<tr>
<th>Business Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Operational excellence to increase customer experience</td>
</tr>
</tbody>
</table>
### 5.2.2 Media and Entertainment

#### 5.2.2.1 [UC5] Data Journalism

<table>
<thead>
<tr>
<th>Summary</th>
<th>Large volumes of data become available to a media organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synopsis</td>
<td>A dataset or multiple datasets requires analysis to derive insight, find interesting stories, and generate material which can be enhanced and ultimately monetised by selling to the media organisation's customers.</td>
</tr>
</tbody>
</table>
| Business Objectives | • Improve quality of journalism and therefore enhance the brand  
• Analyse data more thoroughly for less cost  
• Enable data analysis to be performed by wider constituency of staff than just data scientists |

#### 5.2.2.2 [UC6] Dynamic Semantic Publishing

<table>
<thead>
<tr>
<th>Summary</th>
<th>Scalable processing of content for efficient targeting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synopsis</td>
<td>Using semantic technologies to provide the means to both produce and target content more efficiently.</td>
</tr>
</tbody>
</table>
| Business Objectives | • Manage content and scarce content creation staff more efficiently  
• Add value to data to differentiate from competitors |

#### 5.2.2.3 [UC7] Social Media Analysis

<table>
<thead>
<tr>
<th>Summary</th>
<th>Processing of large user-generated content datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synopsis</td>
<td>Batch and real-time analysis of millions of tweets, images, status updates to identify trends and content which can be packaged in value-added services.</td>
</tr>
</tbody>
</table>
| Business Objectives | • Create value-added services for clients  
• Perform large-scale data processing in a cost-effective manner |

#### 5.2.2.4 [UC8] Cross-sell of Related Products

<table>
<thead>
<tr>
<th>Summary</th>
<th>Developing recommendation engines using multiple data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synopsis</td>
<td>Applications that exploit collaborative filtering, content-based filtering and hybrids of both approaches.</td>
</tr>
<tr>
<td>Business Objectives</td>
<td>• Generate more revenue from customers</td>
</tr>
</tbody>
</table>

#### 5.2.2.5 [UC9] Product Development

<table>
<thead>
<tr>
<th>Summary</th>
<th>Using predictive analytics to commission new services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synopsis</td>
<td>Data mining to support development of new and enhanced products for the marketplace.</td>
</tr>
</tbody>
</table>
| **Business Objectives** | • Offer innovative new products and services  
| | • Enable development to be done in a more quantitative way than is possible currently |

### 5.2.2.6 [UC10] Audience Insight

| **Summary** | Using data from multiple sources to build up a comprehensive 360 degree view of a customer |
| **Synopsis** | Extension of scenario “Product Development” – mining of data external to the organisation for information about customer habits and preferences. |
| **Business Objectives** | • Reduce costs of customer retention and acquisition  
| | • Use insights to feed into commissioning of new products and services  
| | • Maximise revenue from customers |
5.3. Sector Requirements

The list of common requirements for telecom and media was analysed and provided in (Zillner et al., 2014).

The following table maps the use cases to the requirements:

<table>
<thead>
<tr>
<th>ID</th>
<th>Big Data Requirement</th>
<th>Telecom</th>
<th>Media and Entertainment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>UC1</td>
<td>UC2</td>
</tr>
<tr>
<td>R1</td>
<td>Single 360 degree customer view</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R2</td>
<td>Customer profiling and segmentation</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R3</td>
<td>Improve customer experience management</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R4</td>
<td>Product and service usage analytics</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R5</td>
<td>Reduce data infrastructure costs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R6</td>
<td>More efficient service delivery management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R7</td>
<td>Ease of solution implementation</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R8</td>
<td>Usability of tools for non-technical users</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R9</td>
<td>Improve speed and quality of data-driven innovation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R10</td>
<td>Stable regulatory framework</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R11</td>
<td>Reduce speed of processing new data sources and formats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R12</td>
<td>Data analysis for patterns, entities, sentiment etc. in data</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R13</td>
<td>Transform data from diverse sources or formats</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R14</td>
<td>Ingest and process data from networks, sensors and wearables and other automatically-generated data sources</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>R15</td>
<td>Manage data through full lifecycle</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R16</td>
<td>Real-time data analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R17</td>
<td>Process heterogeneous data sources</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R18</td>
<td>Content enrichment and aggregation</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R19</td>
<td>Tailored product and service offerings</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>R20</td>
<td>Recommendations and cross-sell based on multiple factors</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Telecom and media requirements and use cases mapping

The needs of telecom and media firms that are moving to data-driven business models are complex, but the requirements can be generalised to a certain extent, and then prioritised against objective criteria.

Some requirements underpin many different use cases; therefore the technical solution could have multiple applications and benefit diverse organisations. An example is processing of heterogeneous data sources – this is essential for all kinds of businesses whether they specialise in analysing externally-generated data or curating their own (or do both, of course).

Naturally, some requirements are also salient for industries beyond the telecommunications, media and entertainment applications described here. The deliverable D2.5 Cross-sectorial
roadmap consolidation will present the most important shared requirements across the ten sectors of the BIG project.

For each requirement, the Technical Working Groups requested extra information to be provided. Obviously the requirement itself is generalised across the sector (and beyond, in some cases) but the individual needs of a business will vary according to many intrinsic and extrinsic factors such as budget, overall strategy and location. For example, in evaluating Quality of Service, some businesses will have much higher tolerance for data errors than others, depending on the nature of the data they are processing.

In general, for the telecom and media sectors, the following aspects or qualities of data should be taken into account when considering the appropriate technological solutions:

- **Variety:**
  As mentioned in (Zillner et al., 2013) the variety of data in the telecom sector is quite large, including network data, billing data, voice recordings, etc. This implies multiple data formats: XML for bills, MP3 for voice recordings, text for customer complaints, proprietary formats for CDR collection, SIP messages, SS7, IP, etc.
  Media organisations of all sizes routinely handle data inputs that are both structured and unstructured, which are generated internally or ingested from external sources such as suppliers or from the public web. (In the case of social media streams such as Twitter, data may be accessed directly or purchased from licensed vendors). In addition, the media sector is increasingly making use of automatically-generated data from sensors, devices, wearable technologies etc. (British Library, 2014).

- **QoS requirements:**
  - Robustness – System failures affect differently depending on the use case. The most important aspect would probably be security, which can never fail. Real-time applications are also sensitive (e.g. prepaid). Tolerance for service outages of data-driven services that are delivered to customers will depend on how critical the data is deemed to be, also how valuable customers deem the services to be.
  - Correctness – Data quality is a very important aspect in the sector. Information is spread among different systems and there is a historical issue related to this. For example, addresses have to be normalised, customer complaints are written with different language styles, sometimes with hyphens, etc., which makes it more difficult to assess the data.
  - Timelines – This varies with the data source and on the use case. For example, roaming records are seldom collected in real time and can be carried over for the next bill. In prepaid, for example, the information has to be timely accurate. For media and entertainment, data processing and usage must be timely as the sector is very competitive, and speed is still a competitive advantage in many sub-sectors (despite the disruptive nature of disintermediation in enabling anyone, not just well-resourced companies, to produce high-quality content).

- **Acquisition**
  - Data can be structured (traditional business information systems information) or unstructured (social media data, voice recordings). Solutions must cope with all these data types, (alternatively, specialised solutions should enable processed data to be integrated with data processed by other systems).
As far as security is concerned, there are regulations for the custody of customer data. Communications are sometimes encrypted (depending on the technology used) and the customer data is kept in non-encrypted databases protected with passwords. As mentioned in (Zillner et al., 2013), there is currently no clear regulation as far as inferred data is concerned. The exploitation of personal data is a hotly-discussed topic within the EU, so technical solutions which enable businesses to store and manipulate personal data should be able to align with current regulations, and be able to cope with future changes in the regulatory landscape.

**Data sources:**

- For the quantification of the volume of data, (Zillner et al., 2013) provides some references for some data sources.
- Data can be provided by customers themselves (voluntary data) or can be observed, inferred by the service provider, as explained in (Zillner et al., 2013).
- Some data are timestamped (e.g. traffic information, rated records). Some data contain localisation information, provided that the user has granted access to it.
- Data changes can be faster or slower depending on the data source. For example, billing information is quite static once produced. However, the degree of customer satisfaction can quickly change, as human mood does. As illustrated by Intel in 2014 (Intel, 2014), in an “Internet Minute” 1,572,877 gigabytes of data globally is transferred over the internet, including web page views, searches, uploads, emails, and social media updates. These sources will change at variable rates.
- The data can be used by the end customer himself (through self-care systems), by customer care agents (through CRM software) or by business users (through business intelligence applications).
- Data consistency is crucial since information is usually spread among different systems (prepaid, postpaid, CRM, billing, network, etc.) All systems must be aligned. Data quality and provenance management are critical to the successful implementation of data-driven products and services.
- Availability requirements depend on the particular business process. For real time processes, availability is a must. For sales, customer acquisition etc, availability is crucial. However, there are other processes that do not require full time availability, such as postpaid rating, although any delay introduced can unchain further incidents that might be catastrophic, so availability is generally desired.
- Storage providers and other business players (e.g. system integrators) have access to the data and the access to information is restricted by confidentiality agreements and its management is usually protected by law.

- Data Visualisation depends on the commercial solution adopted by the service provider. There are a number of available Big Data platforms for the sector described in (Zillner et al., 2014).

- End users need to interact with the data, be able to explore “what if” scenarios, make decisions and change customer attributes or products. These procedures are usually controlled through predefined business rules.
5.3.1 Quantification of data

As far as data is concerned, the telecom sector has experienced a huge data volume growth in the last few years mainly due to the following drivers:

- Increase of popularity of mobile internet services which generate data on the devices, networks and systems:
  - According to a recently published mobile phone forecast from the International Data Corporation (IDC, 2014) Worldwide Quarterly Mobile Phone Tracker, worldwide smartphone shipments will reach a total of 1.2 billion units in 2014, marking a 23.1% increase from the 1.0 billion units shipped in 2013. From there, total volumes will reach 1.8 billion units in 2018, resulting in a 12.3% compound annual growth rate (CAGR) from 2013–2018.
  - Global mobile data traffic grew 81 per cent in 2013. Global mobile data traffic reached 1.5 exabytes per month at the end of 2013, up from 820 petabytes per month at the end of 2012 (Cisco, 2014).

- Context information is increasingly used. The European Digital Agenda (EDA) has targeted an increase of regular internet use to 75% of the population to be reached in 2014, although still ahead of the EDA target year of 2015 (EDA, 2014).

- Rise of social networks used to upload messages, pictures, videos and all sorts of files.

- Mobile applications downloads. These apps are not producing revenue by themselves (since their cost is usually very low), but they drive hardware sales, advertising spending and technology innovation. And they produce lots of data. Statistics shows a forecast for the number of mobile app downloads from 2009 to 2017. In 2009, worldwide mobile app downloads amounted to approximately 2.52 billion and are expected to reach 268.69 billion in 2017. In 2010, earnings of mobile apps providers amounted to 6.8 billion U.S. dollars (Statista, 2014).

- Cloud services in expansion, bringing more remote storage and processing, which is also related to a greater production in traffic across networks. By 2017, nearly two-thirds of all workloads will be processed in the cloud. As for Global cloud traffic, annual global cloud IP traffic will reach 5.3 zettabytes by the end of 2017. By 2017, global cloud IP traffic will reach 443 exabytes per month (up from 98 exabytes per month in 2012) (Cisco 2014).

Telecom operators have access to varied data that come from different sources. Here is a possible classification of data according to the level of complexity in the process of gathering this data:

- **Volunteered data**: This is natural data customers provide for registration.
  - Examples: Name, address, date of birth, gender, profile, preferences, number, SIP address, IP address, SIM, Soft SIM, serial number, device details, subscription date, job, marital status, voice recordings for product subscription, etc.
  
  An operator with a customer base of aprox. 40M subscribers (including interested customers not having activated a contract yet, prepaid, postpaid, etc.) requires minimum 500GB to store this basic information only for mobile.
• **Observed data**: This data is retrieved from the service usage and is based on information at customer care, billing systems, network, etc. Some of these data needs to be stored for a certain period of time according to national laws (e.g. bills).
  
  o Examples: billing information, internet access, call origin and destination, reason of session failure, service delivery completeness, service used, complaints reported, products purchased, locations (GPS or cellular), on/off roaming, etc.

  Mobile data traffic will reach the following milestones within the next five years (Cisco White Paper, 2014)

  - Monthly global mobile data traffic will surpass 15 exabytes by 2018.
  - The number of mobile-connected devices will exceed the world’s population by 2014.
  - The average mobile connection speed will surpass 2 Mbps by 2016.
  - Due to increased usage on smartphones, smartphones will reach 66 per cent of mobile data traffic by 2018.
  - Monthly mobile tablet traffic will surpass 2.5 exabyte per month by 2018.
  - Tablets will exceed 15 per cent of global mobile data traffic by 2016.
  - 4G traffic will be more than half of the total mobile traffic by 2018.
  - There will be more traffic offloaded from cellular networks (on to Wi-Fi) than remain on cellular networks by 2018.

  A telecom operator might produce 1TB of raw packet-switched signal data per day and store it in a file format. After processing, 550 GB xDR signal data is generated per day and is saved in a database format. This data is often saved for a few days or a few months (ZTE, 2012).

  For example, according to (DigitalRoute, 2012), in a mobile network, there are typically online transactions between network gateway nodes and online charging or mediation systems for charging purposes. However, these volumes only represent a fraction of the usage data from the network, which primarily comes in the shape of records that are transferred at discrete intervals. It is common for network operators to set these intervals to 10 to 15 minutes and batch process data at off-peak hours. For end-user communication, this is not an acceptable solution. Customers demand accurate information about their usage through, for example, SMS, IVR, smartphone applications or USSD. Sending these notifications and alerts in a timely manner requires that usage data is continually collected at short intervals. With mobile data speeds of eight to 20 Mbit/s, reaching a 1GB allowance within minutes is a practical reality. A typical tier 2 operator in Europe takes in about 200 million records a day on the BSS side (Billing World, 2014).

  According to Cisco, there will be 7.3 billion M2M connections globally, or nearly one M2M connection per capita, based on a 7.6 billion population by 2018 (Cisco Newsroom 2014).

• **Inferred data**: This requires some analysis on volunteered and observed data.
Examples: most used IP address, preferred contact channels, last known location, destinations visited frequently, etc.

By analysing user profiles, product packages, services, billing, and financial information, operators can obtain precisely control policies. Web pages, messages, pictures and movies, and other traffic delivered through the network can also be analysed to better understand user behaviour.

A marketing portal provides daily and monthly statistical reports on data flow, revenue, subscriber development, warnings, and summary tree structure. The amount of data added each month can reach up to 4 TB. Usually, it takes 26 hours to analyse 4 TB of data using a traditional method that is inefficient and cannot adequately deal with system expansion.

- **Social data**: This is personal social data on social networks such as Linkedin, Facebook or Twitter.
  - Examples: number of followers, number of people following, influence rate per follower/followed, contact intensity, mood attributes, recommendations to and by others, school, lifestyle, opinions about products owned, opinions about customer service, opinion about telcos, etc.

  Facebook users post nearly 2,460,000 pieces of content a minute (Worldpress Trail, 2014)

  According to (#TCBlog, 2011), more than 70% of relevant conversations in the telecom sector are in twitter and in forums, which clearly represents an opportunity.

- **3rd party data**: This involved data coming from business partners such suppliers (points of sale), insurance, airlines or banks, for example:
  - Examples: Bank information, credit cards, consumption habits, lifestyle, frequent flyer information, etc.

The more complex the data, the greater business value that can be extracted from it. This is shown in the following figure:
Besides, the data in this sector present other characteristics:

- Multiple data formats: xml for bills, mp3 for voice recordings, text for customer complaints, proprietary formats for CDR collection, SIP messages, SS7, IP.
- Highly transactional: this implies a regular, high-volume stream of records entering the system.
- Multiple Devices causing high data generation rate.

5.3.2 Quantification of business priorities

The Telco & Media Sector Forum ran a survey among relevant players in the sector in order to help find the priorities that need to be overcome first for the uptake of Big Data in Europe. The aggregated results of this survey are shown in the next picture (please refer Zillnet et al. 2014 to for further details).
Finally, concerning privacy, we highlight the fact that Ovum’s latest Consumer Insights Survey (Ovum) reveals that 68 per cent of the Internet population across 11 countries around the world would select a “do-not-track” (DNT) feature if it was easily available, which clearly highlights some amount of end users’ antipathy towards online tracking. This brings about an important barrier since data must be rich in order for businesses to use it.
5.4. Mapping Requirements to Technology

In order to define a roadmap of research questions for the telecom, media and entertainment sectors, the first step is to identify the technologies available for each requirement. This work was partially covered in (Zillner et al., 2013) and has later been further developed along with the information provided by the technical working groups in the project. For this, we have grouped requirements that correspond to the same technology domain as follows:

<table>
<thead>
<tr>
<th>ID</th>
<th>Big Data Requirement</th>
<th>Domain /technologies</th>
<th>Research questions</th>
</tr>
</thead>
</table>
| R1  | Single 360 degree customer view          | All these aspects are usually related to **Big Data end-user software**, such as CRM packages and business user applications which offer data visualisation and whose results depend strongly on overall data quality along the full data lifecycle. | • Automation of data quality assessment process without human intervention that ensures an efficient data quality process avoiding manual corrections, incorrectness, etc.  
• Inclusion of many data sources and formats in data quality algorithms, including CDRs (timeliness), billing records (accuracy), customer agents free text (hyphens, orthography, etc.), audio (understandable recordings, etc.).  
• Automatic generation of high quality visualisation from semantic representations for end users to have interpreted information. Business users and customer care agents require high-level information in order to be able to take decisions and evaluate/accept next steps.  
• Tools and methods to help make complex modelling techniques accessible to a wide constituency of business users. This is important because it will enable business-critical data to be manipulated by more people in the organisation than just a few data scientists. For example, in the news sub-sector, enabling journalists to curate data using ontological models and other semantic standards without requiring the staff to be programmers or semantic data specialists. |
| R3  | Improve customer experience management    |                                                                                      |                                                                                                                                                      |
| R8  | Usability of tools for non-technical users|                                                                                      |                                                                                                                                                      |
| R19 | Tailored product and service offerings    |                                                                                      |                                                                                                                                                      |
| R2  | Customer profiling and segmentation       | These requirements relate to the ability to improve the **commercial offer** of the service provider based on the available information in traditional systems as well as advanced techniques such as predictive, speech or prescriptive analytics. | • Easily understandable network information at customer care so that customer care agents can include this kind of information in their speech without further technical knowledge concerning the underlying technology.  
• Processing of speech to determine e.g. mood and other subjective patterns. During an interaction with customer care, the mood of the customer changes quickly as conversation goes on also depending on subjective aspects. This |
<p>| R4  | Product and service usage analytics       |                                                                                      |                                                                                                                                                      |</p>
<table>
<thead>
<tr>
<th>ID</th>
<th>Big Data Requirement</th>
<th>Domain /technologies</th>
<th>Research questions</th>
</tr>
</thead>
</table>
| R9  | Improve speed and quality of data-driven innovation |                                                                                     | - Dynamic calculation of degree of satisfaction. This is also related to determining the next best decision dynamically, based on different inputs such as social network information (what they are saying on forums, for example), billing (an expensive bill that has just been issued), etc.  
- Modelling high velocity data. This is needed so that businesses can rapidly make sense of and thus exploit data that is in or near real-time. |
| R20 | Recommendations and cross-sell based on multiple factors |                                                                                     |                                                                                             |
| R17 | Process heterogeneous data sources            | These relate to the data management in general, which 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 or audio, for example). | - Unified information system including also social media data (Inclusion of social media information in SID model). Telecom information model is well established by eTOM, except social media information. It is important to be able to catalogue this information among the rest.  
- Correlation of customer, traffic data and social network information. These data coming from different sources and in different formats, produced by heterogeneous systems have to be processed together.  
- Scalability of semantic data storage e.g. graph databases. How can scalability of graph databases be improved? Currently there are solutions in the marketplace but the leading solution has a maximum capacity of 34 billion nodes⁴, and has known limitations around database sharing that relational databases do not have.  
- Standards development and adoption. The use of standards is very important for data assessment and transformation, including integration of data coming from heterogeneous systems. This is essential to manage data lifecycle and its complexity. How can we define an efficient data management system that has not any a priori knowledge about the type of data, its execution platform or data usage? |
| R13 | Data transformation                           |                                                                                     |                                                                                             |
| R15 | Data management through full lifecycle        |                                                                                     |                                                                                             |
| R12 | Data analysis for patterns, entities, sentiment etc. |                                                                                     |                                                                                             |
| R14 | Process data from networks, sensors and wearables, etc. |                                                                                     |                                                                                             |
| R18 | Content enrichment and aggregation            |                                                                                     |                                                                                             |
| R6  | More efficient service delivery management    | This requirement relies on technologies that enable the automation (when possible) of tasks thanks to Big Data techniques. This requires complex event processing, data interoperability, sector specific management systems, etc. | - Big Data technologies in the domain of Network Function Virtualisation (NFV) and software Defined Networks (SDN). NFV and SDN bring flexibility and vendor independence to service providers by replacing the legacy proprietary hardware solutions by flexible virtualised functions that can scale up and down. Big Data can help analyse the impact of NFV and SDN on final services provided to end users.  
- Retrieval of information across multiple network domains and technologies. Big data solutions have to cope with different telecom technologies, wireless, wireline, and even with several generations and standards for each of these |

---

⁴ https://groups.google.com/forum/#!topic/neo4j/ts6H9Py-2I
<table>
<thead>
<tr>
<th>ID</th>
<th>Big Data Requirement</th>
<th>Domain /technologies</th>
<th>Research questions</th>
</tr>
</thead>
</table>
|    |                      | (LTE, 5G, XDSL, etc.)| • Big Data solutions can benefit from NFV and SDN and even evolve so that analytics is carried out in a distributed manner in the network nodes themselves when required, thanks to virtualisation.  
• Actuation across multiple network domains and technologies. Predictive analysis to be used to actuate over the network in order to prevent SLA violations, for example. Operational excellence can be achieved thanks to Big Data technologies.  
• Ability to automate tasks according to policy. Automation requires business and/or operational rules to be well defined. |
| R7 | Ease of solution implementation | Big data solutions must have **agile implementations** and evolutions for service providers to adopt them. For this, semantic based technologies, ontologies and standards in general are required. | • Standardisation of data makes integration easier. This relates to data management.  
• Addition of abstraction layers to aid the integration of new sources into the data architecture and make them easier to manipulate by data scientists  
• Technologies for integration and correlation of structured and unstructured data sources without the need of structuring the unstructured ones in advance. |
| R11 | Reduce speed of processing new data sources and formats | **Real time processing** can be carried out using technologies that help make sense of incoming data streams. This could include in-memory databases and complex event processing, also speech processing for analysing voice data as it is being generated. There is also a role here for technologies that model streams of data, including machine learning, if this is scalable and fast enough to deal with the speed of the data coming in. | • Performance of high velocity data storage solutions e.g. improving the read/write capability of databases.  
• In-network analytics is a way of achieving faster reaction when possible. This requires further research since network nodes are not currently prepared to carry out analytic tasks (virtualisation can help).  
• Speech processing is a resource consuming process, how can it be done in larger scale and in less resource-intensive ways? Speech processing that is integrated into call centre software could monitor customer calls for actionable insights or fraud  
• Integration of machine learning into databases so that databases become more than just passive stores of data. Machine learning as an integrated application to scan for patterns and identify trends of interest. |
| R16 | Real-time data analysis | **Early data curation** contributes to data quality enhancement within sector-specific software which, in turn, helps reduce storage costs. The selection of the correct data to achieve a concrete Big Data strategy is also a good approach since it reduces the amount of data to be stored. | • Embedding early data curation techniques within sector specific tools. Crowdsourcing platforms.  
• Blended curation at scale using algorithmic and manual approaches in order to filter, classify or enhance large volumes of data at different points of the data lifecycle, leading to efficiencies in storage.  
• Data-agnostic architectures that can adapt as data and business needs |
<table>
<thead>
<tr>
<th>ID</th>
<th>Big Data Requirement</th>
<th>Domain /technologies</th>
<th>Research questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>change.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Application of semantic technologies to raw data e.g. identifying entities and meaning from unstructured datasets.</td>
</tr>
</tbody>
</table>

Table 5. Technologies and research questions for Telecom, Media and Entertainment

The requirement related to the regulatory framework (R10) does not refer to any specific technology but to legal and policy-related aspects, which is why it has not been considered in the technical roadmap but in section 5.6 along with cultural and economic aspects.
5.5. Technology Roadmaps

The following diagrams show a high-level view for the technology roadmaps in the telecom, media and entertainment sectors. These are summaries of Table 5 and are also grouped by functional domains. However, these roadmaps do not highlight the most important areas to address at a strategic level, that is, at EU policy level since this is tackled in section 5.6.

Figure 14: Roadmap for Telecom, Media and Entertainment Sector (1/3)
## Technology Roadmap
**Telecom & Media Sector / 2/3**

### Technical Requirement

<table>
<thead>
<tr>
<th>Machine learning</th>
<th>Unified information system including also social media data (inclusion of social media information in SID model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>eTOM and SID data ontologies</td>
<td>Correlation of customer, traffic data and social network information</td>
</tr>
<tr>
<td>Automated schema mapping</td>
<td>Scalability of semantic data storage e.g. graph databases</td>
</tr>
<tr>
<td>Standardized and open data formats</td>
<td>Standards development and adoption</td>
</tr>
<tr>
<td>Descriptive analytics</td>
<td>Managing data lifecycle scale and complexity</td>
</tr>
<tr>
<td>Platforms for blended automated and crowdsourced curation</td>
<td></td>
</tr>
<tr>
<td>Data abstraction</td>
<td></td>
</tr>
</tbody>
</table>

| Network management systems                           | Big Data technologies in the domain of NFV and SDN                                                      |
| Proprietary network technology                        | Retrieval of information across multiple network domains and technologies                            |
| Complex event processing                              | Virtualisation of analytics functions for in-network processing                                        |
| Descriptive analytics                                 | Actuation across multiple network domains and technologies                                            |
| Data interoperability                                 | Ability to automate tasks according to policy                                                          |
| Distributed processing architectures                   |                                                                                                        |

<table>
<thead>
<tr>
<th>Technology</th>
<th>Research Question</th>
</tr>
</thead>
</table>

*Figure 15: Roadmap for Telecom, Media and Entertainment Sector (2/3)*
The following figure includes different technologies and research questions in a temporal framework. The diagram is based on the information provided by the technology working groups of the BIG project to the questions included in sections 5.5.1 to 5.5.5.
5.5.1 ACQUISITION

Correlation of customer, traffic data and social network information

How do we combine all this information including unstructured data? Do we have to structure everything before processing the data?

In general, the data has to be structured. The data from different data sources can be stored in different data repositories and must integrated before it is used. Here, using Unified Resource Identifiers (URIs) while generating the data can improve the integration process, especially if the URIs are reused.

Retrieval of information across multiple network domains and technologies

Can we automate the acquisition of the data coming from heterogeneous network domains, all of them having different network nodes and protocols?

This is again more an integration problem. Gathering and storing the data is not a problem, as the streaming technology (see Storm for example) provides means to put all these streams together. The more difficult part comes downstream, i.e., when the data has to be put together for the sake of analysis. Here, converting the data to tables or RDF for the sake of entity resolution/link discovery becomes a matter of knowing the protocols exactly (including the meaning of all fields).
Standardisation
What is the situation in terms of standardisation in the domain of data acquisition?

Some protocols are standardized (e.g., AMQP) while other are quasi-standards (e.g., JMS2.0). The transport for this data depends on the devices but usually flows into some HTTP-capable solution at one point (latest when the data has to go to storage). Please refer to (Becker, 2014) for further information.

Data-agnostic architectures that can adapt as data and business needs change.
What are the best solutions to add new data sources with minimal impact on existing big data solutions?

There is no generic answer to this question. From an architectural point of view, most (if not all) solutions provide means to add sources from which the data is to be gathered. If Storm terminology, this would simply mean adding one spout to the processing pipeline, which can be done at runtime.

5.5.2 CURATION

Automation of data quality assessment process without human intervention
What technologies do we have and how can they fulfill this? What do they lack?

Blended algorithmic and human data curation approaches can increase the level of automation on data quality tasks, automatically addressing easier data quality tasks and routing more complex data curation tasks to human curators. Despite the available infrastructure for human computation, such as crowdsourcing, robust infrastructures for the automated detection, classification and correction of data quality tasks remain unsolved.

Inclusion of many data sources and formats in data quality algorithms
Can quality algorithms deal with different formats or do we need one data quality tool for each data source?

Most data curation tools available today are targeted towards structured data. Automated tools for data quality assessment and improvement are typically domain and format-specific.

Making complex modelling techniques accessible to wide constituency of business users
Are there any robust commercial technologies that allow journalists and other information professionals to manipulate data modelled as ontologies or semantic models?

Most of the existing data curation tools are committed to instantiating rigid conceptual models, using form-style interfaces for data insertion. There is a lack of generic methods and tools to support end users generating and curating data.

Standardisation
What is the situation in terms of standardisation in the domain of data curation?
Standardised data model formats such as **RDF** (Resource Description Framework) provide an integration layer for structured data in different formats. **Entity recognition and linking** already provide a basic infrastructure for integrating structured and unstructured data.

The importance of data standards which facilitate data integration and reuse is widely recognised as a relevant component in data curation infrastructures. However, the recognition at the discourse-level does not match the real adoption, as most data management infrastructures associated with data curation initiatives are based on relational data. There is an early-stage adoption in data curation projects, mostly focused on facilitating the reuse and integration of output data. Many interviewees reported the plan to move into an RDF-based data infrastructure.

**Integration**

Entity recognition and linking already provide a basic infrastructure for integrating structured and unstructured data. Infrastructures for **relating extraction and ontology population** still depend on a research effort.

**Ability to automate tasks according to policy**

Algorithmic data curation is still highly domain-specific and consists today of scripting recurrent tasks such as data validation. The application of approaches which applies machine learning techniques to data curation activities, such as curation by demonstration, is a fundamental demand to scale data curation to large datasets.

**Embedding early data curation techniques within sector specific tools. Crowdsourcing platforms**

*What are the main trends in early data curation? Can these be applied in customer care software solutions as the ones used in telco and media?*

Early data curation is strongly associated with delegating part of the curation effort to users, who curate data at source. The minimization of the disruption on existing processes and of the effort involved in the data curation is an important aspect in this scenario. Minimum information models provide the methodological/technical analysis to minimize users’ efforts. Cultural and methodological aspects play a fundamental role in curation at source, where user engagement needs to be fostered.

**Blended curation at scale using algorithmic and manual approaches**

*What technologies are available to handle manual curation at an enterprise level (e.g. a global corporation) when crowdsourced approaches are not viable for reasons of data sensitivity, quality tolerance etc.?*

There are basic infrastructures in place to support the setup of a data curation infrastructure. Data curation frameworks such as **Open Refine** and **Data Tamer** are the main early stage players in this space. **Karma** is a research-level data curation framework. These approaches can be integrated with crowdsourcing-level platforms which target organisational use cases, such as CrowdFlower. **Robust, generic and highly automated data curation infrastructures** are not fully in place and there is a major growth space for the evolution of new functionalities in tools. The integration of **robust privacy mechanisms** in these tools is also a major unaddressed issue.

**Application of semantic technologies to raw data**

*What applications if any exist to allow manual review of extracted semantic information?*
NLP has improved over recent years but are there any step changes on the horizon to improve quality?

A generic tool to support the curation which can be integrated to NLP pipelines is a major demand.

### 5.5.3 STORAGE

As far as standardisation in the domain of data storage is concerned, the ISO/IEC JTC 1 Study Group on Big Data (SGBD) has recently released a final report to JTC 1, the information technology committee of ISO/IEC. In this report the study groups have identified the following potential gaps related to Big Data storage:

- Query languages including non-relational queries to support diverse data types (XML, RDF, JSON, multimedia, etc.) and Big Data operations (e.g. matrix operations)
- Semantics of eventual consistency
- Big Data security and privacy access controls.
- Data sharing and exchange
- Data storage, e.g. memory storage system, distributed file system, data warehouse, etc.
- Interface between relational (SQL) and non-relational (NoSQL) datastores

The study group has identified that these topics are in the scope of existing subcommittees (SCs) of JTC 1 and recommends that they are explored further. The report provides a mapping to existing SCs. In addition SGBD recommends the creation a new working group (WG) that addresses new work items. This includes 1) the definition and vocabulary of Big Data and 2) the definition of a Big Data Reference Architecture.

For these recommendations the decision from JTC 1 is pending, but the activities show that in the next few years standards for managing, storing, protecting and exchanging data may emerge.

**Integration of machine learning into databases to scan for patterns and identify trends of interest**

The technologies here are analytical databases (e.g. SciDB, Rasdaman) as described in (Becker, 2014). It also relates to standardization as ArrayQL is an example of defining an interface that would satisfy this requirement.

**Performance of high velocity data storage solutions**

*What are the most efficient technologies for reading/writing databases? How fast are they? What are the trends?*

There is no absolute answer to these questions. Limitations of random read/write access are addressed by the **Lambda architecture** as an industry best-practice approach to satisfy real-time processing at scale. Depending on the application and budget different solutions can be applied such as e.g. **columnar stores** and/or **in memory databases** (which seems to be the trend). Please refer to (Becker, 2014) for further details.
Another trend consists on trading off query accuracy with response time (Approximate Query Processing) as e.g. Blinkdb.org proposes. However, typical examples are response time of two seconds which may not be considered real-time depending on the application requirements.

The research question improving random/read write performance is pertinent. From a timeline perspective this is hard to describe since continuous improvements are expected by technical experts.

However, as far as real time is concerned, the solutions do not seem to fulfil a real time requirement, which would be essential for example in the case of quality of experience prediction, since this changes as fast as human mood does.

**Scalability of semantic data storage e.g. graph databases**

*How can scalability of graph databases be improved?*

This question seems to be difficult to answer. Graph DB vendors are working on graph databases to detect subgraphs that have minimum connectivity to other subgraphs. This would allow shading the subgraphs to different nodes and minimize communication between nodes. However there are challenges when edges between nodes are added, for example.

The timeline would depend on the particular use case (how many nodes and edges must it support, graph topology, queries involved, etc.)

### 5.5.4 ANALYSIS

There are a whole set of standards related to the Semantic Web technology stack:

- Languages
- Querying
- Reasoning
- Mapping from other types of data sources to semantic data sources.

**Easy and understandable network information at customer care**

*Is there any semantic tool that helps retrieve understandable conclusions out of very technical data?*

One can use OWL reasoning and SPARQL to reason over and query semantic information – assuming that the data is structured in a semantic form. Natural language generation techniques can turn the answers into natural language although this are needs more work (3 more years from the time of writing this document).

IBM’s WATSON technology beat the best of human competition in answering questions on a popular US gameshow. This technology is now being applied in Healthcare and finance. See for example (Techcrunck, 2014) on WATSON’s analytics capabilities.
Dynamic calculation of degree of satisfaction

What are the technologies that help combine all this information and provide a model that determines what could work to satisfy a customer depending on his background? What limitations?

Semantic technologies can be used to efficiently design user profiles where individual features and preferences are specified to be used to provide personalized recommendations. Different types of information can be included in the models along with the integration framework which allows to manage different information coming from different sources, especially technologies associated with Linked Data.

Correlation of customer, traffic data and social network information

How do we combine all this information including unstructured data? Do we have to structure everything before processing the data?

One can provide a common ontology or map between the ontologies related to the data sources. Not all the information needs be structured as such, but documents can be semantically annotated and then related to the same integration framework. Chris Bizer has written about how Linked Data can support a pay-as-you-go data integration framework (Chris Bizer, 2014).

Big Data technologies in the domain of NVF and SDN/Integration

How can we define an efficient data management system that has not any a priori knowledge about the type of data, its execution platform or data usage?

This is essential to manage data lifecycle and its complexity.

A trade-off must be found between categorising the data and having an efficient data management system. It is impossible to build efficient data management systems without any knowledge of the data. A key feature of semantic technologies such as linked data is that the data schema (ontology or vocabulary) is a first class citizen and can be manipulated easily. Unlike standard database systems where important reworking is required when the data schema is altered.

Data-agnostic architectures

What are the best solutions to add new data sources with minimal impact on existing big data models?

This question does not seem to be solved in the Big Data context according to experts. Linked Data based systems enable easy integration handling ‘broad data’ as proposed by (Jim Hendler, 2014) but there is no system that handles integration easily across the three Vs.

5.5.5 USAGE

Automatic generation of high quality visualisation from semantic representations for end users to have interpreted information

What are the current limitations of visualisation tools (e.g. volume of represented data)? What are the trends? What are the tools?
Visualisation of semantics typically is based on the graph structure of semantic representations, but can also come from available metadata. Sophisticated tools for visualisation of graph data and relational data exist. Some semantics can be expressed by labelling or through its relation to temporal or spatial (geographic) data that can already be visualised. High-level information can also relate to trends analysis and ‘clustering analysis’ and abstraction and ‘dimensionality reduction’ (four trends) and it is mainly dependent on the quality of such analysis algorithms more than corresponding visualisations. Regarding speed, GPU-based visualisation systems can easily display tens of millions of data sets on a map in interactive speed (30 refreshes per second) and systems with multiple GPUs scale to the order of a billion data points.

Processing of speech to determine e.g. mood and other subjective patterns.

*What are the main tools for speech analytics in real time and how performing are they?*

Speech analytics is a multi-step process: The initial step for spoken language is automated speech recognition (ASR), transforming audio data into transcriptions. The next step is a pipeline of Natural Language Processing (NLP) modules, analysing the content of the transcribed text. Possible output of the analysis for scalable, big data processing are recognised named entities (e.g., persons, organisations), events (e.g., changes in management, filing of a new patent), relations (e.g., sequences, reasons), sentiment (e.g., positive or negative opinions towards specific subjects) and opinions. Performance is far from comprehensive: coverage and error rates are constantly improving and computing efficiency is already highly scalable for most (machine-learning based) algorithms. Practical applications are built for complex tasks, e.g. the mining of the complete twitter data stream to extract opinions with regards to an election and thus form a snapshot "How is the web voting?"

Full Natural Language understanding is still a topic of basic research, as full models for semantic representation are lacking. The largest on-going project for semantic modelling is probably the Google Knowledge Graph which is not easily available.

Modelling high-velocity data

*Real time tools: what kind of information can they cope with? how much? How fast?*

This applies to complex event processing (CEP). Systems look at a given time window of a "continuous" data stream, anywhere from a few milliseconds to (typically) a few minutes to hours. Processing complexity is limited, the typical processing model is the application of business rules to extract data and recognise high-level events that can, e.g., trigger alarms. Updates are fast, depending on the application below one second up to under 10 microseconds. Up to some 100k data points per second can be processed and tools scale to medium numbers of CPUs (e.g., 32) for parallel processing. For large data sets, in-memory computing is the most successful technology.

Ability to automate tasks according to policy

*Decisions to actuate should be taken automatically but going too fast could unstabilize the network*

Essentially, this is again complex event processing with respect to the automation of actuating based on the recognition by business rules. Keeping the network stable is a problem for control theory that can probably not be solved formally, but would be addressed by simulations that give statistical guarantees for the stability of the rules used by the CEP system.
Natural language processing at scale

Speech processing is a resource consuming process, how can it be done in larger scale and in less resource-intensive ways?

Current audio-reduction technologies and protocols are sufficient to pre-process (low-level processing) audio in smartphones and send it through (low-bandwidth) mobile data connections to servers for further, high-level processing. Prime examples are the smart assistance technologies in Google Now, Siri, and Cortana. Large scale-applications depend on balancing quality of performance (which is far below 100% in any case) with computing efficiency. Such parameterised machine-learning algorithms exist that can trade accuracy vs. computing efficiency.

Full Natural Language understanding is still a topic of basic research, as full models for semantic representation are lacking. The largest on-going project for semantic modeling is probably the Google Knowledge Graph which is not easily available.

Standardisation of data

What is the situation in terms of standardisation in the domain of data usage?

That is an extremely wide question. Standardisation indeed would be needed for (i) data integration, but also for (ii) data security and privacy, (iii) frameworks and tools, and (iv) benchmarks for comparison of tools and solutions. Currently, there is scarcely any standardisation directly related to Big Data or specifically Big Data Usage. As far as data integration is concerned, a representation framework for the semantics of data exists: RDF (Resource Description Framework) and Linked Data, however, the actual instances, i.e., ontologies and terminologies are not standardised. A very big data base would be needed as a comprehensive, normative standard. Such de-facto standards exist only in a few sectors, e.g., in the Health and Life-Sciences sectors. The Google Knowledge Vault (Graph) is a very big collection of some 1.6 billion general facts that would have the potential as a standard, however, it is not freely available.

5.5.6 Related EU projects

In Commissioner Neelie Kroes’ keynote speech at ICT 2013 in Vilnius¹, she talked of how big Data was “not just a new sector, but a new asset class. One that sits as a pillar of our economy, like human resources or financial capital.” She went on to talk about the work the EU is doing with regard to legislation, investment and building ecosystems.

Three relatively new projects in particular are helping to support the EU’s objectives in promoting the strategic value of Big Data to Europe’s businesses:

BYTE²

“The Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities (BYTE) project will assist European science and industry in capturing the positive externalities and diminishing the negative externalities associated with Big Data in order to gain a greater share of the Big Data market by 2020.”

² http://byte-project.eu/
RETHINK big

“The objective of the RETHINK big Project is to bring together the key European hardware, networking, and system architects with the key producers and consumers of Big Data to identify the industry coordination points that will maximize European competitiveness in the processing and analysis of Big Data over the next 10 years.”

MICO² (Media In Context)

“MICO is a European Union part-funded research project to provide cross-media analysis solutions for online multimedia producers. MICO will develop models, standards and software tools to jointly analyse, query and retrieve information out of connected and related media objects (text, image, audio, video, office documents) to provide better information extraction results for more relevant search and information discovery.”

5.6. Cultural, policy and wider economic aspects for Telco and Media

As mentioned in section 5.4.8 of (Zillner et al., 2014), the telecom and media sectors do not only have technical requirements. There are a number of non-technical aspects that should be considered in the roadmap. The following table summarises and condenses some of these crucial aspects:

<table>
<thead>
<tr>
<th>Aspect nature</th>
<th>Blocker</th>
<th>Some potential solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culture</td>
<td>Top Level Management scepticism concerning Big Data technologies</td>
<td>Standardisation</td>
</tr>
<tr>
<td></td>
<td>End user privacy concerns</td>
<td>Clear regulation and regulation adoption</td>
</tr>
<tr>
<td></td>
<td>Marketing buzz around “Big Data”</td>
<td>Standardisation</td>
</tr>
<tr>
<td>Policy</td>
<td>End user privacy concerns</td>
<td>Market regulation as proposed by the EU (<a href="http://ec.europa.eu/justice/data-protection/index_en.htm">http://ec.europa.eu/justice/data-protection/index_en.htm</a>) and (European Commission)</td>
</tr>
<tr>
<td></td>
<td>Unclear regulatory framework</td>
<td></td>
</tr>
<tr>
<td>Economy</td>
<td>Lack of specialized professionals</td>
<td>- University programs to keep pace with new technologies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- More European initiatives are required in order to support entrepreneurs and SMEs throughout the EU to get access to the resources and people they need to foster innovation and growth</td>
</tr>
</tbody>
</table>

Table 6: Non-technical aspects for telecom and media

¹ http://www.rethinkbig-project.eu
² http://www.mico-project.eu
5.7. Conclusion and Recommendations

The Telecom, Media and Entertainment sector of deliverable (Zillner et al., 2014) has highlighted the most important requirements for the main application scenarios of the sectors. They were then analysed with regard to technology solutions and to outstanding research questions that must be examined to give the sector (and others) the greatest opportunities for success in Europe over the coming years.

The most critical theme that emerges from our analysis of the Telco, Media and Entertainment sectors is that businesses competing in these markets need to be able to handle diverse data sources. In many cases, these sources change rapidly and in large volumes. For example, during the Brazil versus Germany football World Cup semi-final, 35.6 million tweets were sent\(^1\). This is still far too much data for all but the most technically advanced media businesses to successfully exploit.

Secondly, the need to derive insights from data, regardless of where it’s come from and what it’s about, which is important across many of the application scenarios in both the Telco and Media sectors. Analysis solutions are needed for all levels of insight such as descriptive, diagnostic, predictive and prescriptive.

Thirdly, the strategic driver to add value to data, for instance through semantic enrichment or tailored product offerings, is a particular feature of many aspects of operating in Telco or Media. The disruptive changes in these industries over the last 20 years mean that while there are many opportunities, there is equally more potential competition.

Finally, as section 5.6 details, these sectors do not work in perfect and/or unregulated markets. Wider cultural, economic, and policy concerns affect the choices that businesses make, and therefore by extension, what technologies they choose to meet their business objectives.

\(^1\) [http://ftw.usatoday.com/2014/07/germany-brazil-most-tweeted-event-history-miley-cyrus](http://ftw.usatoday.com/2014/07/germany-brazil-most-tweeted-event-history-miley-cyrus)
## 5.8. Abbreviations and acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSS</td>
<td>Business Support Systems</td>
</tr>
<tr>
<td>CDR</td>
<td>Call Data Record</td>
</tr>
<tr>
<td>CRM</td>
<td>Customer Relationship Management</td>
</tr>
<tr>
<td>DWH</td>
<td>Data Warehouse</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>LTE</td>
<td>Long Term Evolution</td>
</tr>
<tr>
<td>NFV</td>
<td>Network Function Virtualisation</td>
</tr>
<tr>
<td>OSS</td>
<td>Operation Support Systems</td>
</tr>
<tr>
<td>OTT</td>
<td>Over the Top</td>
</tr>
<tr>
<td>PDN</td>
<td>Packet Data Network</td>
</tr>
<tr>
<td>P-GW</td>
<td>PDN Gateway</td>
</tr>
<tr>
<td>QoE</td>
<td>Quality of Experience</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>SDN</td>
<td>Software Defined Networks</td>
</tr>
<tr>
<td>SID</td>
<td>Shared Information/Data</td>
</tr>
<tr>
<td>SIP</td>
<td>Session Initiation Protocol</td>
</tr>
<tr>
<td>SS7</td>
<td>Signalling System nº 7</td>
</tr>
<tr>
<td>XDSL</td>
<td>Digital Subscriber Line (any technology)</td>
</tr>
<tr>
<td>XML</td>
<td>eXtensible Markup Language</td>
</tr>
</tbody>
</table>
5.9. References


Chris Bizer (2014): Pay-as-you-go Data Integration on the public Web of Linked Data


Techcrunck (2014):


WT Stead lecture by Emily Bell, British Library (2014):
http://www.bl.uk/whatson/podcasts/podcast161799.html


6. Energy & Transport Sectors

6.1. Introduction

During the course of analysing the energy and transportation requirements regarding big data technologies, and the technology roadmapping over the course of two years we have also had the opportunity to follow the transformation in mindset of the stakeholders. At the beginning some were still hung up on whether Big Data is a hype or not – some have had already acknowledged that the masses and variety of data in the industrial businesses is a consequence of digitization and automation along their value chains. The degree of digitization and automation in the sectors, as well as the progress of liberalization has made a clear difference on what end of the spectrum the stakeholders were.

Whilst transportation is more advanced in digitization and the participation of the always-connected end-user, the energy sector has a longer history of automation, and liberalisation just starts to change how electricity is being generated, delivered and consumed. These two sectors are very close, and also close to telecommunication, since they deliver the resource and infrastructure backbone for our society regarding energy, mobility, and information. Technically they are comparable with respect to big data because they can be modelled and represented as multimodal flow networks, intelligent infrastructures with potential sources of big data but also with storage and computing capabilities installed along them to produce smart data. Many of the efficient data structures and algorithms to tackle big data computing and analytics will be usable for both sectors, at the same time these two sectors will have many cross-optimization points.

We have distilled the many different use cases and specific requirements onto bigger applications scenarios. From those application scenarios we have selected major requirement classes that currently prevent the energy and mobility sectors from moving ahead, to requirement classes that need to be fulfilled such that full digitization and automation are feasible. The technology roadmap presented here is a starting point for the stakeholders of energy and transportation sectors who want to move ahead and generate value from the inevitable data tsunami which will culminate through digitization. Along the process, especially during the workshops with stakeholders, we have identified a multitude of non-technical requirements. These are not considered in the technology roadmap, but will be input into the EU BYTE project, which aims at analysing the technological paradigm shifts including the socio-economic and legal perspectives to refine the research roadmap as well as define and communicate a sound policy framework for the European Digital Single Market. In EU BYTE, both energy and transportation case studies are included as well as the smart city case study where both energy and transportation scenarios intertwine.

Gartner has published a recent survey of more than 2800 CIOs across the different sectors, and it is evident that:

"This isn't just a high-tech story, a U.S. story, a private-sector story, a large-company story or a startup story. Digitalization is transforming all types of companies and public sector agencies. More often than not, these transformations represent both massive opportunities and substantial challenges. Digitalization is not only a way to gain a competitive edge, but also provides a powerful ability to flip disadvantages into advantages."

So whilst the technology roadmap presented here can only be a starting point to remedy the symptoms, i.e. how to get from big data to smart data and generate value along the process, the stakeholders are advised that they need to take into account the cultural transformation that needs to take place in their organizations and in the organizations or mindsets of their customers.
The presentation of technology roadmap for energy and transportation is as follows: In section 6.2 we recap the application scenario classes of **Operational Efficiency**, **Customer Experience**, and **New Business Models**.

Section 6.3 gives a short recap of the energy and transportation sector requirements regarding their quantification along the three dimensions of big data, volume, velocity, and variety.

Section 6.4 maps those requirements along the data value chain of data acquisition, data management, data analytics, and data usage.

Finally, we develop the technology roadmap in Section 6.5: First, we selected major requirements based on their importance for (a) overcoming fundamental barriers to make data available for digital value creation and (b) for creating the data value chain in energy and transportation sectors that will assure scalability of technology as well as their business benefit lifecycles. Together with the technical working groups we then selected major technologies to fulfil these requirements, and assessed the open research questions associated with utilizing these technologies on energy and mobility data. The final step of roadmapping, defining a timeline, is done based on expert opinion; and as discussed before will be refined through taking into account the accompanying social, economic, legal, and policy roadmaps and development – which certainly has a huge impact on data value in the European sectors of energy and transportation.

### 6.2. Application scenarios

In the pursuit of collecting the many sectorial requirements towards a European Big Data economy and its technology roadmap, Big Data applications have been analysed\(^1\). A finding that is also congruent with Gartner’s study on the advancement of analytics (Kart, 2013) (Figure 18) is that big data applications can be categorized in “Operational Efficiency,” “Customer Experience,” and “New Business Models.”

<table>
<thead>
<tr>
<th></th>
<th>Mfg</th>
<th>Education</th>
<th>Banking</th>
<th>Insurance</th>
<th>Government</th>
<th>Energy &amp; Utilities</th>
<th>Comm, Media &amp; Svs</th>
<th>Retail</th>
<th>Transportation</th>
<th>Healthcare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process efficiency/cost reduction</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Enhanced customer experience</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Improved customer service</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Identifying new product/market opportunities</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

*Figure 18: Excerpt from Gartner’s study on business problems to solve with big data analytics (1= top prio). Heat map also indicates that the highest priority business applications powered by big data analytics can be categorized into Operational Efficiency, Customer Experience, New Business Models*

New data sources come into play mainly through increasing levels of digitization and automation. **Operational efficiency** is the main driver (Kart, 2013) behind the investments for

\(^1\) [http://big-project.eu/sites/default/files/BIG_D2_3_2.pdf](http://big-project.eu/sites/default/files/BIG_D2_3_2.pdf)
digitization and automation. The need for operational efficiency is manifold, such as revenue margins, regulatory obligations, or the retiring skilled workers.

Once pilots of big data technologies are setup to analyse the masses of data for operational efficiency purposes, industrial businesses realise that they are building a digital map of their businesses, products, and infrastructures — and that these maps combined with a variety of data sources also deliver insight to asset conditions, end usage patterns etc. Customer experience becomes the next target of value generation — just because it is now in their reach to find out more about how their products, services or infrastructure is being utilized and where frustrations might be remedied, how customers can even be excited. Stakeholders in customer-facing segments of the two sectors may start with customer experience related big data projects right away, but soon discover the close interactions with operational efficiency related big data scenarios.

Literally playing around with data makes creative. Some analytical insights reveal that systems are used in ways that they were not planned for, or that there is value in some niche, which previously was unknown, or that there is value in own data for some other organization possibly in another sector. Hence, new business models are the third pillar of big data application. Departments such as strategy or business development in the companies also benefit from this data-driven creativity.

In the following is a recap of the scenarios with the available or potential big data sources and the key challenge that mainly prevents this scenario from uptake in Europe. The distilled requirements are to be found in D2.3.2.

Global spending on power utility data analytics is forecast to top $20 billion over the next nine years, with an annual spend of $3.8 billion globally by 2020 (Leads, 2012). Internationally, there is also movement in the real-time monitoring & control of distribution and transmission grids. GTM Research found that annual rate of data intake in terabytes will be high through automation of the electricity distribution system and highest through the installation of phasor measurement unit technology that enhance the monitoring, protection, and control of power transmission grids.

6.2.1 Operational Efficiency

Operational efficiency subsumes all use cases that involve improvements in maintenance and operations in real-time or a predictive manner based on the data which comes from infrastructure, stations, assets, and end users of energy. Technology vendors, who develop the sensorisation, i.e. digitisation and automation, of the infrastructure are the main enablers. The market demand for such enhanced technologies is increasing, because it also helps the businesses in energy to better manage risk: The complexity pan-European interconnected electricity markets with the integration of renewables and liberalisation of electricity trading requires more visibility of the underlying system and of the energy flows in (near) real-time.

As a rule of thumb – anything with the adjective “smart” falls into this category: smart grid, smart metering; smart cities, smart (oil, gas) fields. Some examples of big data use cases in operational efficiency are:

a) Predictive and real-time analysis of disturbances in power systems and cost-effective countermeasures;

b) Operational capacity planning, monitoring and control systems for energy supply and networks, potentially with dynamic pricing or

1 [http://dqbasmyouzl2.cloudfront.net/content/images/articles/BigData_AmiToSmartHome.jpg](http://dqbasmyouzl2.cloudfront.net/content/images/articles/BigData_AmiToSmartHome.jpg)
c) Optimizing multimodal networks in energy as well as transportation especially in urban settings, such as city logistics or eCar-sharing for which the energy consumption and feed-in in the transportation hubs could be cross-optimised with the logistics.

All of the scenarios in this category have the main **big challenge of breaking up silos**: be it across departments within vertically integrated companies or across organisations along the value chain of electrification, which now are liberalised and oftentimes have conflicting objectives. The big data use cases in the operational efficiency scenario require seamless integration of data, communication and analytics across a variety of data sources, which are owned by different stakeholders.

### 6.2.2 Customer Experience

Understanding big data opportunities regarding customer needs and wants is especially interesting for companies in liberalised, consumerised markets such as electricity, where entry barriers for new players as well as the margins are decreasing. Customer loyalty and continuous service improvement is what enables energy players to grow in these markets. There are, however, also B2B use cases mainly from technology vendors, but also from energy data or mobility data start-ups.

Some examples of using big data to improve customer experience are:

a) Continuous service improvement and product innovation, e.g. individualized tariff offerings based on detailed customer segmentation using smart meter or device level consumption or feed-in data;

b) Predictive lifecycle management of assets, i.e. data from machines & devices in energy combined with enterprise resource planning and engineering data to offer services such as intelligent on-demand spare-parts logistics, for example; or

c) Industrial demand-side management, which allows for energy efficient production and hence increases competitiveness of manufacturing businesses.

**Big challenge is handling confidentiality and privacy of private and business customers** whilst getting to know and anticipate their needs. Data originator, data owner, and data users are different stakeholders, who need to collaborate and share data to realize these application scenarios.

### 6.2.3 New Business Models

New Business Models revolve around monetising available data sources and existing data services in entirely different ways. There are quite a few cases in which data sources or analysis results from one sector represent insights for stakeholders of another sector. The analysis of energy and mobility data start-ups shows that there is a whole new way of generating business value if the resources are owned by the end user. Then the business is entirely customer- and service-oriented; whereas the infrastructures of energy and transportation with their existing stakeholders are utilized as part of the service. These are called intermediary business models.
Energy consumer segment profiles, such as prosumer profiles for PVs, CHPs; or actively managed demand side profile, etc., from metering service providers could also be offered for smaller energy retailers, network operators or utilities who can benefit from improvements on the standard profiles of energy usage but do not otherwise have access to high resolution energy data of their own customers yet.

The big challenge is the unclear regulation around the secondary use of energy and mobility data. Additionally, one of the other scenarios needs to be realised ensuring the digitisation of most parts of the infrastructure and the sourcing of end usage data. The connected end user is the minimal prerequisite for the consumer focused new business models.

A disturbing current market development is that Europe's ten largest power companies are set to lose billions of Euros after over-investing in fossil fuels rather than renewable energy: For example, E.ON's earnings from power generation dropped by two thirds in the first nine months of 2013 compared with the same period in 2012. The Greenpeace publication claims that leading energy companies, including E.ON, RWE, Iberdrola, and EDF, have failed to deal with the challenges of slowing electricity demand, rapidly falling Renewables costs, increased competition from people producing their own energy (prosumers) via solar panels or small wind turbines, and air quality legislation that will force the closure of dozens of coal-fired plants.

On the upside: the market in this new segment is getting diversified through big data spirited energy start-ups like Next Kraftwerke, who "merge data from various sources such as operational data from our virtual power plant, current weather and grid data as well as live market data. This gives Next Kraftwerke an edge over conventional power traders."

Cars are parked 95% of the time. Hence many actors in the electricity, transportation and automobile industry are trying to establish new business models around this technological opportunity (Barter, 2013).

Considerable is the business move towards of car-sharing from incumbents such as Daimler and BMW. According to a recent study, one car-sharing vehicle displaces 32 new vehicle purchases (AlixPartners, 2014) – and business that previously revolved around the product, now becomes all about data-driven services. On the contrary to the energy sector, this bold moves show the readiness of the incumbents to seize the big data value potential of a data-driven business.

New players in multimodal transportation are like Israeli startup Waze and US startup Moovit, who closed a $28 million round for its app that combines public transport data with input from the community, to give commuters a real-time perspective of what their trip will be like and suggest the best routes.

6.3. Sector Requirements

In the past years we have interviewed stakeholders and researched trends regarding big data in Energy and Transportation sectors. It is safe to say that whilst at the beginning some were hang up on whether Big Data is a hype or not – today, a considerable part of the stakeholders have acknowledged that the masses and variety of data in the industrial businesses is a consequence of digitization and automation along their value chains. In the following we recap the three dimensions of Big Data in Energy and Transportation with some numbers and the

2 http://www.next-kraftwerke.com/
3 Car2go
4 DriveNow
5 acquired by Google for $1.3 billion
resulting needs of business and end users with respect to big data technologies and their usage:

**Volume** of available data from digitisation and automation along the infrastructure as well as at the supply and demand side is steadily increasing: e.g. Phasor measurement units (PMUs) deliver up to 100 Hz resolution GPS-synchronized data on grid parameters – a current SCADA system with traditional RTUs, which poll data every 2-4 seconds, has to process less than one per cent of the data volume of the entire PMU data accumulating in the same area (NERC, 2010). Smart metering results in 3,000-fold increase in data when 15 minutes intervals are measured (Picard, 2013). The software deployment for the smart infrastructure additionally generates data in forms of error logs, etc.

**Velocity** of data, i.e. the rate at which data is produced and consumed is also increasing (Heyde, 2010), since the balancing of energy supply and demand as well as the provision of capacity becomes increasingly more real-time due to the added uncertainties through Renewables, decentralization, and active or mobile consumption.

**Variety** of data arises due to the diversification of energy consumption and generation as well as consumerisation and liberalisation. Market communication alone results in the exchange of a variety of data also coming from new source and roles. With the always connected end user, online sources such as social media as well as smart phones add to this inherent variety in directly and indirectly-related data sources (GTM Research, 2012).

- **In electricity industry**: Data is coming from digitalized generator substations, transformer substations, and local distribution substations in an electric grid infrastructure of which the ownership has been unbundled; information can come in the form of service and maintenance reports from field crews about regular and unexpected repairs, or health sensor data from self-monitoring assets: end usage and power feed-in data from smart meters, high resolution real-time data from GPS-synchronized phasor measurement units or intelligent protection and relay devices.

  Example use case EDF (Picard, 2013): Currently, most utilities do a standard meter read once a month. With smart meters, utilities have to process data at 15-minute intervals. This is about a 3,000-fold increase in daily data processing for a utility, and it’s just the first wave of the data deluge. Data: individual load curves, weather data, contractual information, network data 1 measure every 10 min for a target of 35 million customers (currently only 300,000)- Annual data volume: 1,800 billion records, 120 TB of raw data. The second wave will include granular data from smart appliances, electric vehicles and other metering points throughout the grid. That will exponentially increase the amount of data being generated.

  Example use case Power Grid Corporation of India (Power Grid Corporation India, 2012): Unified Real-time Dynamic State Measurement with a target integration of around 2,000 PMUs (starting with 9 PMUs). The PMUs provide time-stamped measurements of local grid frequency, voltage, and current at 25 samples per second (up-to 100 Hz possible). This amounts to about 1 TB of disk space per day. Traditional SCADA systems poll data from RTUs every 2-4 seconds. For comparison: A traditional SCADA system for wide area power system management has 50 times more data points than a system that locally processes high-resolution data received from PMUs. Yet, the SCADA system has to process less than 1% of the data volume of the entire PMU data accumulating in the same area (NERC, 2010).

- **In oil & gas industry**: Data comes from digitalized storage and distribution stations, but also wells, refineries, and filling stations are becoming data sources in the intelligent infrastructure of an integrated oil & gas company. Finer resolution seismic data, data from actual drilling and logging activity; down hole sensors from production sites deliver data on a real-time bases including pressure, temperature and vibration gauges, flow meters, acoustic and electromagnetic, circulation solids; data coming from sources such as vendors, and tracking service crews, truck traffic, equipment and hydraulic fracturing.
water usage; SCADA data from Valve events and Pump events, Asset operating parameters, Out of condition alarms; Unstructured reserves data, geospatial data, safety incident notes, surveillance video streams.

Example Use Case Shell (Mearian, 2012): optical fibre attached down hole sensors generate massive amounts of data that is stored at a private isolated section of the Amazon Web Services. They have collected 46 petabytes of data and the first test they did in one oil well resulted in 1 petabyte of information. Knowing that they want to deploy those sensors to approximately 10,000 oil wells, we are talking about 10 Exabyte’s of data, or 10 days of all data being created on the internet. Because of these huge datasets, Shell started piloting with Hadoop in the Amazon Virtual Private Cloud.

Others (Nicholson, 2012): Chevron proof-of-concept using Hadoop for seismic data processing; Cloudera Seismic Hadoop project combining Seismic Unix with Apache Hadoop; PointCross Seismic Data Server and Drilling Data Server using Hadoop and NoSQL; University of Stavanger data acquisition performance study using Hadoop.

- **In transportation:** data comes from digitalized (air/sea) ports, train/bus stations in transportation, logistics hubs and warehouses; employed Electronic On Board Recorders (EOBRs) in trucks delivering data on load/unload times, travel times, driver hours; truck driver logs, pallet or trailer tags delivering data on transit and dwell times; information on port strikes; public transport timetables, fare systems and smart cards, rider surveys, GPS updates from bus/truck/car fleet, higher volumes of more traditional data from established sources such as frequent flier programs, etc.

Example use case City of Dublin (Tabbitt, 2014): road and traffic department is now able to combine Big Data streaming in from an array of sources - bus timetables, inductive-loop traffic detectors, and closed-circuit television cameras, GPS updates that each of the city’s 1,000 buses transmits every 20 seconds - and build a digital map of the city overlaid with the real-time positions of Dublin’s buses using stream computing and geospatial data. Some interventions have led to a 10-15 per cent reduction in journey times.

**Business user needs** can be derived from the above categories of business needs:

- **Ease of use:** Big data technologies employ new paradigms and mostly offer programmatic access only, e.g. R the statistical programming language or MapReduce, a framework for programming massively parallel task execution. Without software development skills and a deep understanding of distributed computing paradigm as well as application of data analytics algorithms within such distributed environments, scalable analytics of mass, streaming data is currently not possible.

- **Semantics** of correlations and anomalies that can be discovered and visualised via big data analytics need to be made accessible. Currently only domain and data experts together can interpret the data outliers, business users are often left with guesswork when looking at data analytics results.

- **Veracity** of data needs to be guaranteed before it is used in smart grid applications, such as distribution automation or billing of variable tariffs. Because the increase in data that will be used for these applications will be magnitudes bigger, simple rules or manual plausibility checks no longer are applicable.
Smart Data often is used by industrial stakeholders emphasizing that an industrial business user needs refined data—not necessarily all raw data (big data)—but without losing information by concentrating only on small data that is of relevance today. In cyber-physical system as opposed to online businesses, there is information and communication technology (ICT) embedded in the entire system instead of only in the enterprise IT backend. Especially infrastructure operators have the opportunity to pre-process data in the field and aggregate levels or distribute the intelligence for data analytics along the entire installed ICT infrastructure to make best use of computing and communication resources to deal with volume and velocity of mass sensor data. The need for smart data is driven by the need for cost-effective usage of all installed ICT resources along the infrastructures of energy and transportation.

Decision support or automation becomes a core need, as the pace and structure of business is changing. European grid operators today need to intervene almost daily to prevent potentially large-scale blackouts, e.g. due to integration of Renewables or liberalized markets. Business users need more than the information that something is wrong. Visualisations can be extremely useful, but the question of what needs to be done remains to be answered either in real-time or in advance of an event, i.e. in a predictive manner.

Strong need for scalable, advanced analytics, hence, will push the envelope of state-of-the-art, such as smart metering data analytics (Picard, 2013): e.g. Segmentation based on load curves, forecasting on local areas, scoring for non-technical losses, pattern recognition within load curves, predictive modeling, etc. or real-time analytics as fast and as reliable in order to control such delicate and complex systems as electricity (Heyde, 2010).

End users needs are not directed towards big data neither towards energy or mobility data as is. However, big data scenarios in these sectors offer many improvements for the end users: Operational Efficiency ultimately means energy and resource efficiency, which will improve quality of life—especially in urban settings—for all end users. Customer Experience and New Business Models related big data scenarios are entirely based on better serving the end user of energy. However, both scenarios need personalised data in higher resolution, geo- or time-tagged. We have been discussing the real big data value in cross-combining such variety in data, which on the downside can make pseudonymisation or even anonymisation ineffective in protecting the identity and behavioral patterns of individuals or business patterns and strategies of companies. New business models based monetising once collected data, with currently unclear regulations, leaves end users entirely transparent, uninformed, and unprotected against secondary use of their data for purposes they might not agree with, e.g. insurance classification, credit rating, etc.

Reverse transparency is at the top of the wish list of data-literate end users: Not end users should become transparent through data analytics, but data analytics ought to empower end users to grasp the usage of their data trails, personal or business consequences, and the efficient dynamic configurability of data access and usage. End users need practical access to information on what data is used by whom for what purpose in an easy-to-use, manageable way. Rules and regulations are needed for granting such transparency for end users; data protection laws that are minimally consistent, but also allow individualization for maximal privacy are required.

Data access, exchange, and sharing needs, hence, apply on both business and end users. In today’s complex electricity markets, there is almost no scenario where all required data for answering a business or engineering question comes from one department’s databases. Nonetheless, most of the currently installed advanced metering infrastructures, for example, have a lock-in of the acquired energy usage data in the utilities’ billing systems. The lock-in makes it cumbersome to use the energy data for other valuable analytics. Such silos
have traditional roots from the recent era where most of European utilities were vertically integrated companies. Also, the amount of data to be exchanged was much less, so that interfaces, protocols, and processes for data exchange have been rather rudimentary.

6.4. Mapping Requirements to Technology

The big data refinery pipeline for infrastructure- and resource-centric systems of energy and mobility businesses consists of three main phases: data acquisition, data management, and data usage. Data analytics, as indicated by business user needs, is implicitly required within all steps, and is not a separate phase:

- **During data acquisition**, analytics supports the data cleansing and ensures data quality. Privacy and confidentiality requirements can be assured through analytics methods already at acquisition time. Sensor data, which is the data source type that mostly defines the volume and velocity dimensions of big data in energy and transportation (see Section 6.3), is sparse data.

- For sparse data, analytics can enable efficient event-specific data compression and increase data quality through anomaly detection at acquisition time. This is important because at big data scale, it is manual data curation processes are not feasible.

- For data usage, business and engineering questions need to be formulated by business users and translated into suitable models. These models among other aspects imply which types of mass storage and data handling are cost-efficient for the specific purpose.

So, each step of the pipeline is setup to refine the data, but the methods of refinement vary depending on the data type.

Currently deployed big data technologies mainly focus on scalable storage, batch and ad-hoc querying capabilities for petabyte scale data, SQL-like querying capabilities and increasingly so near real-time analytic processing capabilities. Below is a short reference list – a thorough discussion is available in the “D2.2.2 Big Data Technical Working Group Whitepaper.”

- Batch processing: Google File System & MapReduce, Apache Hadoop HDFS & MapReduce
- Ad-hoc processing: Amazon Dynamo, Google BigTable, Apache HBase, Apache Cassandra
- SQL-like access: Apache Hive, Pig, PrestoDB, H-Store, Google Spanner,
- Stream processing: Hadoop Streaming, Google BigQuery, Google Dremel, Apache Drill, Apache Flume/HBase, Apache Kafka/Storm, Samza
- Real-time analytical processing: Apache Spark, Amazon Kinesis, Google Dataflow

These technologies have been pioneered and proven by big data natives like Google (Search), Amazon (eCommerce), and Facebook (Social), are also available as open source, and have been well adopted by IT vendors. Hence, many requirements of the sectors can be satisfied by

---

1 http://big-project.eu/sites/default/files/BIG_D2_2_2.pdf
state of the art cost-efficient handling and value generation from massive amounts of data coming at high speed.

However, skills to adapt and apply these state of the art technologies in big data computing for the traditional businesses of energy and transportation sectors are scarce. Thus, the adoption of these “new” technologies requires ramp up time.

6.4.1 Data Acquisition Requirements

In the data acquisition phase, data should be checked for quality features, such as missing data or implausible data, e.g. in time series. Depending on whether data is generated during an automated process, or by a human, e.g. repair crew in the field or customer call agent, data is multi-structured, i.e., highly structured or unstructured and of very different volumes and speed, i.e. it is acquired as continuous streams of high frequency samples or sent irregularly (see Section 6.3).

Data acquisition in the energy and transportation scenarios is concerned with how to ingest data from installed data sources owned by the infrastructure providers, by connected smart devices of end users, or any other source potentially including information about usage, such as twitter messages of end users. Many data sources and data needs to be shared with the stakeholders in order to realise the scenarios. Data, acquired in very different ways and various formats, needs to be accessible and interpretable across applications, departments, or increasingly more so across organizations.

The great differentiator of big data scenarios in traditional sectors such as energy and transportation, especially in Europe, is the concern that data should be collected for a known purpose. This may sound conservative at first, and actually is orthogonal to the big data mentality. And with current technological capabilities, it is very restrictive in practice. However, it actually gives way to formulate a central requirement that will drive Europe to thrive both technologically as well as in terms of digital rights on individual and business level:

**Dynamic configurability of data access** about which data can be collected for what purpose in what granularity and time span and location must be given along the entire intelligent infrastructure of electrification. The configuration or the effects thereof must be easily comprehensible for the data owners. The configuration must be dynamic in the sense that service providers are able to receive the data in the required granularity and with the allowed privacy and confidentiality protection settings defined – on demand.

Privacy and confidentiality preserving data analytics at the sources of data are required to enable the service provider to retrieve the knowledge without violating the agreed upon granularity of data or the allowed privacy or confidentiality settings.

6.4.2 Data Management Requirements

Different types of data management may be considered for the same data source type depending on for what purpose it will be used: Graph databases can yield faster results for geospatial analysis, key-value stores may prove most efficient for feature discovery purposes in time series. Regarding the generation of value from a variety of data sources the data storage step needs to enable the cost-efficient integration of the different data sources in an extensible
Flexible abstractions, multiple layers of abstractions for data and analytical models will be required for the data ingestion and storage step.

Lightweight semantic data models that represent the multiple links within various data sources, such as Linked Data, may provide such cost-efficient abstractions, especially when the data domain is complex and highly interlinked. For high-dimensional data, multi-dimensional data structures such as tensors and space-filling curves may be required to support the design of more efficient analytics algorithms, which are capable of exploring multiple relations at once. These are only some examples of a few methods from machine learning, semantic data technologies, information extraction, and distributed computing. What will be required is a flexible toolbox, an analytics engine, which scales with the increasing business need to gain more value from ever increasing amounts and variety of data without having to change the entire analytics application:

Abstraction from the actual big data infrastructure is required to enable (a) ease of use and (b) extensibility and flexibility. The analysed use cases have such diverse requirements that the future energy and mobility businesses will require to flexibly bundle technology stacks into solutions to gain insights from massive amounts of available data. Adaptive data and system models are needed so that new knowledge extracted from domain analytics or the ever changing circumstances of the system can be redeployed into the data analytics framework without disrupting the analytics applications.

6.4.3 Data Analytics Requirements

As stated in the beginning of this Section, data analytics is part of every stage in the data refinery pipeline. State of the art technologies for analysing big data at scale are there and see high adoption rates in web-related sectors, however, have not found adoption in energy and transportation sectors. All of the use cases referenced in Section 6.3 are concerned with scalable storage and access to big data. Below is a short reference list of state of the art for scalable big data analytics - a thorough discussion is available in the “D2.2.2 Big Data Technical Working Group Whitepaper.”

- Historical and online analytical processing: Apache Mahout
- Real-time and in-stream analytical processing: Apache Spark, Amazon Kinesis, Google Dataflow

Historical and online analytical processing of big data will be adopted eventually, since 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. Additionally, real-time insights are of negligible value, if the stakeholders, e.g. operators, have no options to take actions based on that insight in real-time, too. If their only option is to implement improvements for the future, then state of the art capabilities for historical analyses or online analyses, and predictive analytics are sufficient. The Gartner Hype Cycle for Emerging Technologies in 2013 found that predictive analytics is reaching plateau of productivity within less than 2 years.

---

1. [http://big-project.eu/sites/default/files/BIG_D2_2_2.pdf](http://big-project.eu/sites/default/files/BIG_D2_2_2.pdf)
6.4.4 Data Usage Requirements

Although data usage is the last phase in the refinement process, the business and engineering questions formulated in this phase are the starting point for choosing the technology stack. This phase represents the reason for setting up a particular connection through the pipeline to the required data sources. This connection motivated by the usage scenarios relies on the above formulated requirements for data acquisition, management, and analytics to be fulfilled.

In the industrial domains of energy and transportation, analytics for data usage, reaches from descriptive analytics such as dashboards, to predictive and prescriptive analytics such as forecasting for portfolio management or simulation of interconnected power systems, intermodal transportation systems, and extraction of operational intelligence on what actions to take in real-time or predictive manner. Prescriptive analytics, i.e. gaining analytical insights from mass data in order to propose best options to take action, will play a significant role in industrial automation. The Gartner Hype Cycle for Emerging Technologies in 2014\(^1\) finds that **prescriptive analytics will be the innovation trigger in the next years**, whilst Big Data is moving towards the trough of disillusionment.

Thus, industrial data usage requires considerably more precision than data usage in online data businesses – **veracity is a key requirement, the acquired data and the analytics output of knowledge and proposed actions need to be verifiably reliable.** At anytime the results must be traceable.

**Scalable access and data interpretability** must be assured without the constant involvement of domain experts. Hence, expert and domain know-how must be blended into the data management and analytics – otherwise usage will not scale along the 3V dimensions. Self-describing data sources, such as smart devices, are one part of the requirement. Semantic and adaptive data models are the other part.

6.5. Technology Roadmap

The technology roadmap for fulfilling the key requirements along the data value chain for the energy and transportation sectors focuses on not readily available big data technology: i.e., big data technology that needs further research and development in order to fulfill the more strict requirements of industrial applications, or current gaps in big data technology, all of which prevent these innovative and cost-efficient ways of processing big data for insights and eventually for decision automation to be adopted by industrial players from energy and transportation sectors.

Out of scope from D2.3.2 Sectors’ Requisites are the non-technical requirements such as “**new skills**” and requirements which are fulfilled, i.e. at least being prototyped by energy and mobility stakeholders, using current big data technologies. Requirements in this category are, e.g. “**storage of 15-minutes interval data from smart meters with an accumulated annual data volume: 1800 billion records, 120 TB of raw data**” (see Section 6.3). Of course, current big data technologies are enablers of this technology roadmap – and the fulfilling of non-technical requirements are often times prerequisites or part of a framework that will support the technology roadmap presented here. Especially the latter point, which is out of scope in this technology roadmap, will be further examined in EU BYTE\(^2\), the project focusing on socio-economic aspects of the big data paradigm shift and aiming to produce a roadmap that helps

---

2. [http://byte-project.eu/](http://byte-project.eu/)
minimize the associated risks and maximize the opportunities of big data scenarios across the European sectors.

In the following, we give a short recap of the key requirement and its importance, the required technologies and open research questions in scaling these technologies for energy and mobility data analytics applications (also see Figure 19). For a fundamental discussion of the technology characteristics and state of the art please refer to “D2.2.2 Big Data Technical Working Group Whitepaper”¹ and the state-of-the art collection of “Big Data Technologies and Infrastructures” in D1.4 of EU BYTE².

**Technology Roadmap**

**Energy & Transport Sector**

<table>
<thead>
<tr>
<th>Technical Requirement</th>
<th>Linked Data</th>
<th>Automating data linkage, Scalability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Access &amp; Sharing</td>
<td>Linked Data</td>
<td>Automating data linkage, Scalability</td>
</tr>
<tr>
<td>• Interpretability of data &amp; scalability of access</td>
<td>Encrypted Storage</td>
<td>Analytics on encrypted data</td>
</tr>
<tr>
<td>• Privacy &amp; Confidentiality</td>
<td>Data Provenance, Differential Privacy</td>
<td>Efficient computing, Scalability</td>
</tr>
<tr>
<td>Real-time Analytics</td>
<td>Machine Learning (ML)</td>
<td>Deep learning, such as Convolutional neural networks, Tensor modeling</td>
</tr>
<tr>
<td>• spatio-temporal, high-dimensional, high-speed</td>
<td>Distributed stream computing</td>
<td>Infrastructure awareness, Advanced elasticity, Federation</td>
</tr>
<tr>
<td>Prescriptive Analytics</td>
<td>Prediction &amp; Optimization</td>
<td>Efficient computing, Scalability</td>
</tr>
<tr>
<td>• In support of usage for automation</td>
<td>Embedded analytics</td>
<td>Distributed Data Analytics</td>
</tr>
<tr>
<td></td>
<td>Linked Data &amp; ML</td>
<td>Large-scale reasoning &amp; ML in real-time</td>
</tr>
<tr>
<td>Abstraction</td>
<td>Linked Data</td>
<td>Scalable knowledge modeling and retrieval</td>
</tr>
<tr>
<td>• Ease-of-use</td>
<td>Data abstraction layers</td>
<td>Data type agnostic architectures</td>
</tr>
<tr>
<td>• Flexibility &amp; extensibility</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 19: Roadmap for Energy & Transportation*

**Data Access & Sharing:**

_Enable scalable data access and sharing, whilst enhancing interpretability of data as well as preserving privacy and confidentiality:_

Priority rationale of this requirement:

Energy and mobility are resource-centric infrastructure businesses. Access to usage data creates the opportunity to analyze usage of a product or service and improve it or gain efficiency in sales and operations. This usage data however, mostly needs to be cross-combined with other available data to deliver reliable predictive models.

¹ [http://big-project.eu/sites/default/files/BIG_D2_2_2.pdf](http://big-project.eu/sites/default/files/BIG_D2_2_2.pdf)
Currently there is a trade-off between enhancing interpretability of data and preserving privacy & confidentiality, which the following example with mobility data usage with cross-combination of variety of data demonstrates:

(de Montjoye et al, 2014) show that 4 spatio-temporal points, approximate places and times, are enough to uniquely identify 95% of 1.5M people in a mobility database. The study further states that these constraints hold even when the resolution of the dataset is low, mobility datasets and metadata circumvent anonymity.

At the same time, such insufficient privacy protection options can hinder the sourcing of big data in the first place, as experiences from smart metering roll outs in the energy businesses show:

In the EU only 10 per cent of homes have smart meters (Nunez, 2012). Although there is a mandate that the technology reaches 80 per cent of homes by 2020, the European rollouts stagnate: A survey from 2012 (Department of Energy & Climate Change, 2012) finds that “with increasing reading frequency, i.e. from monthly to daily, to half hourly, etc, energy consumption data did start to feel more sensitive as the level of detail started to seem intrusive[ ...]Equally, it was not clear to some [participants] why anyone would want the higher level of detail, leaving a gap to be filled by speculation which resulted in some becoming more uneasy.”

Technologies & Open Research Questions:

Linked Data is a lightweight practice for exposing and connecting pieces of data, information, or knowledge using basic web standards. As such it promises to open up siloed data ownership and is already an enabler of open data and data sharing. However, with the increasing number of data sources already linked, and with the various types of new data that especially will come from intelligent infrastructures and always connected end users in energy and mobility scalability and cost-efficacy becomes an issue. One of the open research questions in this regard is how to (semi-)automatically extract data linkage as to increase current scalability issues.

Encrypted Data Storages can enable integrated, data-level security, which is the basis for big data analytics across organisational siloes. As cloud storage becomes commonplace for end users and commercial users alike better and more user friendly data protection becomes a differentiation factor (Tanner, 2014). Eventually, by helping to preserve privacy and confidentiality encrypted data storages are the most basic enablers of data sharing and shared analytics. However, analytics and encrypted data is still an ongoing research question. The most widely pursued research is called fully-homomorphic encryption. Homomorphic encryption theoretically allows operations to be carried out on the ciphertext. The result also is a ciphertext that when decrypted matches the result of operation on plaintext. Currently, however, there are only basic operations possible feasible.

Data Provenance is the art of tracking data through all transformations, analyses, and interpretations. Provenance assures that data which is used to create actionable insights are reliable. Data sets are reliable when the process used to create them is reproducible and testable. The metadata which is generated to realize provenance across the big variety of data sets from differing sources also increases interpretability of data, which in turn could improve automated information extraction. However, scaling data provenance across the dimensions of big data is an open research question.

Differential Privacy (Dwork, 2014) is the mathematically rigorous definition of privacy (and its loss) with the accompanying algorithms. The fundamental law of information recovery (Dwork, 2014) states that too many queries with too little error will expose the real information. The

---

2 https://github.com/shaih/HElib
The purpose of developing better algorithms is to push this event as far away as possible. This notion is very similar to the now mainstream realization that there is no unbreakable security, but that barriers if broken need to be fixed and improved. The cutting edge research on differential privacy also considers distributed databases and computations on data streams enabling linear scalability and real-time processing for privacy preserving analytics. Hence, this technique could be an enabler of privacy preserving analytics on big data, allowing big data to gain user acceptance in mobility and energy.

**Real-time & multi-dimensional analytics:**

*Enable real-time, multi-way analysis of streaming, spatio-temporal energy and mobility data:*

**Priority rationale of this requirement:**

Examples from dynamic complex cyber-physical systems such as power networks show that there is a clear business mandate: Global spending on power utility data analytics is forecast to top $20 billion over the next nine years, with an annual spend of $3.8 billion globally by 2020 (GTM Research, 2012). However cost-effectiveness of required technologies needs to be proven. Especially monitoring in real-time does not justify costs if actions cannot be undertaken in real-time:

*Phasor measurement technology, enabling high-resolution view of the current status of power networks in real-time is a technology that was invented 30 years ago. Possible applications have been researched for more than 10 years. However, initially there was no business need for it, because the power systems, well engineered and well structured, were predictable. With increased dynamics through market liberalization and integration of power generation technology from renewable but stochastic sources like wind and solar, real-time view of power networks becomes indispensible. However, businesses without the capability of communicating, storing and analyzing so much data (see Section 6.3 for the 3V quantification) in a cost-effective way in real-time are still struggling to reap value. At the same time, state-of-the-art in big data computing can be adapted in this domain now*.  

**Technologies & Open Research Questions:**

*Distributed Stream Computing* currently gains traction. However, there are two different strains of research and development of stream computing: (1) stream computing as in complex event processing (CEP), which has had main focus on analyzing data of high-variety and high-velocity and (2) distributed stream computing, which has been adopted by the big data natives, focusing on high-volume and high-velocity data processing. Complementing the missing 3rd dimension in both strains is the current research direction. We argue that distributed stream computing which already has linear scalability and real-time processing capabilities, will tackle high-variety data challenge with Linked Data. A further open question is then how to ease development and deployment for the algorithms that make use of distributed stream computing as well as other computing and storage solutions such as plain old data warehouses and RDBMS but also parallel computing frameworks through infrastructure awareness and federation. Additionally, since cost-effectiveness is the main enabler for big data value, advanced elasticity, i.e. computing and storage on demand as the algorithm requires across all these possibilities, must also be tackled.

---

**Machine learning** is an interdisciplinary technology of statistics and computer science, where the aim is that algorithm design allows systems to learn from data rather than follow explicitly programmed instructions, e.g. finding clusters of similar data versus predefinition of how to classify data, or finding patterns instead of defining what an event is. This is a fundamental capability needed when dealing with big data and dynamic systems, where a human could not possibly review all data (as they did up to recently with business intelligence, or limit the data so they could but miss out on insights that competitors get when letting machines learn from all the data), or where humans just lack the experience to define pattern, either because of retiring skilled workers or because the systems they knew are used in completely different ways (e.g. power networks after market liberalization) and increasingly more dynamic with complex network effects that human are just not capable to extract reliable cues from – but only in hindsight after post-mortem data analysis (which takes months or years when done by human data scientists). Deep learning, a research field which gains momentum concentrates on more complex, non-linear data models and multiple transformations of data. Some representations of data are better for answering a specific question then others, meaning multiple representations of the same data in different dimensions may be necessary to satisfy an entire application. Additionally, the complex algorithms are computing intensive. The reason nonetheless for this momentum is that big data storage & computing by utilizing distributed systems makes deep learning techniques feasible.

Traditional industries today are applying large-scale machine learning and other algorithms for the analysis of huge datasets, e.g. for speech recognition, computer vision, and recommender systems. Now the open questions are how to represent the specific energy and mobility data, possibly in multiple dimensions, for designing algorithms that learn the answers to specific questions from the energy and mobility domains better than human operators of these systems can – and verifying this. The main requisites for machine learning are massive amounts of high-sampling data, and cost-effective storage and computing for the design of new efficient data structures and algorithms such as tensor modelling and convolutional neural networks.

**Prescriptive analytics:**

*Enable prescriptive analytics (in real-time) for decision automation in energy and mobility systems:*

**Priority rationale of this requirement:**

As stated also in the example for motivating real-time analytics, dynamic systems – and the increasing dynamicity through liberalization and consumerization – forces players in the energy and transportation domain to increase operational efficiencies by adapting on-demand either due to responding to critical system situations or in order to personalize and differentiate service offerings. In any case the more complex and dynamic the systems are becoming, the faster should insights from data be delivered to enhance decision making. With increasing information and communication technology installed into intelligent infrastructures of energy and transportation, decision automation becomes feasible – and desirable as following example demonstrates:

---

3 [http://deeplearning.net/tutorial/lenet.html](http://deeplearning.net/tutorial/lenet.html)
Modern aircraft are highly sophisticated machines. The Boeing 777(MH370) and Airbus A330 (AF447) are both exceptional aircraft that are used the world over. They provide safe travel to tens of thousands of people every month, without incident [...] However, things can go wrong. In the AF447 case, there were many emergency messages scrolling across the screens demanding pilot attention. The aircraft was in a violent storm and the postmortem analysis indicated the crew got confused, allowing the aircraft to enter an aerodynamic stall, failing to take the proper steps for recovery. Even with all the best technology on the flight deck, humans can still be overwhelmed.

This example of an overwhelmed human operator by a digitized system delivering all actionable information in case of an emergency can be applied in any control center for any human operator. Furthermore, with the increasing digitization, the normal operating state, when all digitized field devices deliver actionable information on how to operate more efficiently, will overwhelm human operators. The only logical conclusion then is either have dependable data structures and algorithms or ignore the insights per second that a human operator cannot handle at the cost of less operational efficiency. Of course the latter would just mean we do not need more digitization, i.e. data, or insight, if all operators are fine with the current state of the art. But thanks to regulations for efficiency assurance and thanks to market liberalization, i.e. competition, it is clear:

Now is the time to invest in, adapt, and improve current techniques to the data and questions of energy and mobility.

Technologies & Open Research Questions
Technologies enabling real-time analytics (see previous requirement) are the basis for prescriptive analytics in cyber-physical systems with resource-centric infrastructures such as energy and mobility. In addition to predictive models from machine learning, scalable optimization techniques are required. With prescriptive analytics the simple predictive model is enhanced with possible actions and their outcomes, as well as an evaluation of these outcomes. In this manner, prescriptive analytics not only explains what might happen, but also suggests optimal set of actions. Simulations and optimization are analytics tools that support prescriptive analytics. In cyber-physical systems, which are enhanced with real-time monitoring, model- and data-driven analytics is the essence of efficient prescriptive analytics, since the correlations that are discovered in the data can additionally be explained or verified and reliably acted upon according to the known (and digitized) system and engineering models.

Currently many system models are not machine readable. Engineering models on the other hand are semi-structured because digital tools are increasingly used to engineer a system. Research and innovation in this line of work will assure that machine learning algorithms can leverage system know-how that today mainly is limited to humans. Linked Data will facilitate the semantic coupling of know-how at design and implementation time, with discovered knowledge from data at operation time, resulting in self-improving data models and algorithms for machine learning.

As emphasized throughout the work on energy and mobility sectors’ requirements and in this document: intelligent infrastructures in these sectors have ICT capability built in, meaning, there is storage and computing power along the entire cyber-physical infrastructure of electricity and transportation systems, not only in the control rooms and data centers at enterprise level. Embedded analytics, and distributed data analytics, facilitating the in-network in-field analytics in concert with analytics carried out at enterprise level will be the innovation trigger in energy and mobility. Real-time analytics as also pointed out is a prerequisite.
**Abstraction:**

**Abstract from underlying big data technologies to enable ease of use for data scientists, as well as from big data analytics to enable ease of use for business users:**

**Priority rationale of this requirement:**

Many of the techniques required for real-time, prescriptive analytics, such as predictive modeling, optimization and simulation are data and compute intensive. Combined with big data these require distributed storage and parallel, or distributed computing. At the same time many of the machine learning and data mining algorithms are not straight forward to parallelize. A recent survey (Paradigm 4, 2014) found that:

> although 49 per cent of the respondent data scientist could not fit their data into relational databases anymore, only 48 per cent have had used Hadoop or Spark – and of those 76 per cent said they could not work effectively due to platform issues.

This is an indicator that big data computing is yet too complex to use without sophisticated computer science knowhow. Remember that machine learning is the interdisciplinary work of statistics and computer science. Computer science through big data computing is seeing a fundamental mainstreaming of research in distributed computing fields, such as peer-to-peer and grid computing. This leap is missing in machine learning, or today can only be made use of the big data natives. One direction of advancement is that abstraction and high-level procedures will be developed hiding the complexities of distributed computing from data scientists. The other direction of course will be: more skilled data scientists who are literate in distributed computing or distributed computing experts becoming more literate in data science and statistics. In any case, it is safe to assume that big data computing is transforming data analytics and machine learning as well. During that transformation an abstraction layer for analytics, such as an analytics engine could provide, will be of value for all stakeholders.

**Technologies & Open Research Questions:**

Abstraction is a common tool in computer science. Each technology at first is cumbersome. Abstraction factors out patterns such that the user (e.g. programmer, data scientist, or business user) can work closer to the level human problem solving, leaving out the practical details of realization. In the evolution of big data technologies there has been several abstractions already allowing using distributed file systems like databases by extracting SQL-like querying languages, or by adapting the style of processing to that of more familiar online analytical processing frameworks.

Linked Data is one state-of-the art enabler for realizing an abstraction level, which may be called “analytics engine” that assures **multiple data and analytics access layers** are in place but the insights are still connected. The semantic linkage of data with a priori knowledge and continuously with discovered knowledge is what will allow **scalable knowledge modelling and retrieval** in a big data setting. On further open questions is how to manage variety of data sources in a scalable way. Future research will establish a thorough understanding of data type agnostic architectures.

---

6.5.1 Preliminary Timeline for Technology Roadmap

In contrast to technology roadmap developments accomplished in the context of a single company, our approach covers the development of a technology roadmap for the European market. As a consequence, it is not possible to come up with a precise timeline of technology milestones and specific products, as the speed of technology development and its adoption on the market relies a) on the degree to which the identified non-technical requirements will be addressed and b) on the extent to which European organizations are willing to invest in big data developments and use case implementations. Table 7 depicts the estimated timeline it may take until the different research challenges in the big data technology adaptation for the energy and mobility domains are solved by indicating expected outcomes (for which the need have been described in the sections 6.3 and 6.4) that build on top of each other.

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Access &amp; Sharing</strong></td>
<td><strong>Real-time Analytics</strong></td>
<td><strong>Prescriptive Analytics</strong></td>
<td></td>
</tr>
<tr>
<td>B2C: Full digitalization and access policy clarification of potential big data (e.g. smart meter data)</td>
<td>Toolbox based on state of the art for satisfying the different notions of real-time in cyber-physical-infrastructure value chains</td>
<td>(Semi-) automated semantic linkage of knowledge and discovered knowledge for semi-automation of optimization &amp; prediction models based on combination of linked data and machine learning</td>
<td></td>
</tr>
<tr>
<td>B2B: Identify available big data (e.g. phasor measurement units) and other sources to be shared to reap full value</td>
<td>*Development of applications reveals requirements about need for semi-automation analytics workflows</td>
<td>Application development that reap value from prescriptive analytics, such as decision automation</td>
<td></td>
</tr>
<tr>
<td>[Mobility is fully digitalized]</td>
<td>*Initial application of data provenance methods reveal scalability requirements</td>
<td>(Semi-) automated data provenance on Linked Data for increased interpretability, and reliability of data and intelligence</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inter organization (across applications) and intra-organizational data sharing policies based on Linked Data, encrypted storage</td>
<td>*Loss of anonymity through increased transparency reveals requirements for differential privacy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Implementation of differential privacy models and algorithms</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upgrade of data sharing policy based on new technical capability of privacy &amp; confidentiality preserving analytics</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Capability of real-time for automation processes in energy and mobility through automated analytics deployment on any computing unit along infrastructure</td>
<td></td>
</tr>
</tbody>
</table>
Abstraction (Data Mgmt & Engineering) - Develop analytics toolbox into a framework ("analytics engine") based on new requirements from scalable semi-automated real-time, prescriptive analytics in energy and mobility applications

Year 1 | Year 2 | Year 3 | Year 4
---|---|---|---
Develop analytics toolbox into a framework ("analytics engine") based on new requirements from scalable semi-automated real-time, prescriptive analytics in energy and mobility applications | Refined abstraction levels between scalable smart data infrastructure, analytics engine, and apps
Semi-automation of analytics workflow and model deployment | Analytics engine capable of self-annotating discovered knowledge

Table 7. Timeframe for the major expected outcomes of technology adaptation and innovation

6.5.2 Related EU projects

**EU CSA BYTE**\(^1\), the project focusing on socio-economic aspects of the big data paradigm shift and aiming to produce a roadmap that helps minimize the associated risks and maximize the opportunities of big data scenarios across the European sectors

**EU Optique** concentrates on reducing these semantic challenges, to give end users scalable semantic access to Big Data. The challenge of variety in Big Data is also the driving business and research question behind the project. As stated on the projects web site: in traditional businesses such as Oil & Gas, 30–70% of engineers' time is spent looking for and assessing the quality of data.

**EU Bigfoot** - Big Data Analytics of Digital Footprints: The aim of BigFoot is to design, implement and deploy a Platform-as-a-Service solution for processing and interacting with large volumes of data coming from ICT Security, Smart Grid and other application areas. The BigFoot stack builds upon and contributes to the Apache Hadoop ecosystem and the Apache OpenStack project.

**EU FET DataSIM**: Data Science for Simulating the Era of Electric Vehicles; aims at providing an entirely new and highly detailed spatial-temporal microsimulation methodology for human mobility, grounded on massive amounts of big data of various types and from various sources (GPS, mobile phones and social networking sites), with the goal to forecast the nation-wide consequences of a massive switch to electric vehicles, given the intertwined nature of mobility and power distribution networks.

6.6. Conclusion and Recommendations

The energy and transportation sectors, from an infrastructure perspective as well as from resource efficiency and quality of life perspectives, are very important for Europe. The high quality of the European infrastructures and global competitiveness also needs to be maintained with respect to the digital transformation and big data potentials.

The analysis of the available data sources in energy as well as their use cases in the different categories for big data value: operational efficiency, customer experience, and new business

\(^1\) http://byte-project.eu/
models, helped in identifying the industrial needs and requirements for big data technologies. Already in the discussion of these requirements, it becomes clear that a mere utilization of existing big data technologies as employed by the online data businesses will not be sufficient. Domain- and device-specific adaptations for use in cyber-physical energy and transportation systems are necessary. Innovation regarding privacy and confidentiality preserving data management and analysis is a primary concern of energy and transportation stakeholders. Without satisfying the need for privacy and confidentiality there will always be regulatory uncertainty and uncertainty regarding user acceptance of a new data-driven offering.

Among the energy stakeholders, there is a common sense that “big data” is not enough: The increasing intelligence embedded in the infrastructures is also able to analyze data to some extent as to deliver “smart data.” This seems to be necessary, since the analytics involved will require much more elaborate algorithms than for analyzing click streams. Additionally, the stakes in energy and transportation big data scenarios are very high, since the optimization opportunities and risks affect critical infrastructures.

There are a few examples in the energy and transportation sectors, where a technology for data acquisition, i.e. a smart device, has been around for many years, or that the stakeholders have been measuring and capturing a lot of data. However, it was not until recently that this potentially big data could be communicated, stored, and processed cost-effectively. Additionally, there was no business need for it (e.g. increased dynamics through Renewables only now). Hence, some stakeholders run the danger of not acknowledging the technology push. On the other hand, the unclear regulation on what is allowed to do with this data keeps them from experimenting. EU BIG and other coordination and support actions such as EU BYTE, have a great opportunity to fill in these gaps by providing necessary analyzed information, by gathering information on white spots, and fostering the exchange among the stakeholders.

Many of the state of the art big data technologies just await adaptation and usage in these traditional sectors of energy and mobility. With this technology roadmap we have identified and elaborated those high-priority requirements and technologies that will take the energy and mobility sectors beyond state of the art, such that they can concentrate on generating value by adapting and applying those technologies within their specific application domains and use cases.
## 6.7. Abbreviations and acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EOBR</td>
<td>Electronic On Board Recorders</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and communication technologies</td>
</tr>
<tr>
<td>SCADA</td>
<td>Supervisory Control and Data Acquisition</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured English Query Language</td>
</tr>
<tr>
<td>PMU</td>
<td>Phasor Measurement Unit</td>
</tr>
<tr>
<td>RTU</td>
<td>Remote Terminal Unit</td>
</tr>
</tbody>
</table>
6.8. References


7. Manufacturing Sector

7.1. Introduction

This chapter builds on previous deliverables of the BIG project to present a final roadmap for Big Data adoption within the manufacturing sector. The following section presents the six most relevant application scenarios from the manufacturing sector with a concise synopsis and identifies their main business objectives. Section 5.3 identifies the requirements derived from these application scenarios and groups the requirements, also providing a mapping from the requirements to the application scenarios. Section 1.4 then projects the requirements to technologies and identifies some of the research questions triggered by the technological requirements. General related issues in culture, policy, and general economic considerations follow in section 1.5 and final conclusions and recommendations are presented in section 1.6.

7.2. Application scenarios

For the manufacturing sector, a number of application scenarios involving Big Data technologies have been identified and analysed, as described fully in deliverables (Zillner et al., 2013) and (Zillner et al., 2014). Concise information is provided here for ease of reference and as a starting point for condensing the requirements, technologies suitable to meet requirements, open research questions, and the resulting roadmap for Big Data application in the manufacturing sector.

All application scenarios have a strong focus on manufacturing in the European and other similar markets. They are oriented towards market leaders, especially in niche markets, in high quality products, specialised products, and less in mass production that relies on cheap labour. These are areas where Europe (and other highly developed countries) have their strong points, are market leaders and have opportunities to established leading markets.

7.2.1 AS1 Optimisation of production, service, and support

<table>
<thead>
<tr>
<th>Summary</th>
<th>This scenario represents the general optimisation of the production process, service and support processes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synopsis</td>
<td>Sensors acquire product data along the production, delivery, and logistics chains, including the eventually deployed product. Many manufacturing companies are working on extending the used data to include their suppliers and partners in the optimisation process.</td>
</tr>
</tbody>
</table>
| Business Objectives | • Reduce production costs  
• Improve product quality  
• Flexible production, customised products, lot size one |
7.2.2 AS2 Optimisation of distribution and logistics

| Summary | This scenario represents the specific optimisations in distribution and logistics based on connecting delivery vehicles with their environment and sensor information on storage. |
|-----------------------------------------------|
| Synopsis | An increasing number of vehicles are equipped with sensors and control modules that acquire data such as energy consumption, state of wear of parts, and positioning data that are transmitted and stored in data bases. Based on such data, scheduling personnel can improve the planning of transports, change routes and loads where needed and minimise maintenance and standstill costs. |
| Business Objectives | • Optimise cost of logistics  
• Increase availability of materials  
• Minimise maintenance costs  
• Minimise standstill times |

7.2.3 AS3 Production plant planning

| Summary | This scenario represents future opportunities to apply integrated models of the entire production process when planning new production plants or remodel existing facilities. Simulations can optimise multiple aspects of the manufacturing process. |
|-----------------------------------------------|
| Synopsis | With the integration of data and data models in the entire engineering life cycle comes the opportunity to extend integrated approaches to the planning of entire production plants. Existing simulation models must be adapted and evolved to support all steps of planning new production facilities that allow for modularized, flexible and reconfigurable manufacturing processes that are supported by cyber-physical systems. |
| Business Objectives | • Optimise production processes  
• Increase flexibility  
• Enable central and decentralised management and control |

7.2.4 AS4 Predictive maintenance

| Summary | This scenario represents the application of Big Data analytics to predict and prescribe maintenance based on sensor data and analytics rather than standardised intervals. |
|-----------------------------------------------|
| Synopsis | Using predictive analysis, expected failure of machine parts can be predicted with greater accuracy and reliability. Current maintenance schedules follow a worst-case scenario and call for a scheduled down-time for maintenance, including the exchange of often expensive tools in regular intervals.  
With on-line supervision of the machine state, such maintenance intervals can be made flexible and be extended to apply to the actual state of the system as opposed to the current worst-case scenario. |
| Business Objectives | • Lower cost of maintenance when avoiding unnecessary maintenance  
• Avoid failure and subsequent standstill when scheduling early but necessary maintenance  
• Develop new business models, offering maintenance as a service by the machine manufacturer |
7.2.5 AS5 Service-based integrated environment

**Summary**
This scenario represents the integration of digital services (Internet of Services), smart, digital products (Internet of Things) and the production process.

**Synopsis**
Hardware and Software will be offered as services, all integrated to support Big Data Usage. A stack of services will provide the environment for technical standards and protocols, for new platforms that enable firms to build specific applications upon a shared framework/architecture.

**Business Objectives**
- Manage the integration of entire infrastructures
- Integrate individual steps with the complete production (value) chain

7.2.6 AS6 (Big Data) user interfaces

**Summary**
This scenario represents the enabling of human operators and managers with efficient access to the data and efficient tools to exploit the data.

**Synopsis**
In a production environment, the data tools may only take attention from actual production to the extent that they provide a quantifiable benefit. See, e.g., the use case for predictive maintenance.

In the area of qualification and capturing corporate knowledge, Big Data offers many opportunities for on-the-job qualification with new assistive technology that can guide workers step-by-step through the individualised manufacturing sequences needed for customised production, up to the extreme case of “lot size 1”. In a digitized environment, such assistive technology can take the current situation and environment into account: machine state, special requirements of the product, individual information about the worker, etc.

**Business Objectives**
- Improve qualification of workforce
- Quality assurance
- Enable customised production, lot size one

From the perspective of the application scenarios, a number of common or recurring aspects can be identified that lead to common requirements and can be translated to common technical requirements, possibly across industry sectors.

Common to all application scenarios is the business situation as described above: a preference for quality and customised products and innovation over cheap mass production, relying on cheap labour.

A recurring theme is the integration of individual production steps and even production lines with business process within a factory (intra-mural) as well as with other elements of the complete manufacturing value chain (inter-mural), leading to challenges in data integration with other companies and related challenges like data and IPR protection.

Data integration is a central theme in many application scenarios, calling for supporting data semantics and data models, standards, document (natural language) analysis etc.

Innovation in manufacturing aims internally at empowering workers and managers, calling for support through Big Data interfaces for decision makers on all levels, beginning at manufacturing plant planning down to supporting workers at individual machines.
Accompanying developments are a general digitalisation and enrichment of manufacturing under the name "Industry 4.0" (or industrial internet in the U.S.) and the impact of "Internet of Things" (IoT) on manufacturing and products.

Future innovations in manufacturing will see a much bigger role of IT in addition to classical engineering.

7.3. Requirements

Requirements for the manufacturing sector were analysed and provided in (Zillner et al., 2014). The following section aims at highlighting the most important requirements in a way that supports the identification of common requirements for the purposes of the roadmap.

Requirements of AS1 Optimisation of production, service, and support

Sensor technology. The application scenario relies on a much extended number of sensors, ranging from sensors in production machinery to sensors in products (including intermediate) stages. With the uptake of the Internet of Things (IoT), a framework for the increase of sensors is available. A recurring special task of sensor is the proper identification of objects, e.g., product parts and raw materials. A special requirement in manufacturing is related to the hostile working environment where sensors must function.

Real-time data transmission. Sensor data as described above must be acquired at appropriate, often high, sample rates and also be transmitted close to real-time to be used effectively. Decisions can be made at central planning, command, and control points or can be made in a distributed fashion at local machinery or plant divisions. Data transmission may additionally be hampered by the hostile working environment in manufacturing.

Data models. To ensure compatibility of data and appropriate semantics of data in analysis and planning processes, proper data models and data exchange standards must be made available and adhered too. This is a challenge in manufacturing where data will come from manufacturing equipment from multiple sources (machine manufacturers).

Data integration. When extending the application scenario to larger parts of the value chain, data needs to be exchanged with suppliers of raw goods and product parts as well as customers. Customers need not be consumers but might integrate products into larger products. Data exchange poses challenges for data formats and integration as well as IPR and privacy issues. Not all product-related data shall be exchanged with all parts of the production chain.

Real-time analytics. With manufacturing being a mostly continuous, real-time process, data analytics must be provided in a close to real-time manner to support flexible manufacturing.
Requirements of AS2 Optimisation of distribution and logistics

Sensor technology. To support logistics in a data rich manufacturing environment of Industry 4.0, sensor data must be available to locate transport vehicles, know their status, understand available inventory and inventory needs. Location sensors must be available to navigate to moving targets, as in a more individualised manufacturing environment the exact target locations of material and products will potentially change often. A recurring special task of sensor is the proper identification of objects, e.g., product parts and raw materials. A special requirement in manufacturing is related to the hostile working environment where sensors must function.

Real-time data transmission. Data transmission must be sufficiently close to real-time, greatly improving on the currently long intervals (hourly or larger) in which inventory data is sampled. Data transmission may additionally be hampered by the hostile working environment in manufacturing.

Analytics processes for planning. To support flexible logistics, big data analytics and algorithms for constantly updated planning must be available that provides practical and close to real-time schedules for logistics processes and at the same time provides the necessary guarantees and redundancies.

Real-time analytics and planning. When production plans change in intervals close to real-time (see AS1), logistics planning must follow closely. Big data analytics and planning algorithms must be available that offer (close to) optimal plans in (close to) real-time.

Requirements of AS3 Production plant planning

Modelling and Simulations. When re-designing production plants, either improving existing ones or developing new ones, existing data describing the production process must be available and be useful for inclusion in modelling and simulation tools.

Spatial and temporal analytics. Modelling for planning of manufacturing processes incudes many spatial aspects, e.g., the interactions of robotic arms, workers, and semi-finished product in car manufacturing.

Visualisation. To support human decision making and guidance, appropriate data visualisation tools must be available and integrated to support browsing, controlling and decision making in the planning process. This applies primarily to general big data but extends to and includes special visualisation of spatio-temporal aspects of the manufacturing process as described in spatial and temporal analytics.

Data integration. To use existing plant data and specifications of future equipment, data must be integrated from a large variety of sources. Data formats and their semantics must be integrate-able.
Requirements of AS4 Predictive maintenance

**Sensor technology.** A proper understanding of the needed features for predictive algorithms is needed and the necessary data must be available in high quality (including sample frequency).

**Real-time data transmission.** Ideally, impending failure and correspondingly needed maintenance can be predicted with long time horizons and based on long-term change of sensor data. In practice, rapidly changing sensor data can also point to possible immediate problems, not detected by existing self-monitoring equipment. This requires real-time data transmission.

**Data storage.** Predictive maintenance can be based on data from all instances of a production machine, possible collected at the manufacturer of the machine. This requires the availability of appropriate data storage.

**Data integration.** On a smaller scale, sensor data from one machine must be integrated. On a larger scale, however, where data from all instances of a production machine is collected, the context data (types of tasks executed by the machine) must be integrated.

**Analytics for prediction.** Appropriate big data algorithms must be available to derive useful predictions from sensor and context data.

Requirements of AS5 Service-based integrated environment

AS5 is an extension of AS1 with some overlap in addressing the complete manufacturing value chain.

**Data integration.** While full, global optimisation of manufacturing cannot be achieved, data integration across all dimensions (from machines to facilities, fleets, networks and along the manufacturing process) is a requirement for all initial steps towards global optimisation.

**Services.** It can be expected that big data applications in this application scenario will rely on external services for all aspects of optimisation.

**Data and service semantics.** To support the integration of (multiple) external services, the data and service semantics must be explicit and compatible to allow for plug-and-play and even automated combination of services.

**Data security, IPR protection and privacy.** As already described in AS1, data and the IPR that is contained in the data must be properly protected and filtered when data is exchanged beyond the boundaries of an individual company.
Requirements of AS6 (Big Data) user interfaces

Natural Language Analytics and Synthesis User interfaces in manufacturing must present data and analytics in a summarizing fashion, including natural language reports and visualisations (below). In a symmetric interface, natural language input must then also be available. Natural language analytics is also required to include the large variety of unstructured documents available in manufacturing (e.g., technical specification documents and manuals).

Visualisation Another part of user interfaces in manufacturing are appropriate visualisations for understanding as well as controlling and browsing data sets and analyses. See also AS3 for spatio-temporal visualisations.

Data security and privacy Analytics users may not have complete or direct access to individual data to protect (even anonymised) private data.

The following table maps the use cases to the requirements:

<table>
<thead>
<tr>
<th>ID</th>
<th>Big Data Requirement</th>
<th>AS1</th>
<th>AS2</th>
<th>AS3</th>
<th>AS4</th>
<th>AS5</th>
<th>AS6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Req1</td>
<td>Availability of sensor data</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Req2</td>
<td>Real-time transmission of sensor data</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Req3</td>
<td>Data integration across manufacturing chain</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Req4</td>
<td>Data integration for planning</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Req5</td>
<td>Real-time analytics for planning and scheduling</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Req6</td>
<td>Real-time analytics for simulation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Req7</td>
<td>Visualisation in user interfaces</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Req8</td>
<td>Natural language processing</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Req9</td>
<td>Data services</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Req10</td>
<td>Data security, privacy, and IPR protection</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Req11</td>
<td>Data storage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 8. Manufacturing requirements and mapping to application scenarios*

The requirements of the manufacturing sector in its move to data-using business models are complex, but the requirements can be generalised to a certain extent, and then prioritised against objective criteria.

Some requirements underpin many different use cases; therefore the technical solution could have multiple applications and benefit many players. Any method at the interface between cooperating companies that helps integrating data will see applications in many similar situations and benefit all parties involved.
Naturally, some requirements are also salient for industries beyond the manufacturing application scenarios described here. The deliverable D2.5 Cross-sectorial roadmap consolidation will present the most important shared requirements across the ten sectors of the BIG project.

### 7.4. Mapping Requirements to Technology

The next step in setting up the technology road map consists of identifying the relevant technologies and the specific research questions posed by the application scenarios AS1 to AS6. A first collection of requirements across application scenarios is already presented in Table 8 above. The following Table 9 takes the next step of grouping related requirements, identifying the corresponding domains and technologies and finally identifying the specific research questions raised by the application scenarios.

<table>
<thead>
<tr>
<th>ID</th>
<th>Big Data Requirement</th>
<th>Domain / Technologies</th>
<th>Research Questions</th>
</tr>
</thead>
</table>
| Req1 | Availability of sensor data | Applies mainly to hardware technologies outside of classical Big Data, such as network and radio hardware and protocols and sensor technology. Identifying parts, materials, and machinery to access data. Distributed storage of data. | • Sensor robustness in hostile working environments  
• Data transmission robustness in hostile working environments  
• Robust identification technologies for parts and materials  
• Object memories |
| Req2 | Real-time transmission of sensor data | | |
| Req3 | Data integration across manufacturing chain | Data semantics and underlying data models for manufacturing. Linking manufacturing data to enterprise information systems and non-technical business process models (ERP etc.). | • Specialised manufacturing data models  
• Ontology standards and links to external standards  
• Document analysis technologies, including text understanding |
| Req4 | Data integration for planning | | |
| Req5 | Real-time analytics for planning and scheduling | These requirements are generic for Data Analytics: big, streaming, and diverse and unstructured data must be analysed. For some applications, results must be available in near real-time. | • Large data sets, especially when analysing whole fleets or all machines of the same model  
• Analytics must be open to user interaction and control  
• Analysis and simulations must deal |
<table>
<thead>
<tr>
<th>Req</th>
<th>Requirement</th>
<th>Description</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Req7</td>
<td>Visualisation in user interfaces</td>
<td>Visual Analytics is the general domain. Application scenarios have specific requirements.</td>
<td>• Interactive visualisation, visual queries in relation to spatio-temporal data</td>
</tr>
<tr>
<td>Req8</td>
<td>Natural language processing</td>
<td>Natural language understanding for processing unstructured documents and data. Natural language interaction for command and control interfaces for Analytics and planning software.</td>
<td>• Connect classical text mining and understanding to manufacturing specific data and semantics models&lt;br&gt;• Interaction in hostile working environments (noisy, dirty, hot, etc.) needs robust interpretation and presentation technologies</td>
</tr>
<tr>
<td>Req9</td>
<td>Data services</td>
<td>Smart services are a new field of research with wide-ranging consequences for business models. They are a model for new markets.</td>
<td>• Service integration, based on data, service and business aspect description languages&lt;br&gt;• Business models of services and service market(place)s</td>
</tr>
<tr>
<td>Req10</td>
<td>Data security, privacy, and IPR protection</td>
<td>Conflicting interests in storing data on products for easy retrieval and protection of data from unauthorised retrieval. Protect IPR as far as it is encoded in product and production data. Regulations for data ownership, e.g., what access may the producer of a production machine have to usage data. Privacy protection for workers interacting in an industry 4.0 environment.</td>
<td>• Integration of data encryption and access control into object memories&lt;br&gt;• Harmonisation of European and worldwide regulations&lt;br&gt;• Data privacy regulations and transparent privacy protection implementations</td>
</tr>
<tr>
<td>Req11</td>
<td>Data storage</td>
<td>Analytics on Table 9machinery can span across entire fleets and include all instances of a machine model that have ever been built. The resulting data sets can be very large and yet must be available for real-</td>
<td>• Expandable, high-availability mass storage</td>
</tr>
</tbody>
</table>
Table 9. Technologies and research questions for the manufacturing sector.

Requirements grouped in Req10 and the corresponding technology domains and “research” questions are mostly actually policy-related and are thus considered for the technology roadmap laid out in section 7.4 but rather in section 7.5.

Requirements grouped in Req11 are mainly a challenge in terms of cost-effectiveness as the necessary technology is available already and developing quickly. However, in many concrete business cases, the investment necessary and the ROI are deciding factors that are not always met by currently available technology.

The smart service related requirements Req9 are the least precise. They relate to a new and rapidly evolving field in the interface between IT technologies and economics and business world. New architectures for the technology as well as the market structure are currently being developed, see, e.g., the presentations in (acatech, 2014).

7.4.1 Technology Roadmap

The following diagram shows a high-level view for the technology roadmap in the manufacturing sector. It is a summary of Table 9 grouped by functional domains.
Based on input from BIG’s Technical Working Groups, the following Figure 21 shows estimates for the development timeline of the most relevant technologies. Times represent estimates for viable technologies being available on the market, unless noted otherwise. Certain technologies cover a wide range of challenges, e.g., data sensors and transmission technologies for hostile working environments and will naturally see a gradual development and adaptation in the market. Capabilities for machine learning and natural language processing will also develop gradually with no clear criteria or milestones. Existing technologies include Support Vector Machines, Gaussian Processes, and Expectation Maximisation; currently developing technologies include Bayesian approaches and Deep Learning. The main area that is not impeded by technology development but company interests is the development of standards. Support for standardization, e.g., in the case of manufacturing ontology standards, can influence creation and adoption times.
Technology Roadmap
Manufacturing Sector

Technical Requirement

<table>
<thead>
<tr>
<th>Sensor data availability</th>
<th>Object Memories in use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust sensors</td>
<td>Sensors available for most hostile environments</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Real-time transmission</th>
<th>Technology available for most hostile environments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust networks</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data integration</th>
<th>Manufacturing data models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology-based integration</td>
<td>Manufacturing ontology standard</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Deep Data Analytics</th>
<th>Spatio-temporal simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine learning</td>
<td>ML for large data sets</td>
</tr>
<tr>
<td>Simulation tools</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Big Data User-interfaces</th>
<th>Full NLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visualisation</td>
<td>Visual analytics</td>
</tr>
<tr>
<td>Natural Language Processing</td>
<td>NL report generation; NL interfaces</td>
</tr>
</tbody>
</table>


Figure 21: Timeline for the technology roadmap

7.4.2 Related EU projects

As identified in other sectors (see, e.g., the telco, media, and entertainment chapter), three projects in particular are helping to support the EU’s objectives in promoting the strategic value of Big Data to Europe’s businesses. A number of projects related specifically to the manufacturing sector are following below.

BYTE\(^1\)

“The Big data roadmap and cross-disciplinary community for addressing societal Externalities (BYTE) project will assist European science and industry in capturing the positive externalities and diminishing the negative externalities associated with Big Data in order to gain a greater share of the Big Data market by 2020.”

RETHINK big\(^2\)

\(^1\) http://byte-project.eu/
\(^2\) http://www.rethinkbig-project.eu
“The objective of the RETHINK big Project is to bring together the key European hardware, networking, and system architects with the key producers and consumers of Big Data to identify the industry coordination points that will maximize European competitiveness in the processing and analysis of Big Data over the next 10 years.”

**MICO**¹ (Media In Context)

“MICO is a European Union part-funded research project to provide cross-media analysis solutions for online multimedia producers. MICO will develop models, standards and software tools to jointly analyse, query and retrieve information out of connected and related media objects (text, image, audio, video, office documents) to provide better information extraction results for more relevant search and information discovery.”

In addition, there are a number of EU projects specifically related to the manufacturing sector that work at various aspects of data and software related to Big Data. These include:

**iPROD**²

“iProd general aim is to improve the efficiency and quality of the Product Development Process of innovative products by developing a flexible and service oriented software framework that, reasoning and operating on a well-structured knowledge, will be the backbone of the computer systems associated with current and new product development processes.”

**ADVANCE**³

“The patterns and dependencies that exist in the 50 million or more data elements created daily in logistics networks can only be meaningfully processed by intelligent data-mining approaches linked to strategic decision making based on longer term analyses of billions of pieces of information. ADVANCE will support networked companies in improving their information collecting and processing infrastructure, enabling strategic planning coupled with instant decision making. The project will deliver the open-source ADVANCE platform and comprehensive accompanying reference material. The ADVANCE software will provide a dual perspective on transport requirements and decision making dependent on the latest snapshot information and the best higher-level intelligence.”

**ProaSense**⁴

“ProaSense is an EU project aiming at the core goal to pave the way for an efficient transmission from Sensing into Proactive enterprises: exploiting the power of big enterprise data, by sensing the whole business ecosystem, extracting the actionable meaning from data, by applying advanced big data analytics, and increasing the strategic value of data analysis for the decision making, by dynamically adapting patterns of interest to be found in the real-time big data streams. Use cases include proactive manufacturing in an automatic production line.”

---

¹ [http://www.mico-project.eu](http://www.mico-project.eu)
² [http://www.iprod-project.eu/](http://www.iprod-project.eu/)
⁴ [http://www.proasense.eu/](http://www.proasense.eu/)
7.5. Market situation

To put our technology assessment and timelines for the Manufacturing sector into perspective, we take a brief look at the current market situation for Manufacturing and Big Data related fields.

As a general setting, big players in the Internet of Things domain that is closely related to Big Data technology for manufacturing, present accumulated estimates like the following:

“[A] global network connecting people, data and machines called the Industrial Internet [has] the potential to add $10 to $15 trillion to global GDP over the next 20 years.” (GE\(^1\)) and “The increased connectedness we’re beginning to experience equates to a US$ 19 trillion global opportunity to create value over the next decade through increased profits for businesses as well as cost savings, improved citizen services and increased revenues for governments and other public-sector organizations.” (John Chambers, CEO Cisco, Jan. 15, 2014\(^2\))

On a more down-to-technology level, sensor and (Big Data) software enhanced machinery can already generate great economic impact. GE states about its latest wind turbine:

“The sensors and software alone can make the GE2.5-120 wind turbine 25 per cent more efficient and 15 per cent more productive than comparable GE models.” (GE\(^1\))

Our core application scenarios AS1, AS2, and AS4 relate to existing optimisation practices in the manufacturing sector. A key paradigm for optimisation is lean manufacturing, or more general lean management. McKinsey has studied this integration of data-driven insights and existing approaches, e.g., the kaizen method of continuous improvement. The key messages from their analysis are a mapping of available data-driven methods to various sub-sectors within manufacturing and a collection of use cases. E.g.,

“a leading steel producer used advanced analytics to identify and capture margin-improvement opportunities worth more than $200 million a year across its production value chain. This result is noteworthy because the company already had a 15-year history of deploying lean approaches and had recently won an award for quality and process excellence.” (McKinsey, 2014\(^3\)).

The table of techniques and sub-sectors shows how diverse the technology needs are and is reproduced below.

---

2 http://forumblog.org/2014/01/are-you-ready-for-the-internet-of-everything/
3 http://www.mckinsey.com/insights/operations/when_big_data_goes_lean
Table 10: Technology needs per industry

7.6. Cultural, policy and wider economic aspects for Manufacturing

Besides the technical requirements collected in the previous section, there are a number of cultural, policy-related and general economic aspects related to the roadmap for the manufacturing sector.
Business culture, particularly at the top-level management is a potential blocker to big data uptake and the implementation of corresponding application scenarios. Hesitation is based on high investments with an unclear time-line and sceptical, i.e., short estimates of the life time of big data implementations. Another blocker is the lack of standardisation of solutions which raises the fear of investing in the wrong technology. An underlying problem is the lack of global strategies for the move to Big Data based manufacturing that this roadmap is trying to help alleviate.

Also related to business culture is a need to move access to technical analysis of data into the hands of business decision makers. This can and must be supported by tools, but also requires changes in the standing of technical skills and acceptance of such skills for and by business decision makers.

Another aspect on the interface of business culture and policy is data privacy for workers, where unions, businesses and policy-makers have to agree on and provide satisfactory and verifiable solutions.

Finally, a general economic aspect related to technology skills, human resources needs for Big Data uptake in manufacturing must be met. Engineering and business education must include more of the technical aspects of data analytics, planning, simulation, etc. and a rising number of open positions for purely technical skills as well as the managements aspects around them is expected, see also (Becker, 2014).

### 7.7. Conclusion and Recommendations

The predecessor document (Zillner et al., 2014) has highlighted the most important requirements for the main application scenarios of the sectors. They are here analysed with regard to technology solutions and to the related research questions that must be examined to give the sector (and others) the greatest opportunities for success in Europe over the coming years.

The main theme from this analysis for the manufacturing sector is data integration, in particular the integration of sensor data into a comprehensive data model that needs to be extended to eventually cover the complete manufacturing chain from raw materials to post-consumer. Also, manufacturing data must be integrated with business software, starting with ERP and business process systems, and also additional tasks such as production plant planning.

A second theme is a set of research questions grouped around the hostile working environments of manufacturing. Sensors, data transmission, and user interaction must function robustly under such hostile environments, calling for redundancy and new technologies where applicable.

Another theme, although non-technical, is grouped around data protection and data privacy. Company IPR is potentially encoded in product (or part, or object) memories and must be protected from unauthorised access. Workers’ data privacy issues must be addressed by the appropriate stakeholders.
Another non-technical issue is the **availability of technical skills**, on the one hand in terms of education of sufficient numbers of data specialists and on the other hand in terms of changing business processes and culture to give business decision makers the technical skills or at least access to the technical skills to make efficient use of Big Data analytics in the manufacturing sector. This requirement is related to technical and research questions regarding appropriate Big Data user interfaces with visualisation, visual queries, and natural language interfaces.

In most European countries, the impact of Big Data in the context of Industry 4.0 will be the requirement to keep and extend current market advances in the areas of high-quality, high-tech and customised products. As automation will help other competitors to raise quality from manual manufacturing levels, the focus will be on high-tech and customised products.

High-tech or smart products with individual IDs and object memories (Internet of Things) will be used in integrated environments, e.g., the automobile and its life cycle.

In a changing market, the optimal strategy will be a dual strategy, combining the creation of **market leaders** with the establishment of **leading markets**. Market leaders build on top of developed and developing base technologies, i.e., Big Data technologies and other, CPS-related manufacturing technologies. In the manufacturing sector, market leaders are machine and plant engineering and construction companies, manufacturers of automation technology and corresponding integrators and service companies in the ICT sector. The leading markets in Europe and the world are the production facilities, including their network of suppliers and services, i.e., their entire value chain. Both perspectives apply to the European market and should thus be combined in a strategy for adapting the manufacturing sector to the Big Data and Industry 4.0 related changes.
## Abbreviations and acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSS</td>
<td>Business Support Systems</td>
</tr>
<tr>
<td>CPS</td>
<td>Cyber-Physical Systems</td>
</tr>
<tr>
<td>CRM</td>
<td>Customer Relationship Management</td>
</tr>
<tr>
<td>DWH</td>
<td>Data Warehouse</td>
</tr>
<tr>
<td>ERP</td>
<td>Enterprise Resource Planning</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>I4.0</td>
<td>Internet 4.0</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communication Technology</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>IPR</td>
<td>Intellectual Property Rights</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>NLU</td>
<td>Natural Language Understanding</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>ROI</td>
<td>Return on Investment</td>
</tr>
<tr>
<td>XML</td>
<td>eXtensible Markup Language</td>
</tr>
</tbody>
</table>
7.8. References


8. Sector Retail

8.1. Introduction

Digital services for customers provided by smart systems will be essential for the success in the future retail business. The store of the future will not only provide services for the retailers. In particular, the retail domain will especially be focused on highly efficient and personalized customer assistance services. Retailers are confronted with the challenge to meet the demand of a new generation of customers who expect information to be available anytime and anywhere. Furthermore, they want to be treated as individuals and with respect to their personal demand. In the future, new intelligent services that make use of Big Data will allow a new level of personalized and high-quality Efficient Consumer Response (ECR).

The retail sector represents a very powerful sector. In Europe retail and wholesale accounted for a share of 11% of the GDP and employed about 33 million persons in 2010. Nowadays Big Data plays only a minor role in stationary stores, but with the increasing number of sensors and actuators installed, a huge amount of data will be generated. Using intelligent analytics on this data will lead to higher gains. Big data can in this case not be presented as Return-On-Investment over a certain period of time, but as Return-on-Information as it creates more differentiated and sustainable competitive advantages and better products and services. McKinsey estimated in 2011 that retailers are able increase their operating margins by more than 60% by making full use of Big Data.

We obtained a wealth of information concerning the present and desired situation in the retail sector in the interviews we conducted – see final version of sector’s requisites (D.2.3.2). Out of that, we identified two scenarios with the highest value creating potential for the retail sector in which Big Data plays a key role. In this chapter, we first describe these important application scenarios by mentioning specific required data sources and functionalities. Afterwards, we present the requirements necessary for the application scenarios followed by the mapping of appropriate technologies to these requirements. Finally, we conclude our findings.

8.2. Application Scenarios

Out of the conducted interviews we could identify two major use cases for retail. The first one “In-Store Precision Retailing” deals with the collection of customer data for marketing purposes. This is already wide spread and well known for online merchandisers, but also local chain stores could benefit from this data. The second use case “Operational Decision Management in Retail” is concerned about the collection and analysis of data for operational decisions and day to day operations. This includes, among other things, the enrichment of the product database with much more complex data than today. In the following, we provide a coarse description of the two use cases. A more detailed description of the application scenarios can be found in the final version of the sector requirements deliverable (D.2.3.2).
8.2.1 AS 1: “In-Store Precision Retailing”

Description

This application scenario demonstrates how physical stores can benefit from Big Data for customized marketing strategies. The goal is to collect all customer data inside the store. This includes not only the path the customer takes through the market but also all customer interactions with products or staff. This data shall be combined with details about the customer that can be extracted from e.g. social networks and third party data like upcoming events.

The key requirement for this application scenario is the generation of individual customer models. Nowadays, the customer as an individual is unknown and only few personal information, like name, address, gender, age and basket history, is available e.g. through loyalty programs. To realize “In-Store Precision Retailing” a more detailed model of the customer is needed to understand the individual customer requirements and wishes. This is possible by acquiring sensor data inside the store and combining it with information about the customer from the internet (e.g. social media). Having individual customer models the customer could e.g. be provided with tailored advertisements inside the store using multiple channels. Additionally it is also possible to send personalized coupons to the customer’s mobile phone or directly to the cash point where he is located. Another alternative is displaying tailored information on a computer installed at the customer’s shopping cart. Real-time notifications could additionally be used to draw employees' attention to certain customers to give personal and individual advice.

For the implementation of the “In-Store Precision Retailing” Scenario a huge amount of data is required that needs to be analysed. The following section provides an overview of the needed information.

Sources

The required data can be classified according to the source of the data. We can distinguish between data that can be acquired inside the store, online data, data coming from product or customer databases and third party data. The following table lists necessary data for this specific use case according to above mentioned categories.

<table>
<thead>
<tr>
<th>Required Data Sources:</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-store data:</td>
</tr>
<tr>
<td>• entry &amp; dwell time</td>
</tr>
<tr>
<td>• position/ movement data of customers/ staff collected by sensors</td>
</tr>
<tr>
<td>• floor &amp; space plan</td>
</tr>
<tr>
<td>• customer - staff interaction</td>
</tr>
<tr>
<td>• customer - product interaction: dwell time, conversion rate</td>
</tr>
<tr>
<td>Product database:</td>
</tr>
<tr>
<td>• semantic information (e.g. for product recommendations)</td>
</tr>
<tr>
<td>• ingredients (e.g. to identify customers' preferences)</td>
</tr>
<tr>
<td>Customer database:</td>
</tr>
<tr>
<td>• personal data: name, address, gender, age, …</td>
</tr>
<tr>
<td>• shopping history</td>
</tr>
</tbody>
</table>
family context
• customer segmentation
• history of service requests

Online data:
• online customer profiles in social networks
• reviews (from a specific customer, for a specific product or from third parties)

Third party data:
• demographic data for the residential area of a customer
• weather (e.g. for advertisement/ recommendation)
• upcoming events like football cup
• trends from social networks (→ combine with user profile and friends' preferences)

Table 11. Required Data Sources in AS1

Implementing this application scenario requires acquisition, deep analysis and storage of the data presented above. In section 8.3 “Sector Requirements” more detailed information about the necessary technical requirements that need to be fulfilled is presented.

8.2.2 AS 2: “Operational Decision Management in Retail”

Description

This application scenario demonstrates how Big Data in retail can be used for operational decisions as well as for day-to-day operations. The goal is to automatically collect and analyse in-store data, like product placements, customer-product interactions, customer-staff interactions and cash point data.

In-store data like heatmaps of customer movement can be used to optimize floor plans, cash point management, employee schedules and advertisement areas. Product data can be connected to different manufacturers and real-time information on delivery processes. Warehouse information allows more control on the inventory, intelligent shelves can help to control the inventory and direct staff to the most efficient actions. More complex information in a database can be useful for comparisons between similar products, more specific usage data for a single product, product ratings based on characteristics like dwell time or service requests and comparisons via online reviews. The combination of these product details can be useful for stocking decisions, pricing strategies and other operational decisions.

The key requirements for this scenario are Embedded Systems for real-time shopping data. Nowadays, only few sensors are installed in shopping environments hence only little real-time data is available. For example the stores can only react to out-of-stock warnings but not to out-of-shelf notifications because there is no information available about how many products are located inside the shelf. Another drawback today is that acquiring information about the customer inside the market, like e.g. customer-staff or customer-product interactions, requires manual recording of the data. Furthermore, the analysis and usage of sale data is limited to Business Intelligence. To realize “Operational Decision Management in Retail” a seamless integration of smart embedded systems in shopping environments is necessary.
For the implementation of the “Operational Decision Management in Retail” Scenario a huge amount of data is required that needs to be analysed. The following section provides an overview of the needed information.

Sources

The required data can be classified according to the source of the data. We can distinguish between data that can be acquired inside the store, online data, data coming from product or customer databases and third party data. The following table lists necessary data for this specific use case according to above mentioned categories.

<table>
<thead>
<tr>
<th>Required Data Sources:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In-store data:</strong></td>
</tr>
<tr>
<td>• customer movements</td>
</tr>
<tr>
<td>• customer behaviour: dwell times in product departments, conversion rates, product interaction</td>
</tr>
<tr>
<td><strong>Product database:</strong></td>
</tr>
<tr>
<td>• producer, delivery details, inventory</td>
</tr>
<tr>
<td>• history about products (e.g. changed ingredients)</td>
</tr>
<tr>
<td>• history on packaging details</td>
</tr>
<tr>
<td>• efficiency &amp; pricing of product placements</td>
</tr>
<tr>
<td><strong>Online data:</strong></td>
</tr>
<tr>
<td>• reviews (e.g. for a specific product or from third parties)</td>
</tr>
<tr>
<td>• feedback/ “likes” on a store</td>
</tr>
<tr>
<td><strong>Third party data:</strong></td>
</tr>
<tr>
<td>• ingredients</td>
</tr>
</tbody>
</table>

Table 12. Required Data Sources in AS2

Implementing this application scenario requires acquisition, deep analysis and storage of the data presented above. In section 8.3 “Sector Requirements” more detailed information about the necessary technical requirements that need to be fulfilled is presented.

8.3. Sector Requirements

Realizing the described application scenarios is only possible when all technical requirements are fulfilled. In this section we first explain the technical requirements for the two application scenarios and then map them to the different use cases. At the end we define the most important requirements for the retail sector.
Requirements of AS1: “In-Store Precision Retailing"

The goal of this application scenario is to improve the services for the customer inside the store by adapting systems/services to specific customers. In the following the technical requirements to realize this use case are described.

Real-time Data Transmission

It is important that the data from sensors inside the store is acquired in real-time. This includes e.g. visual data from cameras and customer locations from positioning sensors.

Natural Language Analytics

Modelling the customer requires a huge amount of personal data from the customer that can be extracted from social networks using natural language analytics.

Spatial & Temporal Analytics

Spatial and temporal analytics are needed to identify the users’ location inside the store.

Usage Analytics

Besides extracting personal information about the customer from social media, the interaction of the customer with products and with staff needs to be analysed to be able to segment the customers and to generate fine-grained customer models.

Adaptive Visualization

Based on the generated user profiles from sensor data inside the market and information extracted through natural language analysis from the internet, the visualization has to be adapted to the customer using the service. An example for this are tailored advertisements.

Data Improvement

Both sensor data and data extracted from the internet, like product data and customer personal data, is error prone and needs to be checked for trustworthiness. Therefore data improvement procedures are required that help to remove incorrect/redundant data and noise.

Real-time Insights

Real-time insights about customer personal data, like the location inside the shop, and product data are necessary to be able to give real-time product recommendations or to launch intelligent pricing.
Data Security and Privacy

Personal information about the customer extracted from the internet and information collected from sensors inside the store, like shopping history, product and staff interactions and user location need to be saved securely. Before saving customer-related information the retailer has to obtain the permission from the customer.

Requirements of AS2: “Operational Decision Management in Retail”

The goal of this application scenario is to optimize floor plans, cash point management and employee schedules by automatically collecting and analysing in-store data, like e.g. product placements, customer movements, customer-product and customer-staff interactions and cash-point data. In the following the technical requirements to realize this use case are described.

Real-time Data Transmission

For this use case several types of data have to be acquired in real-time. This includes besides visual data from cameras also product position data from location sensors.

Natural Language Analytics

Retailers are able to optimize their operational decisions when they have information about local special data like upcoming regional events, weather data or even potential natural disasters. Additionally, information from social networks including information/feedback about products can help the retailer to adapt the range of products optimally. Extracting all this information from the internet is only possible using natural language analytics.

Semantic/ Data Modelling

Services that extract product information from the internet to optimize operational decisions require the presence of semantic annotation of products and standardized product ontologies. With this information it is e.g. possible to obtain information about similar products belonging to the same category.

Spatial & Temporal Analytics

Visual data from cameras can be analysed to extract customer movement patterns.

Usage Analytics

Customer segmentation can be performed by analysing customer-product interactions and customer-staff interactions.
Prescriptive Analytics

Prescriptive analytics are required to allow intelligent inventory, intelligent staff scheduling and floor plan/product location optimization. This can be realized by analysing product locations, customer-staff interactions, customer-product interactions and customer movements.

Adaptive Visualization

On the basis of the performed customer segmentation extracted from customer interactions inside the shop, the visualization of services can be adapted to the specific user group.

Data Improvement

Same as in the previous application scenario, both sensor data and data extracted from the internet, like product data and customer personal data, is error prone and needs to be checked for trustworthiness. Therefore data improvement procedures that help to remove incorrect/redundant data and noise are required.

Real-time Insights

Real-time insights about customer movements and user groups are necessary to detect movement patterns and to allow intelligent pricing.

Modelling and Simulations

Customer movements provide useful information to run queuing simulations which can be used to optimize customer-cash distribution.
Mapping Requirements to Application Scenarios

Looking at the previously presented requirements it is noticeable that a great overlap exists between the technical requirements of the two application scenarios. The following figure shows how the requirements are mapped to the application scenarios.

![Diagram showing the mapping of application scenarios to technical requirements]

**Figure 22: Mapping Application Scenarios to Technical Requirements**

The figure also illustrates the categories the different requirements belong to. Overall five main categories have been identified as important for the two application scenarios: Data Management Engineering, Deep Data Analysis, Data Visualization and User Experience, Data Quality and Data Security and Privacy. A closer look at these main requirements is provided in the next section.
# 8.4. Mapping Requirements to Technology

In this section the previously presented main requirements of the retail sector are mapped with appropriate technologies. The mapping has been performed in cooperation with the technical working groups. The following figure illustrates the main requirements and appropriate technical solutions to realize them. Furthermore open research challenges are listed. The timeline illustrates the moment in time a solution to the existing research challenges is expected.

The figure shows that most of the technologies that are necessary to implement the described application scenarios already exist. The task of the retailers mainly is to put these technologies into practice.

## Technology Roadmap

**Retail**

<table>
<thead>
<tr>
<th>Technical Requirement</th>
<th>Entity Matching / Data Alignment with Human Validation</th>
<th>Scalable Triple Stores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Management Engineering</td>
<td>Triple Stores for Product Information</td>
<td></td>
</tr>
<tr>
<td>Deep Data Analysis</td>
<td>Information Extraction, Linked Data, Named Entity Recognition, Sentiment Analysis, Machine Learning</td>
<td>Real-time Performance at Scale</td>
</tr>
<tr>
<td>Data Visualization and User Experience</td>
<td>Apply User Modelling Techniques to Visual Analytics</td>
<td>High Performance, Large Scale Visualization Based on Adaptive Semantic Frameworks</td>
</tr>
<tr>
<td>Data Quality</td>
<td>Semantics, Human Validation via Curation</td>
<td>Efficient Crowd-Sourced Data Curation Process, Robust Data Curation Infrastructures, Standardization of Product Ontologies</td>
</tr>
<tr>
<td>Data Security &amp; Privacy</td>
<td>Encrypted Storage (Lambda/ NoSQL), Metadata Description for Handling Data Privacy Information, Anonymization Algorithm</td>
<td>Best Practices, Breakthrough</td>
</tr>
</tbody>
</table>

*Figure 23: Roadmap for Sector Retail*

## Data Management Engineering

Data management engineering includes real-time data transmission and semantic data modelling. There are two appropriate solutions to store the semantic data models, (1) triple stores for product descriptions and (2) key/value stores for image data. The research challenge using triple stores is the scalability.

Semantic data modelling can be performed using OWL or RDF/RDFS. Curation of the data can be accomplished through entity matching and data alignment with human validation.
Deep Data Analysis
The analysis of the data can be performed using information extraction, linked data, named entity recognition, sentiment analysis and machine learning. Here the challenge is the real-time performance at scale.

Data Visualization and User Experience
Adaptive data visualization can be achieved through applying user modelling techniques to visual analytics. The research challenge is high performance and large scale visualization based on adaptive semantic frameworks.

Data Quality
Data quality can be ensured through the usage of semantics and human validation via curation. The challenge is to automatically remove large amounts of noise at scale and scalable semantic validation.

Data Security and Privacy
Personal data from the customer needs to be managed and stored securely. Therefore encrypted storage is necessary. There exist two appropriate storage solutions depending on the scale. The first one is the Lambda architecture approach, which can deal with high volume and high velocity of incoming data. NoSQL databases represent the second approach for storing and accessing raw data and accessing customer model data. Research challenges regarding storage solutions are high performance, best practices and the query execution on encrypted data. Additionally metadata descriptions for handling data privacy information and anonymisation algorithms are helpful to preserve data security.

8.5. Conclusion and Recommendations

In this chapter we have described two major business scenarios that we identified during the conducted interviews in the retail domain, “In-Store Precision Retailing” and “Operational Decision Management in Retail”. Additionally, we have defined the necessary requirements to accomplish these application scenarios and classified them into five main categories. At the end we have presented appropriate technical solutions to fulfil these five main requirements in the retail sector and mentioned existing research challenges.

Our findings show that in future it will be important to be able to address individual customer requirements like it is already possible in online stores. This requires the stationary retailer to collect personal information about the customer both through information extraction from the internet, like e.g. from social networks, and through specific sensors installed inside the market. This information helps the retailer to classify the customer into a specific user group or even to generate a specific user model, which is necessary to be able to provide tailored services. Acquiring information about the customer in-store requires increasing the number of sensors. These smart systems need to be embedded seamlessly into the environment to realize ambient intelligence.
### 8.6. Abbreviations and Acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OWL</td>
<td>Web Ontology Language</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>RDFS</td>
<td>Resource Description Framework Schema</td>
</tr>
</tbody>
</table>
8.7. References


McKinsey Global Institute (2011): “Big data: The next frontier for innovation, competition, and productivity”.