BDV BIG DATA VALUE ASSOCIATION

BDV SRIA

Big Dava Value Strategic Research and Innovation Agenda

VERSION 3.0 January 2017

The New Economic Asset for Europe

www.bdva.eu

Executive Summary

This Strategic Research and Innovation Agenda (SRIA) defines the overall goals, main technical and non-technical priorities, and a research and innovation roadmap for the European Public Private Partnership (PPP) on Big Data Value. The SRIA has been developed by the Big Data Value Association (BDVA) an industry-led organisation representing European large and SME industry and research.

The SRIA explains the strategic importance of Big Data, describes the Data Value Chain and the central role of Ecosystems, details a vision for Big Data Value in Europe in 2020, and sets out the objectives and goals to be accomplished by the PPP within the European research and innovation landscape of Horizon 2020 and at national and regional levels.

The multiple dimensions of Big Data Value are described, and the overarching strategic objectives for the PPP are set out. These embrace data, skills, legal and policy issues, technology leadership through research and innovation, transforming applications into new business opportunities, acceleration of business ecosystems and business models, with particular focus on SMEs, and successful solutions for the major societal challenges Europe is facing such as Health, Energy, Transport and the Environment. The objectives of the SRIA are broken down into specific competitiveness, innovation and technology, societal and operational objectives.

The implementation strategy for addressing the goals of the SRIA involves four mechanisms: i-Spaces, Lighthouse projects, technical projects, and cooperation & coordination projects. I-Spaces are cross-organisation cross-sector interdisciplinary Innovation Spaces to anchor targeted research and innovation projects. They offer secure accelerator-style environments for experiments for private data and open data, bringing technology and application development together. I-Spaces will act as incubators for new businesses and the development of skills, competence and best practices. Lighthouse projects are large-scale data-driven innovation and demonstration projects that will create superior visibility, awareness and impact. The four mechanisms will foster the development of the European data ecosystem in three distinct phases by establishing an innovation ecosystem, pioneering disruptive big data value solutions, and setting long-term ecosystem enablers.

The strategic and specific goals, which together will ensure Europe's leading role in the data-driven world, are supported by key specific technical and non-technical priorities. Five technical priority areas have been identified for research and innovation: data analytics to improve data understanding; optimized architectures for analytics of data-at-rest and data-in-motion; mechanisms ensuring data protection and anonymisation, to enable the vast amounts of data which are not open data (and never can be open data) to be part of the Data Value Chain; advanced visualization and user experience; and, underpinning these, data management engineering. The complementary non-technical priorities are skills development, business models and ecosystems; policy, regulation and standardization; and social perceptions and societal implications.

Finally, the expected impact of the objectives is summarised, together with KPIs to frame and assess that impact. The activities set out in this SRIA will deliver solutions, architectures, technologies and standards for the data value chain over the next decade, leading to a comprehensive ecosystem for achieving and sustaining Europe's role, for delivering economic and societal benefits, and enabling a future in which Europe is the world-leader in the creation of Big Data Value.

SRIA Version Changes:

Significant updates of content between SRIA Version 2 and this SRIA, Version 3, are indicated through [green square]. A top level itemisation of the changes is provided in Appendix 6.5.

Note: This version of the SRIA is pending formal ratification by the BDVA.

Contents

1 Introduction

The recent developments of the European data market have been reflected in this section.

1.1 Strategic Importance of Big Data Value

The continuous and significant growth of data together with better data access and the availability of powerful ICT systems led to intensified activities around Big Data Value. **Powerful data techniques and tools** allow collecting, storing, analysing, processing, and visualizing vast amounts of data. **Open data initiatives** gain momentum, providing broad access to data from the public sector, business and science.

The European data market measured by the value of the data products and services bought by European businesses and consumers is a fast growing multibillion euro business. According to IDC[1] the compound annual growth rate (CAGR) of the EU Data market over the period 2014-2020 may be as high as 14% under the most favourable scenario. This would mean that the size of the data market in Europe is expected to more than double in the next years boosted by sustained economic recovery and swift adoption of data-driven technologies, thus reaching a value of around 111 billion EUR by 2020.

The exploitation of Big Data in various sectors has socio-economic potential far beyond the specific Big Data market. Therefore, it is essential to embrace new technology, applications, use cases, and business models within and across various sectors and domains. This will ensure rapid adoption by organizations and individuals, and provide major returns in growth and competitiveness. In particular, the efficiency gains made possible by Big Data will also have a profound **societal impact**. As an example, the OECD[2] reports of 380 megatonnes of CO_2 emissions may be saved worldwide in transport and logistics, while the utility sector may see a CO_2 reduction of more than 2 gigatonnes.

The volume of data is rapidly growing. By **2020, there will be more than 16 zettabytes** of **useful data** (16 Trillion GB)[3], which implies growth of 236% per year from 2013 to 2020. This data explosion is a reality that Europe must both face and exploit in a structured, aggressive and ambitious way to create value for society, its citizens, and its businesses in all sectors.

Large companies and SMEs in Europe are clearly seeing the fundamental potential of Big Data for disruptive change in markets and business models, and are beginning to explore the opportunities. IDC confirms that Big Data adoption in Europe is accelerating[4]. According to recent IDC findings[5], the European data market and economy in 2013–2014 was second in value only to the U.S. (with ca. half the market size), and was growing almost as fast. Companies intending to build and to rely on data-driven solutions appear to have started successfully addressing challenges that go well beyond technology. Successful adoption of Big Data requires changes in business orientation and strategy, processes, procedures and the organizational setup. European enterprises are creating new knowledge and start hiring new experts, enhancing a new ecosystem.

Economic and social activities have long relied on data. But the increased volume, velocity, variety, and social and economic value of data signal **a paradigm shift towards a data-driven socio-economic model**.

1.2 The Role of Big Data Value in Digitizing European Industry (DEI)

The **Digitising European Industry (DEI)** initiative[6] recognizes that digitisation of all sectors of the economy is needed for the EU to reinforce its competitiveness, build a strong industrial base, and manage

[1] European Data Market – SMART 2013/0063 D6- First Interim Report, IDC, October 2015, available under: https://idc-emea.box.com/s/sywez4wpbrbjggaliupensocoittxc6y
[2] *"Exploring Data-Driven Innovation as a New Source of Growth – mapping the policy issues raised by "Big Data"",* OECD, 2013
[3] "*The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*" Vernon Turner, John F. Gantz, David Reinsel, and Stephen Minton, Report from IDC for EMC April 2014.
[4] "*The European Data Market*", *Gabriella Cattaneo, IDC, presentation at the European Data Forum in Luxembourg, November 2015 http://2015.data-forum.eu/sites/default/files/1140-1155_Gabriela%20Cattaneo_SEC.pdf*
[5] European Data Market – SMART 2013/0063 D6- First Interim Report, IDC, October 2015, available under: https://idc-emea.box.com/s/sywez4wpbrbjggaliupensocoittxc6y
[6] EC Communication (COM(2016) 180)

the transition to a smart economy. In particular, this requires strengthening leadership in digital technologies and in digital industrial platforms across value chains in all sectors of the economy.

One key such value chain is the Big Data Value chain as depicted in **Error! Reference source not found.**. Europe needs strong players along this Big Data Value Chain, ranging from data generation and acquisition, through to data processing and analysis, then to curation, usage, service creation and provisioning. Each link in the entire value chain has to be strong so that a vibrant Big Data Value ecosystem can evolve.

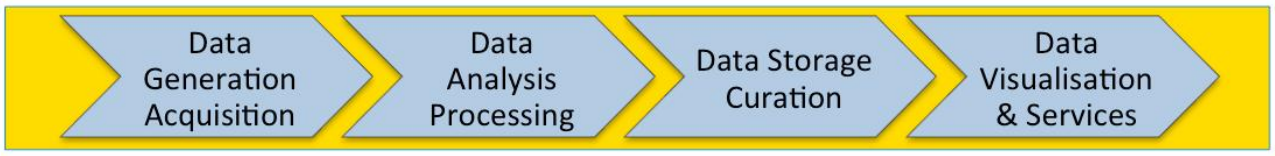

Figure 1: The Big Data Value chain[7,8]

There are strong companies in Europe that provide services and solutions along the Big Data Value chain. Some of them generate and provide access to huge amounts of data including structured and unstructured data. They acquire or combine real-time data streams from different sources, or add value by pre-processing, validating, augmenting data and ensuring data integrity. There are companies specialized in analysing data and recognizing correlations and patterns. Furthermore, some companies use these insights for predictions and decisions in various application domains.

However, despite the growing number of companies active in the data business, strengthening an economic Big Data Value ecosystem by bringing organisations together along the Big Data Value Chain at the European-level is required. Data usage is growing, but in both businesses and science, it is treated and handled in a fragmented way. To ensure a coherent use of data, a wide range of stakeholders along the Data Value chain need to be brought together to facilitate cooperation.

The stakeholders that will form the basis for interoperable data-driven ecosystems as a source for new businesses and innovations using Big Data are:

- Vendors of the ICT industry (Large and SME)
- Users across different industrial sectors (private and public)
- Big Data Value companies that do not exist yet and will emerge (start-ups)
- Researchers and academics who can provide knowledge and thought leadership

The cross-fertilisation involving many stakeholder and many datasets is a key element for advancing the Big Data economy in Europe.

1.3 The Multiple Dimensions of Big Data Value

In order to sustain the Big Data growth and remain competitive with other countries and regions, Europe needs to foster, strengthen and support the development and wide adoption of Big Data Value technologies, successful use cases and data-driven business models. At the same time, it is necessary to deal with many different aspects of an increasingly complex landscape. The main issues that Europe must tackle for creating and sustaining a strong Big Data ecosystem concern the following dimensions:

- **Data**: Availability of data and the access to data sources is paramount. There is a broad range of data types and data sources: structured and unstructured data, multi-lingual data sources, data generated from machines and sensors, data-at-rest and data-in-motion. Value is created by acquiring data, combining data from different sources, and providing access to it with low latency while ensuring data integrity and preserving privacy. Pre-processing, validating, augmenting data and ensuring data integrity and accuracy add value.

[7] "Competitive Advantage –Creating and Sustaining Superior Performance", Michael E. Porter, New York, 1998
[8] M. Cavanillas, E. Curry, W. Wahlster: New Horizons for a Data-Driven Economy – A Roadmap for Big Data in Europe, Springer International Publishing, 2016.

- **Skills**: In order to leverage the potential of Big Data Value, a key challenge for Europe is to ensure the availability of highly and rightly skilled people who have an excellent grasp of the best practices and technologies for delivering Big Data Value within applications and solutions. There will be the need for data scientists and engineers who have expertise in analytics, statistics, machine learning, data mining and data management. These experts will need to be combined with other experts having strong domain knowledge and the ability to apply this know-how within organisations for value creation.
- **Legal:** The increased importance of data will intensify the debate on data ownership and usage, data protection and privacy, security, liability, cybercrime, Intellectual Property Rights (IPR) and the impact of insolvencies on data rights. These issues have to be resolved in order to remove the adoption barriers. Favourable European regulatory environments are needed to facilitate the development of a genuine pan-European Big Data market.
- **Technical:** Key aspects such as real-time analytics, low latency and scalability in processing data, new and rich user interfaces, interacting with and linking data, information and content, all have to be advanced to open up new opportunities and to sustain or develop competitive advantages. Interoperability of data sets and data-driven solutions as well as agreed approaches is essential for a wide adoption within and across sectors.
- **Application:** Business and market ready applications should be the target. Novel applications and solutions must be developed and validated in ecosystems providing the basis for Europe to become the world-leader in the creation of Big Data Value.
- **Business:** A more efficient use of Big Data, and understanding data as an economic asset, carries great potential for the EU economy and society. The setup of Big Data Value ecosystems and the development of appropriate business models on top of a strong Big Data Value chain must be supported in order to generate the desired impact on the economy and employment.
- **Societal:** Big Data will provide solutions for major societal challenges in Europe, such as improved efficiency in healthcare information processing or reduced CO_2 emissions through climate impact analysis. In parallel, it is critical for an accelerated adoption of Big Data to increase awareness of the benefits and the Value that Big Data can create for business, the public sector, and the citizen.

Creating a favourable **ecosystem** for Big Data and pushing for its accelerated adoption requires an interdisciplinary approach addressing all of the aforementioned dimensions of Big Data Value.

1.4 The Big Data Value PPP (BDV PPP)

Europe must aim high and mobilise stakeholders in society, industry, academia and research to enable a European Big Data Value economy, supporting and boosting agile business actors, delivering products, services and technology, while providing highly skilled data engineers, scientists and practitioners along the entire Big Data Value chain. This will result in an innovation ecosystem in which value creation from Big Data flourishes.

To achieve these goals, the **European contractual Public Private Partnership** on **Big Data Value** (**BDV PPP**) was signed on 13 October 2014. This signature marks the commitment by the European Commission, industry and academia partners to build a data-driven economy across Europe, mastering the generation of value from Big Data and creating a significant competitive advantage for European industry, boosting economic growth and jobs. The **Big Data Value Association (BDVA)** is the private counterpart to the EU Commission to implement the BDV PPP program. BDVA has a well-balanced composition of large and small and medium-sized industries and enterprises as well as research organizations to support the development and deployment of the PPP work programme and to achieve the Key Performance Indicators (KPI) committed in the PPP contract. The BDV PPP has commenced in 2015 and was operationalized with the launch of the LEIT work programme 2016/2017. The BDV PPP activities will address technology and applications development, business model discovery, ecosystem validation, skills profiling, regulatory and IPR environment and social aspects. The BDV PPP will lead to a comprehensive innovation ecosystem for achieving and sustaining European leadership on Big Data, and for delivering maximum economic and societal benefit to Europe – its business and its citizens.

1.5 BDV PPP Vision for Big Data

Structured along the dimensions introduced in Section 1.3, the Big Data Value PPP vision for Europe in 2020 is:

- **Data:** Zettabytes of useful public and private data will be widely and openly available. By 2020, smart applications such as smart grids, smart logistics, smart factories, and smart cities will be widely deployed across the continent and beyond. Ubiquitous broadband access, mobile technology, social media, services, and IoT on billions of devices will have contributed to the explosion of generated data to a global total of 40 zettabytes. Much of this data will yield valuable information. Extracting this information and using it in intelligent ways will revolutionize decision-making in businesses, science, and society, enhancing companies' competitiveness and leading to new industries, jobs and services.
- **Skills:** Millions of jobs will have been established for data engineers and scientists, and the Big Data discipline is integrated into technical and business degrees. The European workforce is more and more data-savvy seeing data as an asset.
- **Legal:** Privacy & Security can be guaranteed along the Big Data Value chain. Data sharing and data privacy can be fully managed by citizens in a trusted data ecosystem.
- **Technology:** Real-time integration and interoperability among different multilingual, sensorial, and non-structured datasets is accomplished, and content is automatically managed and can be visualised in real-time. By 2020, European research and innovation efforts will have led to advanced technologies that make it significantly easier to use Big Data across sectors, borders and languages.
- **Application**: Applications using the BDV technologies can be built which will allow anyone to create, use, exploit and benefit from Big Data. By 2020, thousands of specific applications and solutions will address data-in-motion and data-at-rest. There will be a highly secure and traceable environment supporting organisations and citizens and having the capacity to support various monetization models.
- **Business:** A true EU single data market will be established allowing EU companies to increase their competitiveness and become world leaders. By 2020 Value creation from Big Data will have a disruptive influence on many sectors. From manufacturing to tourism, from healthcare to education, from energy to telecommunications services, from entertainment to mobility, Big Data Value will be a key success factor in fuelling innovation, driving new business models, and supporting increased productivity and competitiveness.
- **Societal:** Societal challenges are addressed through BDV systems, addressing the high data volume, the high motion of data, the high variety of data, etc.

These will impact the European Union's priority areas as follows:

- **Economy:** Competitiveness of European enterprises will be significantly higher compared to their worldwide competitors with improved products and services, and greater efficiency based on Big Data value. A true EU single data market will be established allowing EU companies to increase their competitiveness and become world leaders.
- **Growth:** There is a blossoming sector of growing new small and large businesses with a significant number of new jobs that create value out of data.
- **Society:** Citizens benefit from better and more economical services in a trustful economy where data can be shared with confidence. Privacy & security will be guaranteed throughout the lifecycle of BDV exploitation.

This foreseen evolution demands rethinking technologies around Big Data. Data collection, storage and processing must be improved to allow much more efficient access to data. Data visualisation and data analytics are also areas where new technologies will be needed. These technologies have different innovation cycles (in the range of months for services and applications, and years for ICT infrastructure) implying that architectures, technologies and standards cannot be designed based on pre-defined requirements. It is necessary to make challenging working assumptions on major basic technical requirements based on today's best knowledge to meet the needs expected in 2020.

Software-based systems provide the flexibility to adapt to new requirements introducing innovation into deployed systems, but the overall architecture and ICT infrastructures for storing and managing data do not offer this flexibility at present. Therefore, for the medium- to long-term perspective, future systems have to offer high flexibility and have to allow for high adaptability to new schemes.

1.6 BDV PPP Objectives

The objectives have been described in a more specific manner.

As laid out in the Contractual Arrangement (CA) of the BDV PPP[9], the overarching **general objectives** are:

- to foster European Big Data technology leadership for job creation and prosperity by creating a Europe-wide technology and application base and building up competence and the number of European data companies, including start-ups
- to reinforce Europe's industrial leadership and ability to compete successfully in the global data value solution market by advancing applications converted into new opportunities, so that European businesses secure a 30 % market share by 2020
- to enable research and innovation work, including activities related to interoperability and standardisation, for the future basis of big data value creation in Europe
- to facilitate the acceleration of business ecosystems and appropriate business models with a particular focus on SMEs, enforced by Europe-wide benchmarking of usage, efficiency and benefits
- to provide and support successful solutions for major societal challenges in Europe, e.g. in the fields of health, energy, transport and the environment, and agriculture
- to demonstrate the value of big data for businesses and the public sector and increase acceptance by citizens, by involving them as 'prosumers' and accelerating take-up
- to support the application of EU data protection legislation and provide for effective mechanisms to ensure its enforcement in the cloud and for big data

Complementing these general objectives, **specific objectives** on improved competitiveness, innovation, societal issues, and operational concerns have been agreed, these are:

Objectives for improved competitiveness:

- enable European suppliers to secure a 30% share of the global big data market by 2020
- develop solutions leading towards the use of big data value technology for increased productivity, optimised production, more efficient logistics (inbound and outbound) and effective service provision from public and private organisations
- implement Europe-wide strategic ('lighthouse') projects for specific reference deployments of existing or near-to-market technologies that demonstrate the potential impact of big data value creation across sectors
- create new big data ecosystems and markets between data providers, knowledge providers and consumers that will profit from collaboration between sectors, organisations and individuals
- develop and diffuse a better understanding of the business opportunities of the big data sector
- drive the take-up and integration of big data value services in private and public decision-making systems

Innovation objectives:

- optimise architectures for real-time analytics of data at rest and in motion, enabling data-driven decision-making 'on the fly' with low latency, and improve the scalability and processing of data validation and information discovery, especially in heterogeneous datasets

[9] http://ok-bdva.iais.fraunhofer.de/sites/default/files/BDVPPP_Contractual_Arrangement_.pdf

- validate technologies from a technical and business perspective through early trials in cross-organisational, cross-sector and cross-lingual innovation environments
- integrate advanced visualisation of data and analytics for augmented user experience and prepare platforms, technologies and tools for disruptive changes in the management of data
- develop and provide validated technology and tools for 'deep data analysis' to improve the understanding, deep learning and meaningfulness of data, by raising awareness of the importance of data definition for optimal information content
- structure a Big Data cluster and value chain that will allow for future coordination of actors (suppliers, R&D centres, Large and Small Industry, NGOs, public actors ...)

Societal objectives:

- support widespread know-how, education and skills in Europe through curricula to stimulate higher education provision taking into account and establishing appropriate collaboration links with the "Grand Coalition for Digital Job provision"[10], and making appropriate use of the "European Guidelines and Quality Labels for new Curricula Fostering e-Leadership Skills"[11], and the European eCompetence Framework[12]
- increase the number of European data workers by 100 000 by 2020
- face Europe's societal challenges through 'lighthouse' projects and i-spaces in areas such as personalised medicine
- create new personalised and enhanced products and services adapted to citizens' and organisations' needs that will respect security and ensure privacy and personal data protection of individuals in the framework of relevant EU rules
- foster trust in the data-driven economy, including through incentivizing the application of the principles of privacy and security by design as well as the cooperation with relevant authorities in case of data breaches and cyber incidents
- address European framework aspirations such as IPR rights, liability, etc. within the Digital Single Market and pan-European innovation environments
- address acceptance of new Big Data technologies by society and Consumers by identifying and removing potential barriers, raising awareness, and ensuring that individuals are appropriately informed on the use of their personal data in compliance with EU data protection legislation.

To achieve the general and specific objectives of the BDV PPP, the research and innovation strategy requires dedicated actions and mechanisms along the multiple dimensions of Big Data Value (see Section 1.3). This SRIA aims to provide further details on the key building blocks and actions for such research and innovation strategy.

1.7 BDV SRIA Document History

To establish a contractual counterpart to the European Commission for the implementation of the PPP, the Big Data Value Association, a fully self-financed non-for-profit organisation under Belgian law, was founded by 24 organisations including large, SMEs and research organisations. The objectives of the Association are to boost European Big Data Value research, development and innovation and to foster a positive perception of Big Data Value. As of Jan 2017 the BDVA has 158 members representing Big Data Value stakeholders from across the European Union.

This **Strategic Research and Innovation Agenda (SRIA)** defines the main technical and non-technical priorities to achieve the BDV PPP objectives (see Section 1.6), and it describes a research and innovation roadmap for the BDV PPP. The BDV SRIA was prepared using an extensive process that has heavily engaged with the wider Big Data Value community. Wide ranges of stakeholders have contributed to the

[10] https://ec.europa.eu/digital-agenda/en/grand-coalition-digital-jobs
[11] http://www.eskills-guide.eu/home/
[12] http://www.ecompetences.eu/

SRIA in different forms of engagement (see Appendix 6.4). It is built upon inputs and analysis from SMEs and Large Enterprises, public organisations, and research and academic institutions. Stakeholders include suppliers and service providers, data owners, and early adopters of Big Data in many sectors. The process included multiple workshops and consultations to ensure the widest representation of views and positions including the full range of public and private sector entities. These have been carried out to identify the main priorities with approximately 200 organisations and other relevant stakeholders physically participating and contributing. Extensive analysis reports were then produced which helped both formulate and construct this SRIA. A compilation of the workshop results is provided as an integrated SWOT analysis for the European market in Appendix 6.4.

The BDVA is responsible for providing regular (yearly) updates of the SRIA to ensure it remains relevant to the priorities of the community. Within the update process, the BDVA engages with the BDVA members as well as the wider community to ensure a comprehensive perspective concerning the technical and business impact of the SRIA technical and non-technical priorities as well as to identify emerging priorities with high impact. Further details on the SRIA Update process can be found in Appendix 6.3

2 Implementation Strategy

Given the broad range of objectives around the many aspects of Big Data Value, a complete implementation strategy is needed. In this section, we set out such a strategy that is the result of a very broad discussion process involving a large number of relevant European BDV stakeholders.

The result is an interdisciplinary approach that integrates expertise from the different fields necessary to tackle both the strategic and specific objectives. To this end, European cross-organisational and cross-sector environments have to be incubated, such that large enterprises and SMEs alike will find it easy to discover economic opportunities based on data integration and analysis, and then develop working prototypes to test the viability of actual business deployments.

The growing number and complexity of valuable data assets will drive existing and new research challenges. Cross-sectorial and cross-organisational environments will enable research and innovations in new and existing technologies. Business applications that need to be evaluated for usability and fitness for purpose can be deployed within these environments ensuring the practical applicability of the applications. This, in turn, will require validations, trials and large-scale experiments in existing or emerging business fields, the public sector, industry, and jointly with end-user and individual consumers.

To support validations, trials and large-scale experiments, access to valuable data assets needs to be provided with low obstacles in environments that simultaneously support legitimate ownership, privacy and security related to data owners and their customers. These environments will ease experimentation for researchers, entrepreneurs, SMEs and large ICT providers.

2.1 Four kinds of mechanisms

In order implement the research and innovation strategy, and to align technical with cooperation and coordination aspects four major types of mechanisms are recommended:

- **Innovation Spaces** (i-Spaces): Cross-organisational and cross-sectorial environments – will allow challenges to be addressed in an interdisciplinary way and will serve as a hub for other research and innovation activities.
- **Lighthouse projects:** These will help raise awareness of the opportunities offered by Big Data and the value of data-driven applications for different sectors and will act as an incubator for data-driven ecosystems.
- **Technical projects:** These will take up specific Big Data issues addressing targeted aspects of the technical priorities as defined in Section 3.
- **Cooperation and coordination projects:** These projects will foster international cooperation for efficient information exchange and coordination of activities.

2.1.1 European Innovation Spaces (i-Spaces)

Definition of i-Spaces was updated in accordance to the discussion in the I-Spaces Taskforce.

Extensive consultation with many stakeholders of relevant areas related to Big Data Value (BDV) has confirmed that besides technology and applications, a number of key issues have to be considered. First, infrastructural, economic, social and legal issues have to be addressed. Second, the private and the public sector will have to be made aware of the benefits that BDV can provide, thereby motivating them to be innovative and to adopt BDV solutions.

To address all these aspects, European cross-organisational and cross-sectorial environments, which rely and build upon existing national and European initiatives, will play a central role in a European Big Data ecosystem. These so-called ***European Innovation Spaces*** (or ***i-Spaces*** for short) are the main elements to assure that research on BDV technologies and novel BDV application will be quickly tested, piloted and thus exploited in a context with maximum involvement of all stakeholders of BDV ecosystems. As such, i-Spaces will enable stakeholders to develop new businesses facilitated by advanced BDV technologies, applications, and business models. They contribute to the building of a community and catalyse this community engagement. That act as incubators and accelerators of Data-Driven Innovation.

In this sense, i-Spaces are hubs to unite technical and non-technical activities, for instance by bringing technology and application development together with the development of skills, competence, and best practices. To this end, i-Spaces will offer both state-of-the-art as well as emerging technologies and tools from industry and open source software initiatives, they will also provide access to data assets. By doing so, i-Spaces will foster community building and interdisciplinary approach for solving BDV challenges along the core dimensions of technology, applications, legal, social, and business, data assets and skills.

The creation of i-Spaces will be driven by the needs of large and small companies alike to ensure they easily discover the economic opportunities based on BDV and develop working prototypes to test the viability of actual business deployments. This does not necessarily require moving data assets across borders. Rather data analytic tools and computation activities could be brought to the data. Thereby, valuable data assets are made available in environments that simultaneously support the legitimate ownership, privacy and security policies of corporate data owners and their customers, while facilitating ease of experimentation for researchers, entrepreneurs and small and large IT providers.

Concerning the discovery of value creation, i-Spaces will support various models: at one end, corporate entities with valuable data assets will be able to specify, business relevant data challenges for researchers or software developers to tackle; at the other end, entrepreneurs and companies with business ideas to be evaluated, will be able to solicit the addition and integration of desired data assets from corporate or public sources. I-Spaces contribute also to fill the skills gap in Europe is facing in providing (controlled) access to real use cases and data assets to education and all skills improvement initiatives.

The i-Spaces themselves will be data-driven both at the planning and at the reporting stage. At the planning stage, they will prioritise the inclusion of data assets that, in conjunction with existing assets, present the greatest promise for European economic development (while taking full account of the international competitive landscape); at the reporting stage, they will provide methodologically sound quantitative evidence on important issues such as increases in performance for core technologies or reduction in costs for business processes. These reports will foster learning and continuous improvement for the next cycle of technology and applications.

The particular European value-add of i-Spaces is that they will federate, complement and leverage activities of similar national incubators/environments, existing PPPs and other national or European initiatives. With the aim of not duplicating existing efforts, complementary activities considered for inclusion will have to stand the test of expected economic development: new data assets and technologies will be considered for inclusion to the extent that they can be expected to open new economic opportunities when added to and interfaced with the assets maintained by regional or national data incubators or existing PPPs.

The successive inclusion of data assets into i-Spaces will, in turn, drive and prioritise the agenda for addressing data integration or data processing technologies. One example is the existence of data assets of homogenous qualities (such as geospatial, time series, graphs and imagery), which requires optimising the performance of existing core technology (such as querying, indexing, feature extraction, predictive analytics and visualization). This requires methodologically sound benchmarking practices to be carried out in

appropriate facilities. Similarly, business applications exploiting BDV technologies will be evaluated for usability and fitness for purpose, thereby leading to continuous improvement of these applications.

Due to the richness of data that i-Spaces will offer, as well as access to a large variety of integrated software tools and expert community interactions, the data environments will provide the perfect setting for the effective training of data scientists and domain practitioners. They will encourage a broader group of interested parties to engage in data activities. These activities will be designed to complement the educational offerings of established European institutions.

While economic development is the principal objective of BDV, this cannot happen without taking into proper account the legislative requirements pertaining to the treatment of data, as well as ethical considerations. In addition, BDV will create value for society as a whole by systematically supporting the transfer of sophisticated data management practices to domains of societal interest such as health, environment, or sustainable development, among others. Especially when it comes to SMEs, the issues of skills and training, reliable legal frameworks, reference applications and access to an ecosystem become central for a fast take-up of the opportunities offered by BDV. In this holistic interdisciplinary approach, i-Spaces will be a key mechanism that targets BDV challenges along the dimensions as depicted in Figure 2. The i-Spaces will be instrumental to test, showcase and validate new technology, applications and business models. The central need for availability of open and industrial data assets will be catered for as well as for skills development, best practices identification, requirements for favourable legal, policy and infrastructural frameworks and tools across sectors and borders.

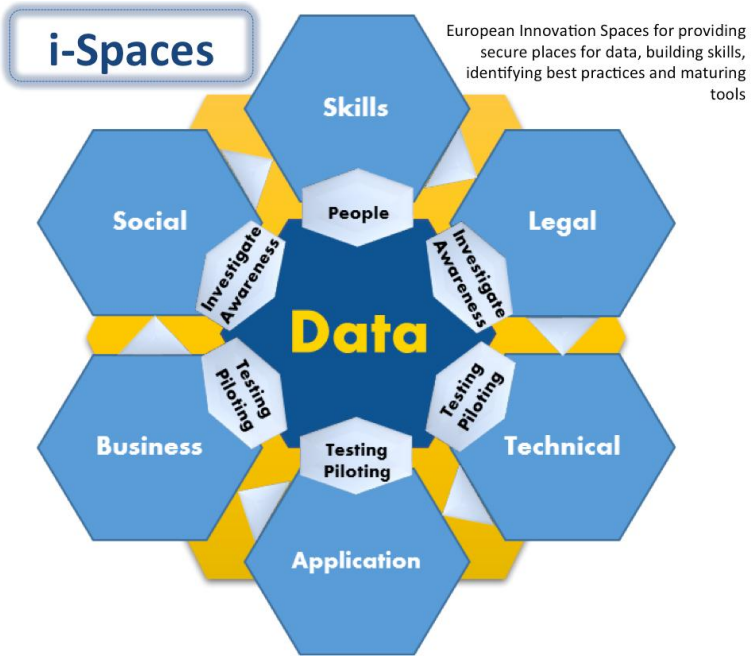

Figure 2: Interconnected Challenges of the PPP within i-Spaces

All i-Spaces will provide a set of basic services to support Lighthouse Projects, Technical Projects, as well as Collaboration and Coordination Projects. These basic services include:

- **Community Building:** Contribution to identification and management of stakeholder ecosystem communities on thematic and/or regional dimensions. This activity capitalizes on existing initiatives thematic and/or regional
- **Asset Support:** Supporting data providers in integrating data sets in a quality-secured way while maintaining a catalogue of available data assets.
- **ICT Support:** Providing basic ICT assistance as well as focused support by Big Data scientists, data specialists, and business development during research and innovation projects. This includes assistance in benchmarking data sets, technologies, applications, services, and business models.
- **On-boarding**: Running an induction process for new project teams.
- **Resourcing:** Allocating the resources (computing, storage, networking, tools, and applications) to individual research and innovation projects and scheduling these resources among different projects.

- **Protection:** Data protection including ensuring the compliance with laws and regulation as well as the deployment of leading-edge state-of-the-art security technologies in protecting data and controlling data access.
- **Privacy:** Data privacy and anonymisation with handling and deletion of personally identifiable information (PII) in compliancy with laws and regulations such as the EU GDPR (General Data Protection Regulation) - as well as the deployment of anonymisation technologies for preventing processing of PII when necessary.
- **Data Governance :** taking into account Privacy and Protection define the rules to access and share data. This includes standardization of sharing metadata, definition of the (smart) contract between stakeholders, technologies such as encryption and probably blockchain , as well as the necessary solutions to orchestrate the agreed governance
- **Federation:** Supporting linkages to other innovation spaces and facilitating experiments across multiple innovation spaces. An effective federation will help to support research and innovation activities accessing and processing data assets across national borders (data spaces).
- **Business support:** Facilitating Start-ups and SME inclusion in the value creation process in leveraging community engagement
- **Incubation and Acceleration:** Delivering all forms of adequate support to Data Driven value creation projects in liaison with existing initiatives thematic, National or regional.

The i-Spaces should also be understood as incubator environments, where research outcomes into novel technologies and applications can be quickly tested and piloted in a context with the maximum involvement of stakeholders in the ecosystems including business innovators and end-users.

Summarizing, the main characteristics of the i-Spaces are:

- Being the **hubs for bringing technology and application developments together** and cater for the development of skills, competence, and best practices. The environments will offer new and existing technologies and tools from industry and open source software initiatives as a basic service to tackle the Big Data Value challenges.
- Ensuring that **data is at the centre** of Big Data Value activities. The i-Spaces will make accessible data assets based on industrial, private and open data sources. i-Spaces will be secure and safe environments that will ensure the availability, integrity, and confidentiality of the data sources.
- Serving as **incubators for testing and benchmarking** of technologies, applications, and business models. This will provide early insights on potential issues and will help to avoid failures in the later stages of commercial deployments. In addition, it can be expected that this activity will provide **input for standardization and regulation**.
- **Developing skills and sharing of best practices** will be an important task of the i-Spaces and their federation, they will also link with other existing initiatives at European and national-levels.
- New **Business Models and Ecosystems** will be emerging from exposing new technologies and tools to industrial and open data. The i-Spaces will be the playground to test new business model concepts and emerging ecosystems of existing and new BDV "players".
- Getting early insights into the **social impact** of new technologies and data-driven applications and how they will change the behaviour of individuals and the characteristics of data ecosystems.
- Act as a catalyser to foster data driven **communities** in the ecosystem and accelerate value creation.

Setup of i-Spaces

To ensure that i-Spaces will achieve their ambitious objectives, the following design considerations will be taken into account when setting up i-Spaces:

- Strong Relation to **Data Innovation ecosystem** in particular to industrial and institutional Data Owners
- Availability of a team providing basic IT **assistance**, as well as focused **support** by Big Data experts.
- **Business Development** resources to initiate and materialize projects taking into account value, legal and technical dimensions
- A well-managed **IT infrastructure**, including remote access capabilities.
- **Secure and trustworthy** data hosting and access.
- **Project Management** resources to ensure delivery.

Key elements for the implementation of i-Spaces include at least the following:

- **A multidisciplinary team** able to manage the value creation process from community building to the final delivery of projects. In particular, this can include coaching of status and SMEs participating in I Spaces Activities and identify and coach potential technology transfer from research. This is performed in full coordination with existing initiatives thematic or regional
- **Secure access to data storage** that provides the necessary security mechanisms needed by industry, and other data asset owner, to trust sharing their data assets with scientists and data specialists for experimentation. At the same time, open data will be made available. Support for running experiments on site or remotely as well as data governance mechanisms for leveraging the different rights and duties (roles) within a data space.
- **Offering hybrid-computing models**. The Cloud paradigm will be one important computing model for Big Data Value technology and thus i-Spaces. Yet, it will not be the only model. For instance, due to the volume and velocity of data, transferring this data from data sources (such as IoT sensors) to Cloud providers might not be feasible. This means that i-Spaces infrastructures will also consider other computing models, such as "distributed computing", "high-performance computing", as well as "computing at the edge".
- **Delivering platforms and tools** from different sources, including open source and proprietary, for enabling data scientist and data engineers to develop and run new technology and applications. It is envisaged that i-Spaces will start from the "state-of-the-art" and continuously evolve, incorporating new technology as it becomes available. Contribution to establishment of data life cycle management and data organization to enable methods for data preservation and curation as well as data sharing.
- **Tool for continuous benchmarking** so that businesses, and in particular start-ups and SMEs, can evaluate whether their products and services will work in a real-world context.
- **Tools for anonymization** in order to protect the identification of sensitive information which may interfere with the right to secrecy (e.g. industrial trade information) and privacy.
- **Business (model) benchmarking.** This type of benchmarking may among other aspects focus on process, financial or investor perspective aspects.
- **Technical benchmarking.** This type of benchmarking is about determining how the performance or operational cost of a product or service compares with existing products or services.
- **User experience benchmarking.** Besides performance and cost, the customer and user experience of a product and service is key for success. The quality of the user centric approach when it comes to products and services is vital for it to become a success.
- **Data set benchmarking.** The data sets are at the core of i-Spaces. Measuring and ensuring data quality, not only on existing data sets, but also on live data streams, is the main concern to this strand.

2.1.2 Lighthouse projects

Information about the future nature of Lighthouse projects has been added.

Lighthouse projects[13] are Innovation projects with a high degree of innovation, **running large-scale data-driven demonstrations** whose main objectives will be to create high-level impact, and broadcast visibility and awareness driving towards faster uptake of Big Data Value applications and solutions.

Lighthouse projects will be the major mechanism to demonstrate Big Data Value ecosystems and sustainable data marketplaces that lead to increased competitiveness of established sectors as well as the creation of new sectors in Europe. Lighthouse projects will propose replicable solutions by using existing technologies or very near-to-market technologies that could be integrated in an innovative way and show evidence of data value.

Lighthouse projects shall lead to explicit business growth and job creation and thus, all projects will be required to define clear indicators and success factors that can be measured and assessed in both qualitative and quantitative terms against those goals.

Increased competitiveness is not only a result of the application of advanced technologies; it results from a combination of changes that expands the technological level, adding political and legal decisions, among others. Thus, Lighthouse projects are expected to bring a combination of decisions centred on data, including the use of advanced big data-related technologies but also in other dimensions. Their main purpose will be to make visible results to a wide and high-level and high-breath audience thus accelerating changes, therefore making explicit and visible the impact of big data in a specific sector, and / or a specific economic or societal ecosystem.

Lighthouse projects will be defined through a set of well-specified goals that will materialize through large-scale demonstrations deploying existing and near-to-market technologies. Projects may include a limited set of research activities if that is needed to achieve the goals, but it is expected that the major focus will be on data integration and solution deployment.

Lighthouse projects are different than proof-of-concepts (which are more related to technology or process) or pilots (which are usually an intermediate step on the way to full production): they should pave the way for faster market roll-out of technologies, should be conducted at a large scale and should use their success to rapidly transform the way an organization thinks or processes are run.

Sectors or environments to be included are not pre-decided but should be in line with the aforementioned impact. For example, deployment of new e-health services along the EU through a large-scale implementation of EHR[14] is aligned with the expected impact; however, this would require not only application of some big data technologies but also pushing for political and regulatory decisions in the fields of Data privacy and interoperability.

The first call for lighthouse projects in the context of this PPP resulted in two actions in the domains of Bioeconomy (including agriculture, fisheries and forestry) and Transport and logistics. Therefore, even though additional use cases in those domains could be valuable, we aim at diversifying the sectorial approach of the PPP and ensure that benefits of Big Data technologies expand over different industries. That is why it is recommended that future lighthouses do not repeat these sectors and address specifically the challenges defined by stakeholders in other high potential data-intensive industries. A short overview of these projects is provided at the end of this section.

Lighthouse projects will operate primarily in a single domain, where a meaningful (as documented by total market share) group of EU industries from the same sector will jointly provide a safe environment in which they will make available a proportion of their data (or data streams) and demonstrate, in a large scale, the impact of big data technologies. It is expected that projects use data sources other than the ones of the specific sector addressed, therefore contributing to break silos. In all cases projects are supposed to have access to appropriately large, complex and realistic data sets.

[13] sometimes also labeled as large scale demonstration or pilots
[14] EHR stands for Electronic Health Record

One of the expected outcomes of this approach is data interoperability. Solutions at EU level (i.e. going beyond national boundaries) and that avoid vendor lock-in will be especially desired in an attempt to reach economies of scale.

Projects will be asked to show sustainable impact beyond the specific large-scale demonstrator/s running along the project duration. This should be done when possible through solutions that can be replicable by other companies in the sector or by other application domains.

All lighthouse projects have to involve, as appropriate, the relevant stakeholders to reach their goals. As a result, it is expected that complete data ecosystems will be developed. When needed, Lighthouse projects may use the infrastructure and ecosystems facilitated by one or more i-Spaces. .

Some of the indicators that will be used to assess the impact of Lighthouse projects will be the number and size of data sets processed (integrated), the number of data sources made available for use and analysis by third parties or the number of services provided for integrating data across sectors. Market indicators will obviously be of utmost importance. Lighthouse projects are expected to contribute to:

- a 20% increased market share in the corresponding sector through selling integrated data and/or data integration services
- the establishment of cross-sectorial standards for data sharing on EU level when applicable

They should make clear the nature of those contributions.

Key elements for the implementation of Lighthouse projects include at least the following:

Use of existing or close-to-market technologies: Lighthouses are not expected to develop completely new solutions; instead, they should make use of existing or close-to-market technologies and should accelerate the roll-out of those ones. Solutions should give answer to real needs and requirements, showing explicit knowledge of the demand side. Even though projects should concentrate on solving concrete problems and this may lead to specific deployments, replicability of concepts should be a priority to ensure impact beyond the particular deployments of the project. Lighthouse projects should address frameworks and tools from a holistic perspective and for e.g. do not only consider analytics but the complete data value chain (data generation, extension of data storing, analyzing, etc.).

Interoperability and Openness: Projects should take advantage of both closed and open data; they can also decide if open source or proprietary solutions are the most suitable to address their challenges. However, they should promote interoperability of solutions in order to avoid locked –in customers.

Some references of solutions that follow that open approach already exist, such as the so-called lighthouse projects funded under the EIP on SCC[15] or the projects within the Future Internet PPP. Of particular relevance are initiatives like OASC[16], based on a lightweight and pragmatic approach to rollout smart city solutions in real environments that do not tie city developments to a specific vendor or technology. It supports Open APIs as well as a driven-by-implementation approach regarding data models. It capitalizes on existing works by adopting a first set of data models; then, further curation of these data models can take place based on feedback from actual usage or experimentation.

Proposals are encouraged to work with open specifications that allow different stakeholders to make their solutions interoperable and compatible with the proposed solutions. The focus should be on the creation of markets in line with the DSS and take advantage of economies of scale. As aforementioned, Pan-European solutions may require in many sectors going beyond pure technological decisions.

Involvement of smaller actors (for example, through opportunities for start-ups and entrepreneurs) that can compete in the same ecosystem in a fair way should be a must. Open APIs could play an important role here (e.g. third party innovation through data sharing). In addition, projects should focus on the usability and reduce barriers/gap from big data methods to end users (break the "BD for data analysts only" paradigm).

Performance: Proposals should contribute to common data collection systems and have a measurement methodology in place. Performance monitoring should last at least 2/3 of the duration of the project.

[15] EIP SCC stands for European Innovation Partnership on Smart Cities and Communities
[16] OASC stands for Open and Agile Smart Cities Alliance

However, longer-term commitment will give value to the proposal. For further information on quantitative impact check WP2016-17.

Set-up of ecosystems: Lighthouses should have a transformational power, i.e. they should not restrict to a very narrow-minded experiment with limited impact. They should demonstrate that they are able to improve (sometimes changing associated processes) the competitiveness of the selected industrial sector in a relevant way. This requires the active involvement of different stakeholders and therefore, attention should be paid to the ecosystem that will enable such changes. Lighthouse projects should be connected from the design phase to communities of stakeholders. Ecosystems should evolve, extend or connect existing networks of stakeholders and hubs.

As it is well known, European industry is characterized by a huge number of small and medium enterprises; lack of consideration of this factor would lead to a non-healthy environment. Thus, adequate consideration of SME integration in the projects is required.

Even though proposals should focus on a sector, use of data from different sources and industrial sectors should be encouraged and priority should be given to avoid the "silo" effect.

Long-term commitment and sustainability: Budgets assigned to the projects should be a seed for a wider implementation plan. It is expected that the proposed activities are integrated into a more ambitious strategy where additional stakeholders and additional funds (preferably private but also possible a combination of public and private) are involved.

As it was aforementioned, two lighthouses have already been selected and recently launched, leading to an evolution of the concept. That is why this updated version of the SRIA suggests more concrete requirements for the upcoming large-scale pilots, in some cases further specifying aspects that were already worked out. See the following list as a guidance and not as a complete list:

- Reuse of technologies and frameworks that are already in the market or close to it (high TRL) in an attempt to avoid the development of new platforms if a good basis already exist (for example, as part of the Open Source community). Projects are especially encouraged to build on top of technologies created by the ongoing projects of the Big Data PPP that fit their requirements (for example, in the area of privacy-preserving technologies)
- Special attention should be paid to interoperability. This applies to all layers of the solution, including data (here, some of the results of the projects funded under the Big Data PPP with a focus on data integration could be particularly useful)
- In January 2017 several Large Scale pilots have been launched in particular technology areas such as IoT (in this case, pilots will run in 5 different domains). The next generation of big data lighthouses is expected to search for synergies with those pilots when it is appropriate and fits the purpose. Some of these synergies could rely on sharing data, bringing solutions developed in one project to the testing environment of the other or taking advantage of the domain communities in the different projects to organize joint dissemination activities or replicate some of the experiments; this list illustrates some examples of potential collaboration but it is by no means complete). Projects are also encouraged to establish tight collaborations with relevant pilots funded under Societal Challenges when they are aligned with their objectives.
- It is expected that projects combine the use of open and close data. While it is understandable that some close data will remain as such, we also expect those projects to contribute to increase the availability of data sets that could be used by other stakeholders, such as SMEs and startups. This could happen under different regimes (not necessarily for free). Projects should state the way they will contribute to this objective by quantifying and qualifying data sets (when possible) and including potential contributions to the ongoing data incubators/accelerators and Innovation Spaces.
- Lighthouse projects have to contribute to horizontal activities of the Big Data PPP as a way to help in the assessment of the PPP implementation and increase its potential impact. Some of the targeted activities include contribution to standardization activities, measurement of KPIs, and coordination with the PPP branding or active participation to training/educational activities proposed by the PPP.

Overview of existing lighthouses in the Big Data Value PPP

Transforming Transport: Big Data Value in Mobility and Logistics

Big Data will have a profound economic and societal impact in the mobility and logistics sector, which is one of the most-used industries in the world contributing to approximately 15% of GDP. Big Data is expected to lead to 500 billion USD in value worldwide in the form of time and fuel savings, and savings of 380 megatons CO2 in mobility and logistics. With freight transport activities projected to increase by 40% in 2030, transforming the current mobility and logistics processes to become significantly more efficient, will have a profound impact. A 10% efficiency improvement may lead to EU cost savings of 100 BEUR. Despite these promises, interestingly only 19 % of EU mobility and logistics companies employ Big Data solutions as part of value creation and business processes.

The Transforming Transport project will demonstrate, in a realistic, measurable, and replicable way the transformations that Big Data will bring to the mobility and logistics market. To this end, Transforming Transport validates the technical and economic viability of Big Data to reshape transport processes and services to significantly increase operational efficiency, deliver improved customer experience, and foster new business models. Transforming Transport will address seven pilot domains of major importance for the mobility and logistics sector in Europe: (1) Smart High-ways, (2) Sustainable Vehicle Fleets, (3) Proactive Rail Infrastructures, (4) Ports as Intelligent Logistics Hubs, (5) Efficient Air Transport, (6) Multi-modal Urban Mobility and (7) Dynamic Supply Chains.

Data Bio: Data-Driven Bioeconomy

The data intensive target sector of DataBio is the Data-Driven Bioeconomy, focusing on production of best possible raw materials from agriculture, forestry and fishery/aquaculture for the bioeconomy industry to produce food, energy and biomaterials taking into account also various responsibility and sustainability issues. Experiences from US show that bioeconomy and specifically agriculture can get a significant boost from Big Data. In Europe, this sector has until now attracted few large ICT vendors. A central goal of DataBio is to increase participation of European ICT industry in development of Big Data systems for boosting the lagging bioeconomy productivity. For this, DataBio proposes to deploy a state of the art, big data platform "on top of the existing partners' infrastructure and solutions - the Big DATABIO Platform. The selected pilots include agriculture (precision horticulture including wines and olives, arable precision farming, and subsidies and insurance), forestry (data crowd-sourcing services, forest health, forest data management services) and fisheries (fishing vessels operational choices, fishing vessel trip and fisheries planning, fisheries sustainability and value).

2.1.3 Technical projects

Technical projects focus on addressing one or few specific aspects identified as part of the BDV technical priorities (also see Section 3). Thereby, technical projects provide the technology foundation for lighthouse projects and i-Spaces. Technical projects may be implemented as Research and Innovation Actions (RIAs) or Innovation Actions (IAs) depending on the amount of research work required to address the respective technical priorities.

2.1.4 Cooperation and coordination projects

Cooperation and coordination projects[17] will work on detailed activities that ensure coordination and coherence in the PPP implementation and will provide support to activities that fall under the skill, business, policy, regulatory, legal and social domains.

[17]for instance Collaboration and Support Actions (CSAs)

2.2 BDV Methodology

The programme will develop the ecosystem in distinct phases of development, each with a primary development theme. The three phases as depicted in Figure 3 are:

- **Phase I:** Establish ecosystem (Governance, i-Space, education, enablers) and demonstrate value of existing technology in high-impact sectors (Lighthouses, Technical Projects) (Work Programme WP 16 - 17)
- **Phase II:** Pioneer disruptive new forms of big data value solutions (Lighthouses, Technical Projects) in high-impact domains of importance for EU industry, addressing emerging challenges of the data economy (WP 18-19)
- **Phase III:** Long-term ecosystem enablers to maximise sustainability for economic and societal benefit (WP 20 –)

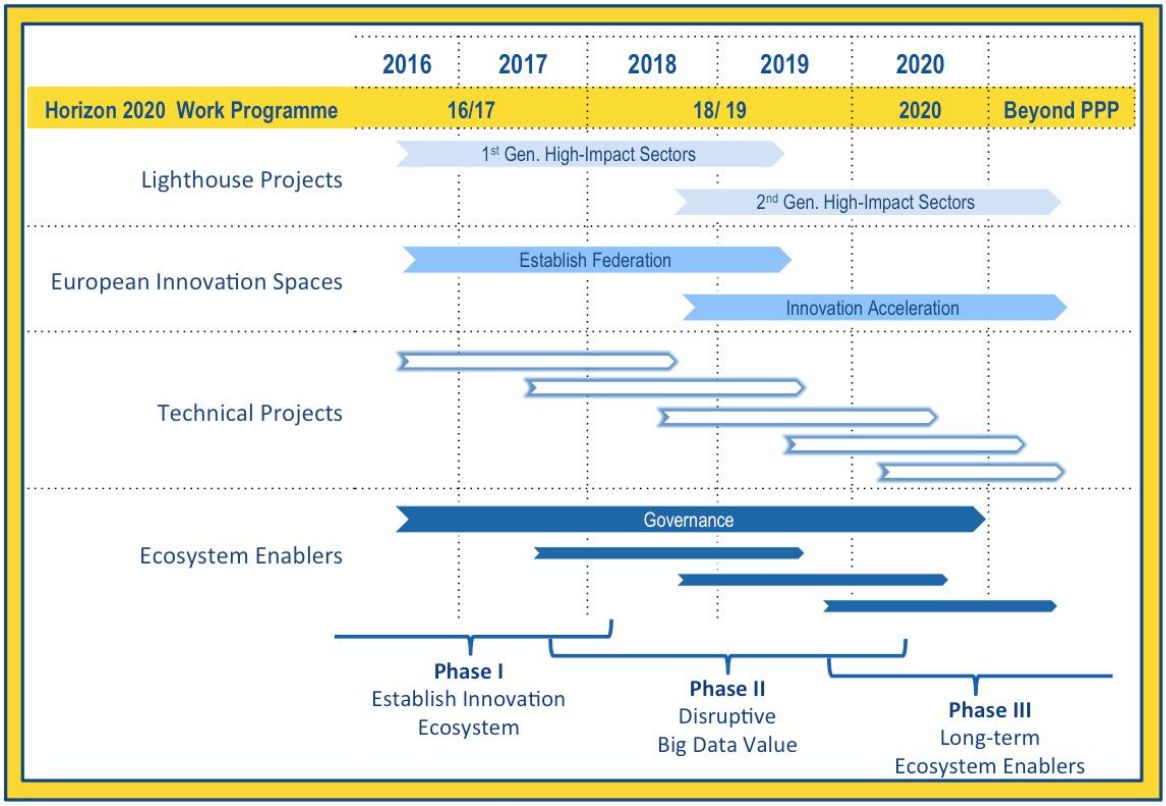

Figure 3: Three-Phase Timeline of the BDV PPP

Phase I: Establish Innovation Ecosystem (2016-2017)

The first phase of the programme will focus on laying the foundations necessary to establish a sustainable European data innovation ecosystem. The key activities of Phase I include:

- Establish a European network of I-Spaces for cross-sectorial and cross-lingual data integration, experimentation, and incubation (ICT14 – 2016-17)
- Demonstrate big data value solutions via Large Scale Pilot projects in domains of strategic importance for EU industry using existing technologies or very near-to-market technologies (ICT15 – 2016-17)
- Tackle the main technology challenges of the data economy by improving the technology, methods, standards and processes for Big Data Value (ICT16 – 2017)
- Advance the state of the art in privacy-preserving big data technologies and explore societal and ethical implications (ICT18 – 2016)
- Establish key ecosystem enablers including programme support and coordination structures for industry skills and benchmarking (ICT17 – 2016-17)

Phase II: Disruptive Big Data Value (2018-2019)

Building on the foundations established in Phase I, the second phase will have a primary focus on Research and Innovation activities to deliver the next generation of Big Data Value solutions. The key activities of Phase II include:

- Innovation projects within a federation of i-Spaces to validate and incubate innovative Big Data Value Solutions and business models with particular focus on accelerating SMEs, start-ups, and entrepreneurs, emerging, and best practices.
- Pioneer disruptive new forms of big data value solutions via Large Scale Pilot projects in emerging domains of importance for EU industry using new platforms, novel tools, and advanced methodologies.
- Tackling the next generation of research and innovation challenges for Big Data Value Solutions as detailed in the updated BDV SRIA.
- Address legal and societal roadblocks and inhibitors for take-up of Big Data Value solutions and big data ecosystem viability including business models.
- Programme support (continuation), networking and cooperation among ecosystem actors and projects. Community building to attract key stakeholders for Phase III. Big data innovative business models and mechanisms to accelerate innovation to the market.

Phase III: Long-term Ecosystem Enablers (2019-)

While the sustainability of the ecosystem has been considered from the start of the PPP, the third phase will have a specific focus on activities that can ensure long-term self-sustainability. The key activates of Phase III include:

- Seed the long-term ecosystems enablers to ensure self-sustainability beyond 2020.
- Ensure continued support for technology outputs of PPP (Lighthouses, R & I, CSA) including non-technical aspects (training) beyond 2020. (i.e. Open Source Community, Technology Foundation).
- Establish a Foundation for European Innovation-Spaces with a charter to continue collaborative innovation activity beyond 2020.
- Liaise with private funding (VCs) to accelerate market introduction and socio-economic impacts including support services to develop Investment ready proposals and support scaling for BDV PPP Start-ups and SMEs to reach the market.
- Tackle the necessary strategy and planning for the BDV Ecosystem until 2030, including the identification of new stakeholder, emerging usage domains, technology, business, and policy road-mapping activity.

2.3 Funded Projects

This section incorporating information about funded projects in 2016 is new in version SRIA 3.0.

The first set of projects from Phase I of the programme have been funded in the calls from 2016. A total of 16 projects in the areas iSpaces (8 projects), Lighthouses (2 projects), Ecosystem Enablers (1 Project), and Privacy-preserving Big Data Technologies (4 projects) will commence their activity in 2017

Project Name	Number	Instrument	Start	End
iSpaces (ICT-14-2016-2017)				
BigDataOcean: Exploiting Ocean's of Data for Maritime Applications	732310	IA	01/01/17	30/06/19
SLIPO: Scalable Linking and Integration of Big POI data	731581	IA	01/01/17	31/12/19
Data Pitch: Accelerating data to market	732506	IA	01/01/17	31/12/19
AEGIS: Advanced Big Data Value Chain for Public Safety and Personal Security	732198	IA	01/01/17	30/06/19
euBusinessGraph: Enabling the European Business Graph for Innovative Data Product and Services	732003	IA	01/01/17	30/06/19

QROWD: Because Big Data Integration is Humanly Possible	732194	IA	01/12/16	30/11/19
FashionBrain: Understanding Europe's Fashion Data Universe	732328	IA	01/01/17	31/12/19
EW-Shopp: Supporting Event and Weather-based Data Analytics and Marketing along the Shopper Journey	732590	IA	01/01/17	31/12/19
Lighthouses (ICT-15-2016-2017)				
DataBio: Data-Driven Bioeconomy	732064	IA	01/01/17	31/12/19
TT: Transforming Transport IA	731932	IA	01/01/17	30/06/19
Ecosystem Enablers (ICT-17-2016-2017)				
BDVe: Big Data Value ecosystem	732630	CSA	01/01/17	31/12/20
Privacy-preserving Big Data Technologies (ICT-18-2016)				
SODA: Scalable Oblivious Data Analytics	731583	RIA	01/01/17	31/12/19
MH-MD: My Health - My Data	732907	RIA	01/11/16	31/10/19
e-Sides: Ethical and Societal Implications of Data Sciences	731873	CSA	01/01/17	31/12/19
SPECIAL - Scalable Policy-awarE linked data arChitecture for prIvacy, trAnsparency and compLiance	731601	RIA	01/01/17	31/12/19
ICT-35-2016				
K-PLEX: Knowledge Complexity	732340	RIA	01/01/17	30/03/18

2.4 BDV Stakeholder Platform

Section 2.4 encompasses an update concerning the collaboration between ETP4HPC and BDVA.

The BDV Stakeholder Platform is one of the main activities of the PPP implementation. Its main objective is to provide a platform to gather BDV stakeholders that cannot commit to regular participation but wish to participate and contribute as they can. Once set-up it will have the capacity to gather and coordinate BDV stakeholder recommendations along the technology, application, skills, ecosystem and social dimensions. It is envisioned that the BDVA Stakeholder Platform will be implemented as part of a cooperation and coordination project.

The BDV Stakeholder Platform will take advantage of already established stakeholder groups and communities, such as those started in BIG [18] and BYTE, and will take them into account wherever appropriate. The stakeholder platform will be open, neutral, independent and representative of the different communities needed to set-up a successful data-driven ecosystem in Europe including technology providers, industrial players both large and SMEs, academia, public sector, users and/or user communities, start-ups etc.

The BDV Stakeholder Platform will address a number of cross-domain and cross-sector topics. As an example, collaboration will be sought with other ETPs such as ETP4HPC, NEM and partnerships like TDL, 5G or EUROGI (see Figure 4). It should also be open to more dynamic agents that can provoke new innovative usages of the data and business models typified by e.g. web entrepreneurs.

While the BDVA is responsible for the overall SRIA, the BDV Stakeholder Platform will provide key inputs for the development of the BDV SRIA. This is due to the activities of the platform being at a more detailed, and

[18] M. Cavanillas, E. Curry, W. Wahlster: New Horizons for a Data-Driven Economy – A Roadmap for Big Data in Europe, Springer International Publishing, 2016.

typically domain-specific, level, thereby complementing insights from BDVA task forces and activities with more detailed perspectives.

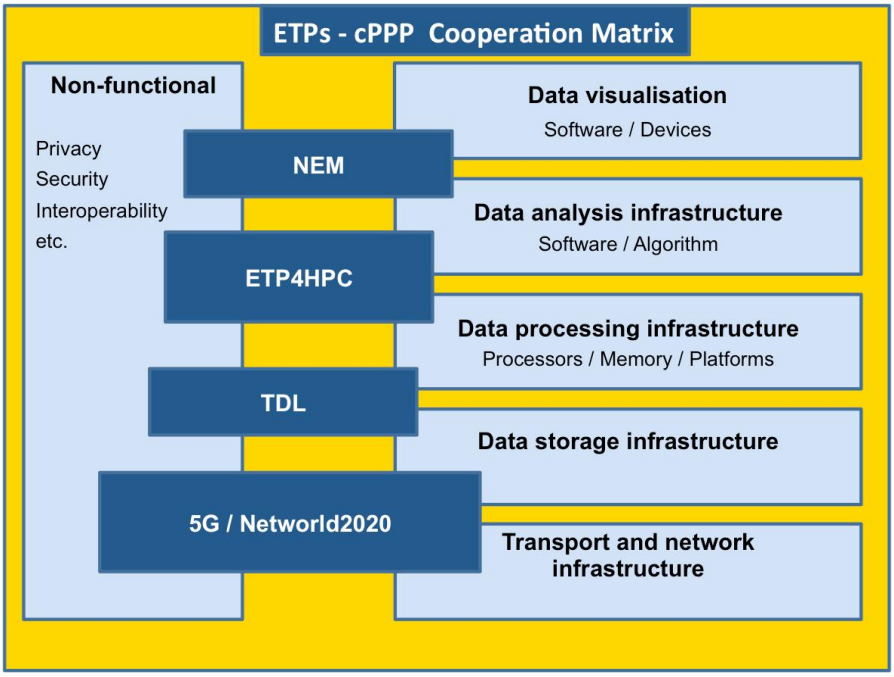

Figure 4: Examples of collaboration with other ETPs and PPPs [19] [20] [21] [22]

BDVA and ETP4HPC Collaboration

To realise the vision of Extreme-Performance Data Analytics (EPDA), there is growing recognition that both the HPC and Big Data ecosystems need to increase collaboration with each other. To this end, the Big Data Value Association (BDVA) and the European Technology Platform for High Performance Computing (ETP4HPC) have approved a joint collaboration charter towards maximising Europe's position. Since then, a number of working sessions have been held (e.g. Barcelona, Valencia, plus numerous virtual meetings) and further scheduled involving technical representatives from both associations, with the commitment to agreeing research themes of common interest in the area of Extreme-Performance Data Analytics (EPDA), including:

- *Existing BDV-PPP Projects:* Inviting an ETP4HPC representative to attend Technical Board meetings which will encompass all BDV-PPP funded projects, including Lighthouses and Innovation Spaces. The presence of an ETP4HPC representative can assist with identifying where HPC capabilities may assist big data analytics within these projects.
- *Cross SR(I)A Influence:* Identify and align research themes common to HPC and Big Data research interests in the area of Extreme Performance Data Analytics (EPDA). This would include aligned submissions to 2018-2020 Big Data Work Programme Consultation.
- *Extreme-Scale Demonstrator (EsDs)*[23]*:* Make a joint recommendation for an 'extreme Big Data' related theme to upcoming Extreme-Scale Demonstrator (EsDs) call for input.
- *Centre of Excellence (CoE) in Computing Applications*[24]*:* Make a joint recommendation for a themed Big Data/HPC investigation in forthcoming call for input.

[19] NEM: New European Media, http://nem-initiative.org/
[20] ETP4HPC: The European Technology Platform for High Performance Computing, http://www.etp4hpc.eu/
[21] TDL: Trust in the Digital World, http://www.trustindigitallife.eu/
[22] 5G: 5G PPP, http://5g-ppp.eu/
[23] http://www.etp4hpc.eu/en/esds.html
[24] https://ec.europa.eu/programmes/horizon2020/en/news/overview-eu-funded-centres-excellence-computing-applications

Additionally, both associations will collaboratively schedule a set of forthcoming interactions involving both ecosystems (e.g. joint events, workshops, and conferences) to advance understanding and definition in these areas.

3 Technical Priorities

The outcome descriptions in Section 3.2-3.6 have been consolidated in order to priorize technical aspects.

A three-way analysis was conducted to identify the key technical priorities that need to be addressed to initiate the development of a European Data Value ecosystem. First, the most important challenges of relevant and representative end-users from various economic sectors were identified by performing a structured needs and requirements analysis as part of a series of sectorial workshops. Second, the outcomes of this needs and requirements analysis was mapped and clustered along the main roles participating in the data value chain. Third, needs and requirements have been crosschecked against existing Big Data technical solutions.

Section 3.1 provides further background, rationale and an in-depth description of the approach that was followed to determine these priorities. The technical priorities resulting from this analysis are presented in Sections 3.2 to 3.6.

3.1 Analysis and Identification of Technical Priorities

Current Situation and European Assets

The fields of Big Data infrastructure and storage techniques are currently dominated by large US IT and Internet companies. Most of the supporting tools and storage architectures are now Open Source (Hadoop, Hive, Spark, Shark, HBase, Riak, Titan Flink, etc.), levelling the playing field for tool vendors in this field. It, therefore, does not seem the most efficient approach to try and overtake or compete with them in these fields by simply repeating what they have already achieved, but rather build on top of the commoditized core that they have established.

The fields of Big Analytics and Data Visualization (such as predictive and decision support systems) in contrast is much more open. The EU has an undeniable competitive advantage here, thanks to the high mathematical and computer literacy level of EU engineers and research scientists, as well as the solid base of industries which own most of the underlying data assets, unlike end-consumer data sets which are dominated by consumer-facing web companies in the US. An example is the domain of IoT applications, where European companies have already established a leading role in different sectors like Transport (e.g. Alstom, CAF, Siemens), Telecommunications (e.g. Ericsson, Nokia, Deutsche Telekom, Telefonica), Smart Cities, Health (e.g. Siemens, Philips), and Aerospace (e.g. Thales, Airbus, Rolls Royce).

This positioning is a major factor of differentiation and a real asset as the added value of Big Data in terms of innovations lies in applications driven by data analysis.

Needs and Stakeholder Analysis

In order to systematically elicit the needs of future Big Data solutions, sectorial workshops have been performed in various fields: geospatial/environment, energy, media, mobility, manufacturing, retail, health, public sector. From the analysis of the results, it is clear that addressing the technical needs of these vertical application markets will require a set of cross-sectors technologies. The main technical needs most often mentioned in the sectorial workshops were:

- Data Integration: Harmonization across different sources (standardized modelling, simplified data access, integration of heterogeneous data sources).
- Data Curation: handling veracity, life-cycle management.
- Handling of data-in-motion: Low latency and real-time data processing.
- Advanced analytics: predictive analytics, graph mining, semantic analysis.
- Data protection and privacy technologies: to make data owners comfortable about sharing data in an experimental environment.
- Advanced visualization, user experience and usability.

To turn Big Data technologies into value, both supply and demand need to be brought together for a mutual benefit. While this will foster the creation of a more competitive Big Data "supplier" industry, it will also take care of developing a European market where benefits will be well documented across a wide range of industrial sectors. The impact of such transverse technologies goes well beyond the vertical sectors described as they require an "ecosystem" that will bring together stakeholders from the European Big Data community (from both demand and suppliers sides) including legal, societal and technical areas.

Three major stakeholder roles relevant from a technical point of view can be identified in the data ecosystem (see Figure 5). For each of these roles, the following main technical needs are identified in Table 1.

Role	What do they do?	How do they make business?	Main technical needs
Data provider	Collect, pre-process, transform data into information and sell or distribute the information	Benefit from wide availability of their data Make a margin on the resale of information	Data management from heterogeneous sources, handling data in rest and data-in-motion
Data processor and Service provider	Buy information, perform deeper analysis to create value and provide services.	Leverages scale effects across multiple clients, service fees	Low latency and data analytics with a good benefit/cost ratio, tools; flexibility to serve multiple clients, wide variety of data sources, predictive analytics
Service consumer	Buy/use data-driven services	Applies decisions and insights derived from analysis to optimization of own business	Privacy and anonymisation, advanced interaction and visualization

Table 1: Roles and activities of ecosystem actors

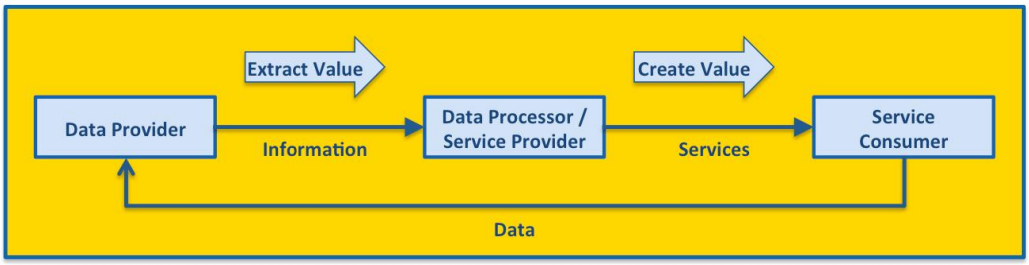

Figure 5: Various technical needs and concerns according to the role in the ecosystems

There are challenges to cope with the volume, velocity, variety, and veracity aspects of data analytics and to integrate novel statistical and mathematical algorithms, as well as prediction techniques into services and applications. Based on the needs gathered, new approaches are required for data management solutions, optimized architectures for both data-at-rest and data-in-motion, data analytics, anonymisation and advanced technologies for visualisation. All are considered as strategic priorities. Real value may stem from the capability to deliver shorter and shorter response times, while analysing more and more complex systems and data sources. Organizations able to handle the increasing complexity and dynamicity of data structures and operations will thus gain a clear competitive advantage.

Based on the needs analysis, the overall, strategic technical goal may be stated as:

Deliver new Big Data technology allowing for deep analytics capacities on data-at-rest and data-in-motion while providing sufficient privacy guarantees, optimized user experience support and a sound data engineering framework.

Achieving this goal requires addressing at least the following technical priorities, which are elaborated in the remainder of this section:

- Principles and techniques for data management.
- Optimized and scalable architectures for analytics of both data-at-rest and data-in- motion with low latency delivering real-time analytics.
- Data analytics to improve data understanding, deep learning, and meaningfulness of data.
- Privacy and anonymisation mechanisms.
- Advanced visualization approaches for improved user experience.

3.2 Priority "Data Management"

Background

More and more data is becoming available. This data explosion, often called "**data tsunami"**, is triggered by the increasing amount of sensor data and social data, born in Cyber Physical Systems (CPS) and Internet of Things (IoT) applications. Traditional means for data storage and data management are no longer able to cope with the size and speed of data delivered in heterogeneous formats and at distributed locations.

Large amounts of data are being made available in a variety of formats – ranging from unstructured to semi-structured to structured formats - such as reports, Web 2.0 data, images, sensor data, mobile data, geo-spatial, multimedia data, For instance, important data types include numeric types, arrays and matrices, geospatial data, multimedia data and text. Much of this data is created or converted and further processed as text. Algorithms or machines are not able to process the data sources due to the lack of explicit semantics. In Europe, text-based data resources occur in many different languages, since customers and citizens create content in their local language. This **multilingualism** of data sources makes it often impossible to use existing tools and to align available resource, because they are generally provided only in the English language. Thus, the seamless aligning of data sources for data analysis or business intelligence applications is hindered by the lack of language support and availability of appropriate resources.

In almost all industrial sectors, isolated and fragmented data pools are found. Due to the prevalence of **data silos**, the seamless integration and smart access to the various heterogeneous data sources is difficult to realize. And still today, data producers and consumers, even in the same sector, are relying on different storage, communication and thus different access mechanisms for their data. Due to a lack of commonly agreed standards and frameworks, the migration/federation of data between pools imposes high-levels of additional cost. Without a **semantic interoperability** layer upon all those different systems, the seamless alignment of data sources cannot be realized.

In order to ensure valuable big data analytics outcome, the incoming **data** has to be of a high **quality**, or at least the quality of the data should be known in order to reason on it accordingly. This requires differentiating between noise and valuable data, thereby being able to decide which data sources to include or exclude in order to achieve the desired results.

Over many years, several different application sectors have tried to develop vertical processes for data management including specific data format standards and domain models. However, a consistent **data lifecycle management**, i.e. the ability to clearly define, interoperate, openly share, access, transform, link, syndicate, and manage data, is still missing. In addition, data, information and content needs to be syndicated from data providers to data consumers whilst maintaining provenance, control and source information including IPR considerations (**data provenance**). Moreover, in order to ensure transparent and flexible data usage, the aggregating and managing of respective data sets enhanced by controlled access mechanism through APIs should be enabled (**Data-as-a-Service**).

Challenges

As of today collected data is rapidly increasing, however the methods and tools for data management do not evolve at the same pace. In this perspective, it becomes crucial to have – at a minimum – good metadata, NLP, and semantic techniques to structure the data sets and content, annotate them, document the associated processes, and deliver or syndicate information to recipients. The following research challenges have been identified:

- **Semantic annotation of unstructured and semi-structured data:** Data needs to be semantically annotated in digital formats, without imposing extra-effort on data producers. In particular unstructured

data, such as videos, images or text in natural language (including multilingual text), or specific domain data, such as Earth Observation data has to be pre-processed and enhanced with semantic annotation.

- **Semantic interoperability**: Data silos have to be unlocked by creating interoperability standards and efficient technologies for the storage and exchange of semantic data and tools to allow efficient user-driven or automated annotations and transformations.
- **Data quality:** Methods for improving and assessing data quality have to be created together with curation frameworks and workflows. Data curation methods might include general-purpose data curation pipelines, on-line and off-line data filtering techniques, improved human-data interaction, standardized data curation models and vocabularies, as well as an improved integration between data curation tools.
- **Data management lifecycle and data governance:** With the tremendous increase of data, integrated data-lifecycle management is facing new challenges: handling the sheer size of data as well as enforcing consistent quality as the data scales in volume, velocity and variability, including support for real-time data management and efficiency in data centres. . Furthermore, as part of the data management lifecycle, data protection and management must be aligned. Control, auditability and life-cycle management are key for governance, cross-sector applications and GDPR.
- **Integration of data and business processes:** a conceptual and technically sound integration of results from the two "worlds" of analytics. Integrating data processes, such as Data Mining or Business Intelligence, on the one side with business processes, such process analysis in the area Business Process Management (BPM), on the other side, is needed.
- **Data-as-a-service:** How to bundle and provision both data and the software and data analytics needed to interpret and process it into a single package that can be provided as (an intermediate) offering to the customer.
- **Distributed trust infrastructures for data management**: Mechanisms to enforce consistency in transactions and data management, for example based on distributed ledger / blockchain technologies. Flexible data management structures based on microservices with the possibility of integrating data transformations, data analysis, data anonymization, etc., in a decentralized manner.

Outcome

- Languages, techniques and tools for measuring and assuring **data quality** (such as novel data management processing algorithms and **data quality governance** approaches that support the specifics of Big Data), and for **data provenance,** control and IPR.
- Principles for a clear **Data-as-a-Service (DaaS) model and paradigm** fostering the harmonization of tools and techniques with the ability to easily re-use, interconnect, syndicate, auto/crowd annotate and bring to life data management use cases and services across sectors, borders and citizens by diminishing the costs of developing new solutions. Furthermore, trusted and flexible infrastructures need to be developed for the DaaS paradigm, potentially based on technologies such as distributed ledgers / blockchains and/or microservices.
- Methods and tools for a complete **data management lifecycle** ranging from data curation and cleaning (including pre-processing veracity, velocity integrity, and quality of the data), using scalable big data transformations approaches (including aspects of automatic, interactive, sharable, and repeatable transformations) to long-term storage and data access. New models and tools to check integrity and veracity of data, through both machine-based and human-based (crowd-sourcing) techniques. Furthermore, mechanisms need to be developed for alignment of data protection and management, addressing aspects such as control, auditability and life-cycle management of data.
- Methods and tools for the sound **integration of analytics results** from **data and business** processes. This relies on languages and techniques for **semantic interoperability** such as standardized data models and interoperable architectures for different sectors enriched through semantic terminologies. In particular, standards and multilingual knowledge repositories/sources that allow industries and citizens to seamlessly link their data with others. Mechanisms to deal with semantic data lakes, industrial data spaces, and development of enterprise knowledge graphs are of high relevance in this context.
- Techniques and tools for handling **unstructured and semi-structured data.** This includes natural language processing for different languages, algorithms for automatic detection of normal and abnormal structures (including automatic measuring, tools for pre-processing and analysing sensor, social, geospatial, genomics, proteomics and other domain orientated data), well as, standardized annotation

frameworks for different sectors supporting the technical integration of different annotation technologies and data formats.

3.3 Priority "Data Processing Architectures"

Background

The Internet of Things (IoT) is one of the key drivers of the Big Data phenomenon. Initially this phenomenon started by applying the existing architectures and technologies of Big Data that we categorise as data-at-rest, which is data stored in persistent storage. In the meantime the need for processing immense amounts of sensor data streams has increased. This type of data-in-motion, that is non-persistent data processed timely on the fly, has extreme requirements for low-latency and real-time processing. What has been hardly addressed is the complete processing for the combination of data-in-motion with data-at-rest.

For the IoT domain these capabilities are essential. This is also needed for other domains like social networks or manufacturing, where huge amounts of streaming data is produced in addition to the available Big Data sets of actual and historical data.

These capabilities will affect all layers of future Big Data infrastructures reaching from the specifications of low level data flows with continuous processing of micro-messages, to sophisticated analytics algorithms. The parallel need of real-time and large data volume capabilities is a key challenge for Big Data processing architectures. Architectures to handle streams of data such as the lambda and kappa architectures will be considered a baseline for achieving a more tight integration of data-in-motion with data-at-rest.

Developing the integrated processing of data-at-rest and data-in-motion in an ad-hoc fashion is of course possible, but only the design of generic, decentralized, and scalable architectural solutions will leverage the true potential. Optimised frameworks and toolboxes allowing the best use of both data-in-motion (e.g. data streams from sensors) and data-at-rest will leverage the dissemination of reference solutions, which are ready and easy to deploy in any economic sector. For example, a proper integration of the data-in-motion with the predictive models based on data-at-rest will enable efficient proactive processing (detection ahead of time). Architectures that can handle heterogeneous and unstructured data are also of importance. When such solutions become available to service providers, in a straightforward manner, they will have the opportunity to focus on the development of business models.

The capabilities of existing systems to process such data-in-motion and answer queries in real-time and for thousands of concurrent users are limited. Special purpose approaches based on solutions like Complex Event Processing (CEP), are not sufficient for the challenges posed by IoT in Big Data scenarios. The problem of effective and efficient processing of data streams (data-in-motion) in a Big Data context is far from being solved, especially when considering the integration with data-at-rest and breakthroughs in NoSQL databases and parallel processing (e.g. Hadoop, Apache Spark, Apache Flink, Apache Kafka). Applications are also required to fully exploit all capabilities of modern and heterogeneous hardware, with parallelism and distribution to boost performance.

To achieve the agility demanded by real-time business and next-generation applications a new set of interconnected data management capabilities is required.

Challenges

There have been advances for Big Data analytics to support the dimension of Big Data volume. Separately stream processing has been enhanced to analytics on the fly to cover the velocity part of Big Data. This is especially important, as business needs to know what is happening now. The main challenges to be addressed are:

- **Heterogeneity:** Big Data processing architectures form places to gather and process various pieces of relevant data together. Such data can vary on several aspects including, various syntactic formats, heterogeneous semantic representations, various levels of granularity, etc. Besides, data can be structured, semi-structured, or unstructured multimedia, audio-visual, and textual data. Hardware can be heterogeneous too (CPUs, GPUs, and FPGAs). A challenge to Big Data processing architectures is to handle Big Data variety at several dimensions.

- **Scalability:** Being able to apply storage and complex analytics techniques at scale is crucial in order to extract knowledge out of the data and develop decision support applications. For instance, predictive systems like recommendation engines must be able to provide real-time predictions while enriching historical databases to continuously train more complex and refined statistical models. The analytics must be scalable with low latency adjusting to the increase of both streams and volume of Big Datasets.
- **Processing of data-in-motion and data-at-rest:** Real-time Analytics through Event Processing and Stream Processing spanning inductive reasoning (machine learning), deductive reasoning (inference), high performance computing (data centre optimisation, efficient resource allocation, quality of service provisioning) and statistical analysis have to be adapted to allow continuous querying over streams (i.e., on-line processing). The scenarios for Big data processing requires also more to cope with the systems which inherently contain dynamics in their daily operation and require its proper management in order to increase the operational effectiveness and competitiveness. Most of these processing techniques have only been applied to data-at-rest and in some cases to data-in-motion. A challenge here is to have suitable techniques for data-in-motion, and also integrated processing for both of them at the same time.
- **Decentralisation:** Big Data producers and consumers can be distributed and loosely coupled as in the Internet of Things. Architectures have to consider the effect of distribution on assumptions underlying them such as loose data agreements, missing contextual data, etc. Distribution of Big Data processing nodes poses the need for new Big Data-specific parallelization techniques and (at least partially) automated distribution of tasks over clusters are crucial elements for effective stream processing. Especially important is an efficient distribution of the processing to the Edge, i.e. local data processing, as a part of the ever increasing trend for Fog computing.
- **Performance:** The performance of algorithms has to scale by orders of magnitude while reducing energy consumption with the best effort integration between hardware and software. It should be possible to utilise existing and emerging high performance computing and hardware-oriented developments like main memory technology with different type of caches, like cloud and fog computing, like software-defined storage with built-in functionality for computation near the data (e.g. Storlets). Also, like data availability guarantees to avoid unnecessary data downloading and archiving, and like data reduction to support storing, sharing, and efficient in-place processing of the data.

Outcome

- **Techniques and tools for processing real-time heterogeneous data sources:** the heterogeneity of data sources for both data-at-rest and data-in-motion requires efficient and powerful techniques for transformation and migration. This includes data reduction and mechanisms to attach and link to arbitrary data. Standardization also plays a key role to address heterogeneity.
- **Scalable and dynamical data approaches:** capabilities of processing very large amounts of data in a very short time (in real-time applications and/or to react to dynamic data) and analysing large amounts of data with the aim of updating the analysis results as the information content changes. It is important to access only the relevant and suitable data, hereby avoiding the access and processing of irrelevant data. Research should provide new techniques that can speed up training on large amounts of data, for example by exploiting parallelisation, distribution, flexible could computing platforms, and moving computation to edge computing.
- **Real-time architectures for data-in-motion:** architectures, frameworks and tools for real-time and on-the-fly processing of data-in-motion, e.g. IoT sensor data, and integrating it with data-at-rest. Furthermore, there is a need to dynamically reconfigure such architectures and dynamic data processing capabilities on the fly, for example, to cope with context changes, changing requirements and optimisation in various dimensions (e.g., performance, energy consumption and security).
- **Decentralised architectures:** architectures that can deal with Big Data produced and consumed by highly decentralised and loosely coupled parties such as in the Internet of Things, with secure traceability such as Blockchain. Additionally, architectures with parallelisation and distributed placement of processing for data-in-motion and its integration with data-at-rest.
- **Efficient mechanisms for storage and processing:** real-time algorithms and techniques are needed for the requirements demanding low latency when handling data-in-motion. Developing hardware and software together for cloud data platforms will in turn enable applications to run with outstanding reliability and energy efficiency.

3.4 Priority "Data Analytics"

Background

The progress of Data analytics of Big Data is not only key to turn Big Data into value, but also to make it accessible to the wider public. Data analytics will have a positive influence on all parts of the data value-chain to increase business opportunities through business intelligence and analytics while bringing benefits to both society and citizen.

Data Analytics is an open emerging field in which Europe has strong competitive advantages with promising business development potential. It was estimated that governments in Europe could save $149 billion[25] by using Big Data analytics to improve operational efficiency. Big analytics can provide additional value in every sector where it is applied, leading to more efficient and accurate processes. A recent study by the McKinsey Global Institute placed a strong emphasis on analytics, ranking it as the future main driver for the US economic growth, before shale oil and gas production[26].

The next generation of analytics will need to deal with the vast amount of information from different types of sources with differentiated characteristics and levels of trust, and frequency of update. Data analytics will need to provide insights into the data in a cost-effective and economically sustainable way. On one hand there is a need to create complex and fine-grained predictive models on heterogeneous and massive datasets such as time series or graph data. On the other hand such models must be applied in real-time on large amounts of streaming data. This ranges from structured to unstructured, numerical to micro-blogs, and streams of data. The latter is extremely challenging because the data, besides its volume, is very heterogeneous and highly dynamic which also calls for scalability and high throughput. For instance, data collection related to a disaster area can easily occupy terabytes in binary GIS formats, and real-time data streams can show bursts of gigabytes per minutes.

Challenges

Understanding data, whether it is numbers, text, or multimedia content, has always been one of the greatest challenges for data analytics. Entering into the era of Big Data this challenge has scaled to a degree that makes the development of new methods necessary. In the following we detail the research areas identified for Data Analytics:

- **Semantic and knowledge-based analysis:** Improvement to the analysis of data to provide a near-real-time interpretation of the data (i.e. sentiment, semantics, etc.). Furthermore, ontology engineering for Big Data sources, interactive visualization & exploration, real-time interlinking and annotation of data sources, scalable and incremental reasoning, linked data mining, cognitive computing
- **Content validation:** Implementation of veracity (source reliability / information credibility) models for validating content and exploiting content recommendations from unknown users.
- **Analytics frameworks & processing:** New frameworks and open APIs for the quality-aware distribution of batch and stream processing analytics with minimal development effort from application developers and domain experts. Improvement of the scalability and processing speed for the aforementioned algorithms in order to tackle linearization and computational optimization issues.
- **Advanced business analytics and intelligence:** All the above items enable the realisation of real and static business analytics and business intelligence empowering business and other organisations to make accurate and instant decisions to shape their market. The simplification and automation of these techniques is necessary especially for SMEs.
- **Predictive and prescriptive analytics**: Machine learning, clustering, pattern mining, network analysis and hypothesis testing techniques applied on extremely large graphs containing sparse, uncertain and incomplete data. Building on results of related research activities within the current EU work-programme, sector-specific challenges and contextualization combining heterogeneous data and data streams via graphs to improve the quality of mining processes, classifiers, and event discovery, need to be addressed. These capabilities will open up novel opportunities for predictive analytics in

[25] "*Big Data: The next frontier for innovation, competition and productivity*", McKinsey Global Institute, June 2011
[26] "*Game changers: Five opportunities for US growth and renewal*", McKinsey Global Institute, 2013

terms of predicting future situations, and even prescriptive analytics in terms of providing actionable insights based on forecasts.

Outcome

The main expected advanced analytics innovations are the following:

- **Improved models and simulations**: Improve the accuracy of statistical models by enabling fast non-linear approximations in very large datasets. Move beyond the limited samples used so far in statistical analytics to samples covering the whole or the largest part of an event space/dataset.
- **Semantic analysis:** Deep learning, contextualization based on AI, machine learning, natural language, and semantic analysis in near-real time. Provide canonical paths so that data can be aggregated and shared easily without dependency on technicians or domain experts. Enables the smart analysis of data across and within domains.
- **Event and pattern discovery:** Discover and predict rare real-time events that are hard to identify since they have a small probability of occurrence, but have a great significance (such as physical disasters, a few costly claims in an insurance portfolio, rare diseases and treatments).
- **Multimedia (unstructured) data mining:** Processing of unstructured data (multi-media, text). Linking and cross-analysis algorithms to deliver cross-domain and cross-sector intelligence.
- **Deep learning techniques for business intelligence:** Coupled with the priorities on visualisation and engineering to provide user-friendly tools which connect to open and other data sets and streams (including a citizen's data), provided intelligent data interconnection for business and citizen orientated analytics, and allow visualization (e.g. diagnostic, descriptive and prescriptive analytics).

3.5 Priority "Data Protection"

Background

Data protection and anonymization is a major issue in the areas of Big Data and data analytics. With more than 90% of today's data being produced in the last two years, a huge amount of person-specific and sensitive information coming from disparate data sources such as social networking sites, mobile phone applications, electronic medical record systems, etc., is being increasingly collected. Analysing this wealth and volume of data offers remarkable opportunities for data owners but, at the same time, requires the use of state-of-the-art data privacy solutions to guarantee the privacy of the individuals who are represented in the data by also adhering to legal privacy regulations. Data protection, while important in the development of any modern information system, becomes crucial in the context of large-scale sensitive data processing.

Recent studies on mechanisms for protecting privacy have demonstrated that simple approaches, such as the removal or masking of the direct identifiers in a dataset (e.g., names, social security numbers, etc.), are insufficient to guarantee privacy. Indeed, such simple protection strategies can be easily circumvented by attackers who possess little background knowledge about specific data subjects. Due to the critical importance of addressing privacy issues in lots of business domains, privacy-protection techniques that offer formal privacy guarantees have become a necessity. This has paved the way to the development of privacy models and techniques, such as differential privacy, private information retrieval, syntactic anonymity, homomorphic encryption, secure search encryption, and secure multiparty computation, among others. The maturity level of these technologies varies, with certain technologies such as k-anonymity being more mature than others. However, none of these technologies has so far been applied to large-scale commercial data processing tasks involving big data.

In addition to the privacy guarantees that can be offered by state-of-the-art privacy-enhancing technologies, another important consideration concerns the ability of the data protection approaches to maintain the utility of the datasets to which they are applied, towards supporting different types of data analysis. Privacy solutions that offer guarantees while maintaining high data utility will make privacy technology a key enabler for the application of analytics over proprietary and potentially sensitive data.

A truly modern and harmonized legal framework on data protection that has teeth and can be properly enforced will ensure stakeholder pay attention to the importance of data protection. At the same time, it should enable the uptake of big data and clearly incentivise privacy-enhancing technologies which can be an asset for Europe as this is currently an underdeveloped market. In addition, users are starting to pay more

attention as to how their data is processed. Hence, firms operating in the digital economy may realize that investing in privacy enhancing technologies could provide a competitive advantage.

Challenges

In this perspective the following main challenges have been identified:

- A more **generic, easy to use, and enforceable data protection approach** suitable for commercial large-scale processing is needed. Data usage should conform to the current legislation and policies. On the technical side, mechanisms are needed in order to provide the data owners with the means to define the purpose, and control the granularity at which their data will be shared with authorized third parties throughout its whole lifecycle (data-in-motion and data-at-rest). Citizens, for example, should be able to decide on the destruction of their personal data (right to be forgotten). Data protection mechanisms also need to be "easy" or at least with a reasonable level of effort in order to be used and understood by the various stakeholders, especially end-users. Technical measures are also needed to enforce and enable auditability of the fact that data is only used for the defined purpose and nothing else, i.e. for usage control in particular of personal information. In distributed settings such as supply chains, distributed trust technologies such as blockchains can be part of the solution.
- **Robust data privacy** with utility guarantees is an important challenge, which also implies sub-challenges, such as the need for state-of-the-art data analytics to cope with encrypted or anonymised data. Scalability of the solutions is also a critical feature. Anonymisation schemes may expose weaknesses exploitable by opportunistic or malicious opponents and thus new and more robust techniques must be developed to tackle these adversary models, therefore, ensuring the irreversibility of the anonymisation of Big Data assets is a key Big Data issue. On the other hand, encrypted data processing techniques such as multiparty computation or homomorphic encryption provide stronger privacy guarantees but can currently only be applied on small parts of a computation due to their great performance penalty. In addition, data privacy methods that can handle different data types as well as co-existing data types (e.g., datasets containing relational data together with sequential data about users), and methods that are designed to support analytic applications in different sectors (e.g., telecommunications, energy, healthcare, etc.) are important. Finally, preserving anonymity often implies removing the links between data assets. However, the approach to preserve anonymity also has to be reconciled with the needs for data quality, on which link removal has a very negative impact. This choice can be on the end user side, who have to balance the service benefits and possible loss of privacy, or on the service provider side who have to offer a variety of added-value services according to the privacy-acceptance of their customers. Measures to quantify privacy loss and data utility can be used to allow end-users to make informed decisions.
- **Risk based approaches** calibrating controllers' obligations regarding privacy and personal data protection must be considered especially when dealing with the combined processing of multiple data sets. It has indeed been shown that when processing combinations of anonymised, pseudonymised even public data sets there is a risk that personal identifiable information can be retrieved. Thus providing tools to assess or prevent the risk associated with such a processing is an issue of significant importance.

Outcomes

- **Complete data protection framework**: Mechanism for data protection within innovation spaces. This includes protecting the cloud infrastructure, analytics applications, and the data from leakage and threats, but also provides easy to use privacy mechanisms. Apart from specification of the intended use of data, also usage control mechanisms should be covered.
- **Mining algorithms:** Developed privacy-preserving data mining algorithms.
- **Robust anonymisation algorithms:** Scalable algorithms that guarantee anonymity even when other, external or publicly available data is integrated. In addition, algorithms that allow the generation of reliable insights by crossing data from a particular user in multiple databases, while protecting the identity of the user. Moreover, anonymisation methods that can guarantee a level of data utility to support intended types of analyses. Last, algorithms that can anonymise datasets of co-existing data types, which are commonly met in many business sectors, such as in energy, healthcare and telecommunications.

- **Protection against reversibility:** Methods to analyse datasets to discover privacy vulnerabilities, evaluate the privacy risk of sharing the data, and decide on the level of data protection that is necessary to guarantee privacy. Risk assessment tools to evaluate the reversibility of the anonymisation mechanisms.
- **Multiparty mining / pattern hiding:** Secure multiparty mining mechanisms over distributed datasets, so data on which mining is to be performed is partitioned, horizontally or vertically, and distributed among several parties. The partitioned data cannot be shared and must remain private, but the results of mining on the "union" of the data are shared among the participants. Design of mechanisms for pattern hiding so data is transformed in a way that certain patterns cannot be derived (via mining) while others can.

3.6 Priority "Data Visualisation and User Interaction"

Background

Data Visualisation plays a key role in exploring and understanding effectively Big Data. Visual analytics is the science of analytical reasoning assisted by interactive user interfaces. Data generated from data analytics processes need to be presented to end users via (traditional or innovative) multi-device reports and dashboards which contain varying forms of media for the end-user, ranging from text, charts, to dynamic, 3D, and possibly augmented reality visualisations. In order for users to quickly and correctly interpret data in multi-device reports and dashboards, carefully designed presentation and digital visualisation is required. Interaction techniques fuse together user input with output to provide a better way for a user to perform a task. Common tasks that allow users to gain a better understanding of Big Data include scalable zooms, dynamic filtering, and annotation.

When representing complex information on multi-devices screens, the design issues multiply rapidly. Complex information interfaces need to be responsive to human needs and capacity[27]. Knowledge workers need relevant information in a just-in-time manner. Too much information, which cannot be efficiently searched and explored, can hide the information that is most relevant. In fast moving time constrained environments they need to be able to quickly understand the relevance and relatedness of information.

Challenges

In the data visualisation and user interaction domain, the tools that are currently used to communicate information need to be improved due to the significant changes brought about with the volume and variety of Big Data. Advanced visualisation techniques must consider this variety (i.e. graphs, geospatial, sensor, mobile, etc.) of data available from diverse domains. Tools need to support user interaction for the exploration of unknown and unpredictable data within the visualisation layer. The following list briefly details the research areas identified for visualisation and user interaction:

- **Visual data discovery**: Access to information is at present based on a user-driven paradigm: the user knows what they need, and the only issue is to define the right criteria. With the advent of Big Data, this user-driven paradigm no longer proves to be the most efficient. Data-driven paradigms will emerge where information is proactively extracted through data discovery techniques and systems are anticipating the user's information needs.
- **Interactive visual analytics of multiple scale data:** There are significant challenges in visual analytics in the area of multiple-scale data. Appropriate scales of analysis are not always clear in advance, and single optimal solutions are unlikely to exist. Interactive visual interfaces have great potential for facilitating the empirical search for acceptable scales of analysis and the verification of results by modifying the scale and the means of any aggregation.
- **Collaborative, intuitive, and interactive visual interfaces**: What is needed is an evolution of visual interfaces towards becoming more intuitive and exploiting the advanced discovery aspects of Big Data analytics. This is required in order to foster effective exploitation of the information and knowledge that

[27] "*The Humane Interface: New Directions for Designing Interactive Systems*", Raskin, J. Addison-Wesley, Reading, MA, 2000

Big Data can deliver. In addition, there are significant challenges for effective communication and visualisation of Big Data insights in organisations to enable collaborative decision-making processes.

- **Interactive visual data exploration and querying in a multi-device context**: A key challenge is the provisioning of cross-platform mechanisms for data exploration, discovery, and querying. How to deal with uniform data visualization on multi-devices and how to ensure access to functionalities for data exploration, discovery, and querying in multi-device settings are difficult problems that require new approaches and paradigms to be explored and developed.

Outcome

The main expected advanced visualisation and user experience are the following:

- **Scalable Data Visualization Approaches and Tools:** In order to handle extremely large volumes of data, interaction must focus on aggregated data at different scales of abstraction rather than on individual objects. Techniques for data summarization in different contexts are of high relevance. There is a need to develop novel interaction techniques that can enable easy transitions from one scale or form of aggregation to another (e.g. from neighbourhood-level to city-level) while supporting aggregation and comparisons among different scales. It is necessary to address the uncertainty of the data and its propagation through aggregation and analysis operations.
- **Collaborative, 3D and Cross-Platform Data Visualization Frameworks:** Novel ways to visualize large amounts of possibly real-time data on different kinds of devices, including augmented reality visualization of data on mobile devices (e.g. smart glasses), as well as realtime and collaborative 3-D visualization techniques and tools.
- **New Paradigms for Visual Data Exploration, Discovery, and Querying:** End-users need simplified mechanisms for visual exploration of data, intuitive support for visual query formulation at different levels of abstractions, and tool-supported mechanisms for visual discovery of data.
- **Personalized End-User Centric Data Reusable Visualization Components:** Plug and Play visualization components that support the combination of any visualization asset in a real time plug and play manner, and can be adapted and personalized to the needs of end users, including also advanced search capabilities rather than predefined visualization and analytics. User feedback should be as simple as possible.
- **Domain-specific Data Visualization Approaches:** Techniques and approaches supporting specific domains in exploring domain-specific data. For example, innovative ways to visualize data in the geospatial domain, such as geo-locations, distances, and space/time correlations (i.e. sensor data, event data). Other example are time-based Data Visualization (necessity to take into account the specifics of time2; in contrast to common data dimensions which are usually "flat", time has an inherent semantic structure and a hierarchical system of granularities which must be addressed). Another example is the visualization of Interrelated/Linked data, exploiting graph visualization techniques to allow easy exploration of network structures.

3.7 Roadmap and Timeframe

This innovation roadmap was updated in accordance to the SRIA survey results in 2016.

In order to achieve the overall, strategic technical goal laid out in Section 1.6 and to address the aforementioned technical priorities, an innovation roadmap defining the expected outcomes has been established. The roadmap is based on the existing Work Programme Calls and the prioritisation of the five technical priorities derived from the internal and external community engagement surveys, as illustrated below, conducted in the context of the SRIA update process (see Appendix 6.3).

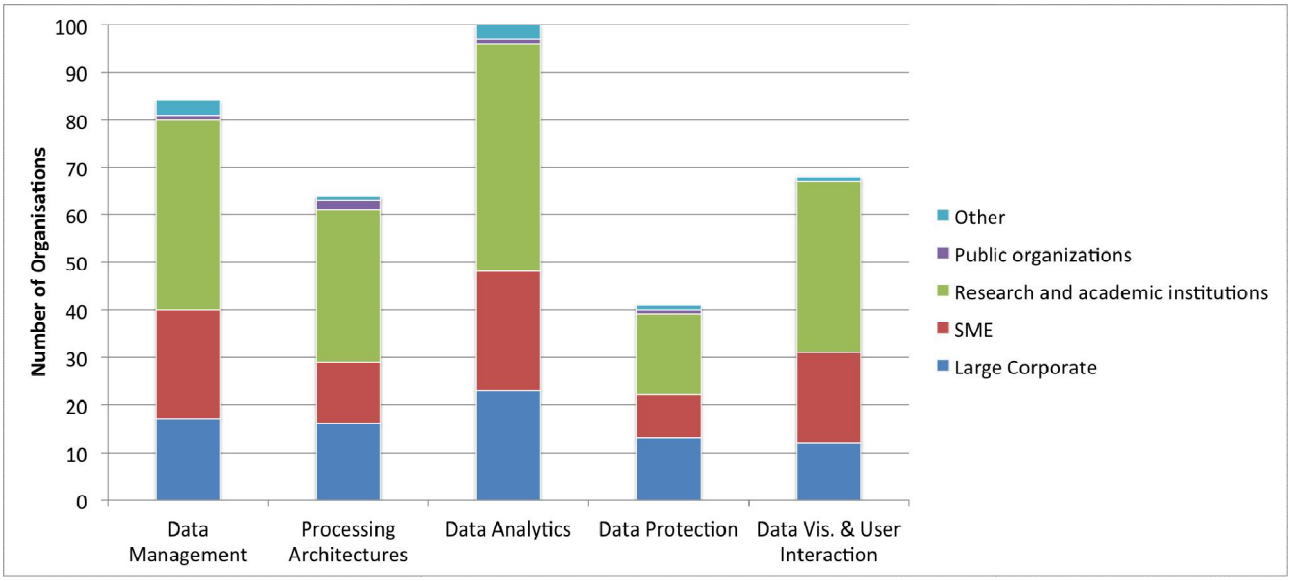

Figure 6 Top Three Technical Priorities for Organisation

The results of this survey are reflected in the definition of the roadmap for technical priorities depicted.

Technical Priority	Year 1	Year 2	Year 3	Year 4	Year 5
Data Management					
Data Processing Architecture					
Data Analytics					
Data Protection					
Data Visualization					

Figure 7 Roadmap for Technical Priorities

4 Non-Technical Priorities

The portfolio of activities of the Big Data Value SRIA needs to comprise support actions that address complementary, non-technical issues alongside the European Innovation Spaces, Lighthouse projects, and the research and innovation activities. In addition to the activities addressing the governance of the PPP[28], the non-technical activities will focus on:

- Skills development.
- Business Models and Ecosystems.
- Policy, Regulation and Standardization.
- Social perceptions and societal implications.

4.1 Skills development

Related activities of the CSA BDVe were incorporated into this section.

In order to leverage the potential of Big Data Value, a key challenge for Europe is to ensure the availability of highly and rightly skilled people who have an excellent grasp of the best practices and technologies for delivering Big Data Value within applications and solutions. In addition to meeting the technical, innovation, and business challenges as laid out in this document, Europe needs to systematically address the need for educating people that are equipped with the right skills and are able to leverage Big Data Value and so enabling best practices. Education and training will play a pivotal role in creating and capitalizing on EU-based Big Data Value technologies and solutions.

Over the past few years several European initiatives have started to fill the gap of profiles and to exactly define the profiles that will be required. In particular The European Data Science Academy (EDSA) is **analysing the required sector specific skill-sets for data analysts** across the main industrial sectors in Europe to develop a **data science curricula** to meet these needs. The EIT-Digital Master on Innovation has launched a major on Data Science as a joint initiative of six European Universities (Universidad Politecnica de Madrid (UPM), Eindhoven University of Technology (TU/e), and Universite Nice Sophia Antipolis (UNS), KTH, TUB).

Building on these activities as part of the PPP the Big Data Value ecosystem (BDVe) CSA will target activities to improve skills, education, and Centers of Excellence around Big Data. It will facilitate coordination between Member States, help to align curricula with industry needs, and accelerates skills development to increase the number of European data scientists by 2020. BDVe will address these challenges by building on top of the work already done and in cooperation with existing initiatives. Specific activities will include:

- A Network of National BDV Centers of Excellence to foster collaboration and share best practices between existing centers and support the establishment of new centers
- To exchange knowledge on data scientist educational programmes across all Member States by delivering a Big Data Value Education Hub as a platform and living repository for knowledge
- A Certification of Curricula and Training Programmes for BDV professionals to ensure their alignment with industry needs
- To stimulate and promote mobility of students, confirmed data professionals and domain experts and mobility opportunities beyond the BDV PPP such as Industrial Internships

Understanding the skills challenge requires a clear definition of the appropriate profiles required to cover the full data value chain. The first SRIA of the BDVA identified data scientists and data engineers. Building on this basis must recognise the individual needs linked to company size. Start-Ups, SME and big industries have very individual requirements in data science. These requirements need to be collected and addressed. However, we were on one hand forgetting about those skills regarding data storage and management and on the other hand, we were using the term engineer to refer to those experienced in business expertise. Consequently, we distinguish now three different profiles, i) to cover the hardware and software infrastructure related part, ii) the analytical part and iii) the business expertise.

[28] Which are described in detail in the Big Data Value PPP proposal.

The educational support for data strategists and data engineers is however far too limited to meet industry requirements, mainly due the spectrum of skills and technologies involved. By transforming the current knowledge-driven approach into an experience-driven one, we can fulfil the industry's needs for individuals capable of shaping the data driven enterprise. The current curricula are furthermore too siloed, leading to communication problems and suboptimal solutions and implementations. The next generation of data professionals needs this wider view in order to deliver the data driven organisation of the future.

Data-intensive Engineers. Successful Data-intensive Engineers control how to deal with data storage and management. They are experts on distributed computing and computing centres, and hence they are mostly at the advanced system administrator levels. They have the know-how to operate large clusters of (virtual) machines, how to configure and optimise load-balancing, how to organise Hadoop clusters, they know about HDFS and RDDs, etc.

Data Scientists. Successful Data Scientists will require solid knowledge in statistical foundations and advanced data analysis methods combined with a thorough understanding of scalable data management, with the associated technical and implementation aspects. They will be the specialists that can deliver novel algorithms and approaches for the Big Data Value stack in general, such as advanced learning algorithms, predictive analytics mechanisms, etc. They are data-intensive analysts. They need to know statistics and data analysis, they need to be able to talk to the data-intensive Engineers, but should be relieved from system administrator problems, and they need to understand how to transform problems into the appropriate algorithms, which may need to be modified slightly. The data scientist benchmarks, selects and optimises these algorithms to reach a business objective. They also need to be able to evaluate the results obtained, following sound scientific procedures. A data scientist curriculum would ideally provide enough insight into the data engineering discipline to steer the selection of algorithms, not only from a business perspective, but also from an operational and technical perspective. For this, Europe needs new educational programmes in data science as well as ideally a network between scientists (academia) and industry that will foster the exchange of ideas and challenges. PPP actions such as BDVe will provide support to facilitate activities towards these goals.

Data-intensive Business Experts. They are the specialists that develop and exploit techniques, processes, tools and methods for developing applications that turn data into value. In addition to technical expertise, data-intensive business experts need to understand the domain and the business of the organizations. This means they need to bring in domain knowledge and are thus working at the intersection of technology, application domains and business. In a sense, they thereby constitute the link between technology experts and business analysts. Data-intensive Business Experts will foster the development of Big Data applications from an "art" into a disciplined engineering approach. They will thereby allow the structured and planned development and delivery of customer-specific Big Data solutions, starting from a clear understanding of the domain, as well as customer and user needs and requirements.

Extensive experience and skills acquired by working on projects in the specified technical priority areas of the SRIA, together with the domain specific knowledge obtained from the development of lighthouse projects will guide the identifying of skill development requirements that can be addressed by collaborating with higher education institutes and education providers to support the establishment of:

- New educational programmes based on interdisciplinary curricula with a clear focus on high-impact application domains.
- Professional courses to educate and re-skill/up-skill the current workforce with the specialised skill-sets needed to be Data-intensive Engineers, Data Scientists and Data-intensive Business Experts. These course will stimulate life long learning in the domain of data and to adopt new data related skills.
- Foundational modules in data science, statistical techniques, and data management within related disciplines such as legal and humanities.
- A network between scientists (academia) and industry that leverages Innovation Spaces to foster the exchange of ideas and challenges.
- Datasets and infrastructure resources, provided by industry, that enhances the industrial relevance of courses.

The regularly updated strategic challenge areas will provide orientation for the development of the required data skills to support building extensive know-how (e.g. by European curricula and sharing of best practices) and skills in Europe for future systems in the industrial and research community.

4.2 Ecosystems and Business Models

Minor updates fostering alignment with ongoing activities in the related taskforce.

The Big Data Value ecosystem (see Figure 8) will comprise many new stakeholders. New concepts for data collection, processing, storing, analysing, handling, visualisation and most importantly usage will emerge and business models will be created around it.

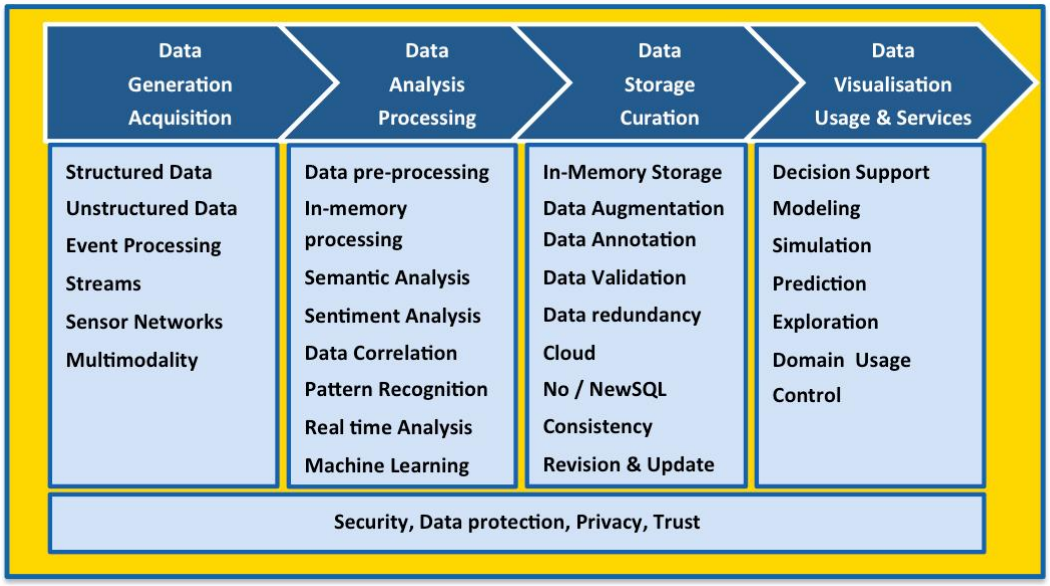

Figure 8 Big Data Ecosystem along the Value Chain

There are three key ways to generate value for companies and along the value chain regardless of sector or domain: optimising and improving core business, white labelling & infrastructure and perhaps most importantly, entirely new business models and business development (see Figure 9).

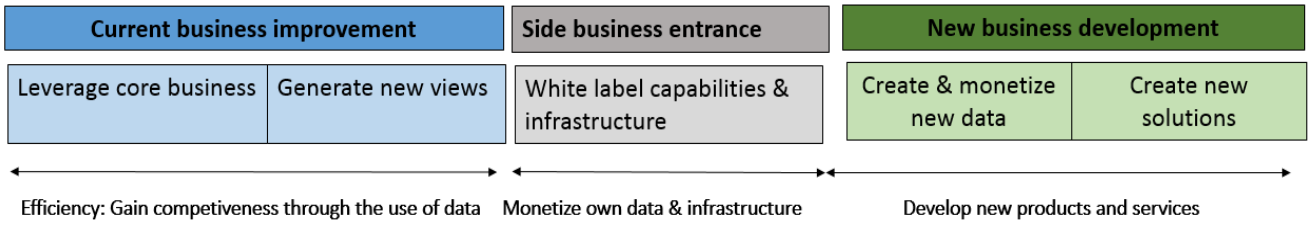

Figure 9 Big Data Value Proposition[29]

Identifying sustainable business models and ecosystems in and across sectors and platforms will be an important challenge. In particular many SMEs that are now involved in highly specific or niche roles will need support to help them align and adapt to new value chain opportunities.

Dedicated projects and activities investigating and evaluating existing and emerging business propositions and models will in part be linked to the innovation spaces where suppliers and users will meet. Those projects will:

- Establish a mapping of technology or platform providers and their value contribution.
- Identify mechanisms by which the value of data can be adequately determined
- Provide a platform for entrepreneurs and financial actors for them to gain adequate levels of understanding about the value chain of big data.
- Scope, describe and validate business propositions and models that might be successful and sustainable in the future data economy.

[29] The table shows different value propositions linked to the different industrial interests (e.g. large companies on business transformation). However it should not be considered as a roadmap

The outcomes of these projects will contribute to the creation of a more stable business environment that enables business, particularly web entrepreneurs and SMEs, to access Big Data markets and ecosystems. Europe needs to raise more and stronger players to make the whole Big Data Value ecosystem strong, vibrant and valuable such that it will lift the entire Europe's economy, The following **key stakeholders** are seen as the main actors along the Big Data Value chain:

- **User Enterprises**, e.g. enterprises in all sectors and off all size that want to improve their services and products using Big Data technology, data products and services.
- **Data Generators and Providers** that create, collect, aggregate, transform and model raw data from various public and non-public sources and offer it to customers.
- **Technology Providers** that provide tools & platforms that offer data management and analytics tools to extract knowledge from data, curate and visualize it.
- **Service Providers** that develop Big Data applications on top of the tools and platforms to provide services to user enterprises.

In addition, the following organisations and communities will have an impact on data-driven ecosystems building on top of the Big Data Value chain:

- **Regulatory bodies** to define privacy and legacy issues related to data usage
- **International/national de jure and de facto standardisation bodies** in order to promote new concepts, systems and solutions for global adoption in international standards.
- **Collaborative networks** where different players in the value chain collaborate to offer value services to their customers based on data value creation.

The current stakeholders in H2020 that operate along the phases of research, innovation, exploitation, and usage will also play an important role in leveraging the Big Data Value chain.

To engage the full value chain of stakeholders, the strategic intention of BDVA is to move forward with analysing and defining new business propositions and models. It will do so by investigating the emergence and evolution of the big data ecosystem in two key ways: the first one addressing the SME, start-ups and entrepreneurship aspect, and the second investigating how the value of big data can be leveraged effectively in (transforming) traditional business and industries. Taken together these tracks will make for a comprehensive analysis of new business propositions.

4.3 Policy, Regulation and Standardisation

Input to policy making and legal support

The PPP has no mandate or competence to be involved directly in policy making for legal or regulatory framework conditions. However, the PPP needs to contribute to the policy and regulatory debate about non-technical aspects of the future Big Data Value creation as part of the data-driven economy. Dedicated projects have to address the circumstance of data governance and usage, data protection and privacy, security, liability, cybercrime, Intellectual Property Rights (IPR), etc.

These projects will initiate activities that are foreseen for exchange between stakeholders from industry, end users, citizen and society to develop input to on-going policy debates where appropriate. Equally it will identify the concrete legal problems for actors in the Value Chain particularly SMEs who have no legal resources. This will establish a body of knowledge on legal issues with a helpdesk for the project participants and ultimately for the wider community. The mentioned projects will:

- Establish an inventory of roadblocks inhibiting a flourishing data-driven economy, e.g. by materializing the value of Big Data collections.
- Make and collect observations about the discovery of new legal and regulatory challenges along with the implementation of state-of-the-art technology and the introduction of new technology.

By doing so, these projects will contribute from the perspective of developments of novel technology and solutions and will have direct contact with the actors to help legislators and regulators make exhaustive considerations about framework conditions. Furthermore these projects will support the BDV actors particularly SMEs to get around legal barriers to integrate into new ecosystems.

Standardisation

Standardisation is essential to the creation of a Data Economy and the PPP will support establishing and augmenting both formal and de facto standards. The PPP will achieve this by:

- Leveraging existing common standards as the basis for an open and successful Big Data market.
- Integrating national efforts on an international (European) level as early as possible.
- Ensuring availability of experts for all aspects of Big Data in the standardisation process.
- Providing education and educational material to promote developing standards.

Standards play a pivotal role in any market to provide customers with a true choice by being able to choose comparable and compatible goods or services from multiple suppliers. In the Big Data ecosystem, this applies to both the **technology** and to the **data**.

Technology Standardisation: Most technology standards for Big Data processing technology are *de facto* standards that are not prescribed (but at best *de*scribed after the fact) by a standards organisation. However, the **lack of standards is a major barrier**. One example is NoSQL databases. The history of NoSQL is based on solving specific technologies challenges that lead to a range of different storage technologies. The large range of choices, coupled with the lack of standards for querying the data, makes it harder to exchange data stores as it may tie application specific code to a certain storage solution. The NoSQL databases are designed for scalability, often by sacrificing consistency. Compared to relational databases, they often use a low-level, non-standardized query interface that makes it harder to integrate in existing applications that expect an SQL interface. The lack of standard interfaces also makes it harder to switch vendors. While it seems plausible to define standards for a certain type of NoSQL databases, creating one language for different NoSQL database types is a hard task with an unclear outcome. The PPP would take a pragmatic approach to standardisation and would look to influence, in addition to NoSQL databases, the standardisation of technologies such as complex event processing for real-time Big Data applications, languages to encode the extracted knowledge bases, computation infrastructure, data curation infrastructure, query interfaces, and data storage technologies.

Data Standardisation: The data "variety" of Big Data makes it very difficult to standardise. Nevertheless, there is a lot of potential for data standardisation in the areas of data exchange and data interoperability.

Big Data is valuable for any organisation across many sectors. Exchange and use of data assets are essential for functioning ecosystems and the data economy. Enabling the seamless flow of data between participants (i.e. companies, institutions, and individuals) is a necessary cornerstone of the ecosystem.

To this end, the PPP would undertake collaborative efforts to support, where possible and pragmatic, the definition of semantic standardized data representation ranging from domain (industry sector) specific solutions, like domain ontologies to general concepts such as Linked Open Data. If such standards for data descriptions and meta-data could be established, it would simplify and reduce the cost of data exchange. Insufficiently described data formats, which are a barrier for global & efficient data exchange and processing, are then eliminated.

4.4 Social perceptions and societal implication

Big Data will provide solutions for major societal challenges in Europe. For an accelerated adoption of Big Data it is critical to increase awareness of the benefits and the Value that Big Data offers, and to understand the obstacles in building the solutions and taking them into use. Lack of trust from end users in Big Data technology is an important barrier that may hinder adoption, which includes aspects like privacy, transparency (ability to understand and interpret), perceived efficacy (the expected benefits), manageability (ease of use and level of control that the user can exert), and acceptability (related to ethical issues that arise when new technology creates new questions, for instance in the case of profiling users for insurance companies; but also willingness to share data: in many cases end users are expected to contribute to the service by providing data themselves). In addition collaboration and co-innovation between organizations, public sector and private people should be enhanced to support value creation from big data solutions.

Societal challenges are covering a wide range of topics:

- How to establish and increase trust in Big Data innovations, addressing transparency, efficacy, manageability and acceptability.
- How to incorporate privacy-by-design principles and create a common understanding amongst the technical community leading to an operational and validated method that is applicable to data-driven innovations development?
- How to develop a better understanding of inclusion and collective awareness aspects of Big Data innovations? How to enable a clear profile of social benefits Big Data Value technology can provide?
- How to identify ethical issues that are created by Big Data innovations, leading to these issues formulated in a clear way, and directions for solutions identified.

By addressing the listed topics, one will assure that citizen's views, and perception is taken into account so that technology and applications are not developed without a chance to be widely accepted. The above actions will be based on and related to work done that address the bridge between ICT and society, for instance, the BYTE and Big Data Europe projects, and other NGOs, such as the Digital Enlightenment forum and national organisations.

5 Expected Impact

5.1 Expected Impact of strategic objectives

The expected impact of the PPP should be recognised in the great enhancement that Big Data analysis techniques will provide to all decision-making processes. From this point of view every sector, private or public, industrial or academic, will be touched as will society. The PPP will show that Big Data Value is not just a new buzzword, but shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions.

The general impact of the PPP is expected in the following lines:

- **Effective service provision** from public and private organisations will be achieved by developing and making available to industry and the public sector technology, applications and solutions for the creation of value from Big Data for increased productivity, optimised production, more efficient logistics (inbound and outbound).
- **Extensive experience and skills will be acquired** and an IPR base will be set-up to support building extensive know-how (e.g. by European curricula and sharing of best practices) and skills in Europe for future systems in the industrial and research community.
- **New Business Models and Optimisation** of existing industries will drive the integration of the BDV services into private and public decision-making systems such as Enterprise Resource Planning and marketing systems.

Significant impact is expected on society with opportunities for a wide range of applications:

- **Big Data Value technologies** will be a key contributor to solutions for major societal challenges, in areas such as health, demographic change, climate change, transport, energy, and cities. Novel Big Data technologies will provide insight on the different aspects of societal challenges and unlock new potential solutions to address them. Similarly, BDV is associated other areas such as the Future Internet and the Internet of Things. In these emerging markets integration of huge volumes of data needs to be supported by solid data-orientated technologies. All these solutions will lead to a transformation of our everyday lives with direct impact on an individual's behaviour and habits. In the future, citizens can expect benefits from a more personalized healthcare system, novel decision-support systems for their everyday life or new ways to interact with companies and administrations, based on Big Data Value solutions.
- **Availability of public government information and open data** will influence educational and cultural services. Large databases containing information on cultural heritage such as digitalized books and manuscripts, photos and paintings, television and film, sculpture and crafts, diaries and maps, sheet music and recordings will be made available and allow for new ways of educating people and novel forms of interaction between people across cultural borders.

- **Big Data technology will improve societal insight** on individual and collective behaviour. Such technologies may allow for greater fact-based decision-making in politics and the economy. Fundamental research will be deeply impacted by the availability of Big Data resources and analysis, providing new insights and new development in many areas such as biology, physics, mathematics, material, and energy. These developments themselves will produce new Big Data and further enhance societal developments.
- **Collaboration**. Big Data Value will help to improve collaboration by providing access to various data sources such as media content, traffic flow, etc. Better services and collaboration will be possible for instance in emergency and crisis situations. Individuals will be empowered by their new role as co–creator or co-innovator as well as generator and provider of personal data.

Industry surveys show that the gains from Big Data Value are expected across all sectors, from industry and production to services and retail. The following are examples of sectors that are especially promising with regard to Big Data Value.

- **Environment:** Better understanding and management of environmental and geospatial data is of crucial importance. Environmental data helps to understand how our planet and its climate are changing and also addresses the role humans play in these changes. For example, the European Earth observation programme, Copernicus, aims to provide reliable and up-to-date information on how our planet's climate is changing to provide a foundation, which will support the creation of sustainable environmental policies. In addition, the EU project Galileo will offer a global network of satellites providing precise timing and location information to users on the ground and in the air. The overall intention is to improve the accuracy and availability of location data to the benefit of the sectors including transport and industry as well as Europe's new air-traffic control system.
- **Energy**: The digitization of the energy system from production, to distribution, to smart meters at the consumer, enables the acquisition of real-time, high-resolution data. Coupled with other data sources, such as weather data, usage patterns and market data, accompanied with advanced analytics, efficiency levels can be increased immensely. Existing grid capacities could be better utilized, and renewable energy resources could be better integrated.
- **Mobility, transport and logistics**: Urban multimodal transportation is one of the most complex and rewarding Big Data settings in the logistics sector. In addition to sensor data from infrastructure, vast amounts of mobility and social data are generated by smart phones, C2x technology (communication among and between vehicles), and end-users with location-based services and maps. Big Data will open up opportunities for innovative ways of monitoring, controlling and managing logistical business processes. Deliveries could be adapted based on predictive monitoring, using data from stores, semantic product memories, internet forums, and weather forecasts, leading to both economic and environmental savings.
- **Manufacturing and production**: With industry's growing investments into smart factories with sensor-equipped machinery that is both intelligent and networked (Internet of Things, Cyber-Physical Systems), the production sectors in 2020 will be one of the major producers of (real-time) data. The application of Big Data into this sector will bring efficiency gains and predictive maintenance. Entirely new business models are expected since the mass production of individualized products becomes possible where consumers may have direct access to influence and control.
- **Public Sector**: Big Data Value will contribute to increased efficiency in public administrations processes. The continuous collection and exploitation of real-time data from people, devices and objects will be the basis for smart cities, where people, places and administrations get connected through novel ICT services and networks. In the physical and the cyber-domain, security will be significantly enhanced with Big Data techniques; visual analytics approaches will be used to allow algorithms and humans to cooperate. From financial fraud to public security, Big Data will contribute to establishing a framework that enables a safe and secure digital economy.
- **Healthcare**: Applications range from comparative effectiveness research to the next generation of clinical decision support systems, which make use of comprehensive heterogeneous health data sets as well as advanced analytics of clinical operations. Of particular importance are aspects such as patient involvement, privacy and ethics.
- **Media and Content**: By employing Big Data analysis and visualisation techniques, it will be possible to allow users to interact with the data, and have dynamic access to new data as they appear in the relevant repositories. Users would be able to register and provide their data or annotations to existing

data. The environment will move from a few state-orientated broadcasters to a prosumer approach, where data and content are linked together blurring the lines between data sources and modes of viewing. Content and information will find organisations and consumers, rather than vice versa, with a seamless content experience.

- **Financial services**: Huge amounts of data are processed to detect fraud and risk, to analyse customer behaviour, segmentation, trading, etc. Big Data analysis and visualization will open up new use cases and permit new techniques to be realised. Possibilities include managing regulation, reporting, audits and compliance, and automatic detection of behaviour patterns and cyber-attacks. Open sources of information can be combined with proprietary knowledge to analyse competitive positions, and recommendation engines will be able to identify potential customers for products.
- **Telecommunications services**: Big Data enables improved competitiveness by transforming data into customer knowledge. Possible use cases can include improvement of service levels; churn reduction, services based on combining location with data about personal context, and better analysis of product and service demand.
- **Retail**: Digital services for customers provided by smart systems will be essential for the success of future retail businesses. The retail domain will especially be focused on highly efficient and personalized customer assistance services. Retailers are currently confronted with the challenge to meet the demand of a new generation of customers who expect information to be available anytime and anywhere. New intelligent services that make use of Big Data will allow a new level of personalized and high-quality Efficient Consumer Response (ECR).
- **Tourism**: Personalized services for tourists are essential for creating real experiences within a powerful European Market. The analysis of real-time and context-aware data with the help of historic data will provide customized information to each tourist and contributes to a better and more efficient management of the whole tourism value chain. The application of Big Data in this sector will enable new business models, services, and tourism experiences.

5.2 Monitoring of objectives

Minor updates, i.e. updated figures based on the 2016 IDC report as well as added numbers of the monitoring report.

Big Data Value generation and the technology for it will have a tremendous impact on industry and economy as a whole. In terms of measuring this impact there are two basic types of measurements with related indicators:

- **Indirect Monitoring**: The monitoring is done using indicators that cannot directly be influenced or monitored by activities resulting from this Big Data Value SRIA. Typically, the monitoring is based on tracking the progress of some developments and is using comparison rather than specific numbers or targets. For example, the proposed Big Data Value activities can provide a research and innovation ecosystem, but ultimately jobs, sales information and business progress will be under the control of individual organisations. Indirect indicators include for instance economic and usage information. The success of the SRIA strategy will be mainly assessed based on indirect indicators.
- **Direct KPIs:** Key Performance Indicators (KPIs) that are directly related to the performance of the SRIA activities themselves and are clearly measurable. For example, providing solutions to the technical priorities or the stimulation of SME participation in research and innovation activities and i-Spaces.

According to the strategic and specific objectives of the Big Data Value SRIA described in Section 1.6 an interdisciplinary and holistic approach will be followed. Consequently, the indicators to be used for assessing the impact of the SRIA have to address strategic, social, competitiveness, and innovation aspects.

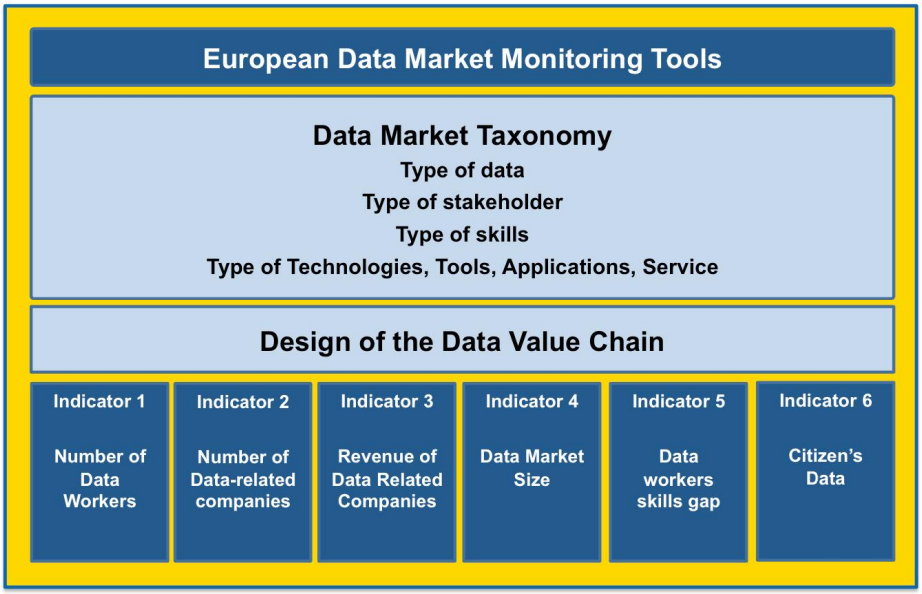

Figure 10: Preliminary European Data Market Indicators; IDC, 2014

Indicators to measure the achievement of the strategic objectives

The development of the BDV market will be pushed by new, innovative and novel products. However, the success of those developments depends heavily on various market conditions and the overall economic climate. IDC[30] has developed a European Data Market Monitoring tool with European Data Market Indicators (see Figure 10) that are in close alignment with certain strategic objectives and KPIs specified in the Contractual Arrangement (CA) of the PPP. Therefore, the SRIA proposes to use this kind of market metrics as indirect indicators for monitoring its strategic impact. The KPIs and their monitoring will need to be adapted in alignment with the on-going development of the European Data Market Monitoring Tool based on the assumption that based on the tool the development of the market can be observed throughout the lifetime of the PPP. The following table lists the relevant KPIs of the CA (the numberings are derived from the arrangement and the IDC report and therefore non-consecutive as the table focuses on the KPIs of the Contractual Arrangement that are relevant in this regard). Furthermore it shows the measurements available for 2013 and 2014:

Strategic Indicators			Societal	Competitiveness	Innovation	Operational
CA-KPI[31]	IDC-KPI	Description and initial measurements				
KPI.CA.1 (II-1)		Market Share of the European suppliers of the global big data market in 2020 The IDC report[30] does not provide any figures to measure this KPI however, it presents a comparison on the assessed indicators for the EU, the US, Japan and Brazil. For details, see the IDC-report on the European Data Market[32, 33].				

[30] "*European Data Market*", Gabriella Catteneo, IDC et al., SMART 2013/0063, D6 – First Interim Report, 16 October 2015

[31] Numbers in parentheses refer to the numbering scheme used in the Monitoring Report of the PPP.

[32] "*European Data Market"*, Gabriella Catteneo, IDC et al., SMART 2013/0063, D6 – First Interim Report, 16 October 2015, p.183ff.

[33] "*European Data Market"*, Gabriella Catteneo, IDC et al., SMART 2013/0063, D8 – Second Interim Report, 09 June 2016

KPI.CA.5 (II-2)	2.1	Number of European companies offering data technology, application, and services, including start-ups. According to IDC[30, 33] the total number of data companies in the EU measured as legal entities based in one EU country increased from 239,846 in 2013 to 243,600 in 2014, and to 249,100 in 2015 representing an total increase of 3.9% since 2013.				
KPI.CA.6 (II-3)	3.1	Increased revenue generated by European data companies compared to baseline in relative and absolute terms. The total revenues of the data companies defined by KPI.CA.5, as estimated by IDC[30, 33], sum up to 47,727 M EUR in 2013 and increased to 51,112 M EUR in 2014 and to 56,029 MEUR resulting in a growth rate of 17.4% since 2013 which represents an absolute increase of 8,302 M EUR.				
KPI.CA.8 (II-5)	1.1	Increased Number of European data workers. IDC[30, 33] estimates an overall number of 5.7 million data workers in 2013 and 6.0 million data workers in 2015 representing an increase of almost 4% since 2013.				

Direct KPIs to measure the achievement of the specific objectives

The SRIA activities will deliver solutions, architectures, technologies and standards for the data value chain over the next decade. The following KPIs are proposed to frame and assess the impact of those SRIA activities.

	Direct KPIs		**Societal**	**Competitiveness**	**Innovation**	**Operational**
Business	KPI.D.1 (II-11)	**At least 50 large-scale experiments are conducted in i-Spaces involving closed data.** Multiple SMEs should be encouraged to perform experiments by using i-Spaces. This will foster their growth from small companies into larger ones and/or their expansion from national markets to the EU (or even global) market. The i-Space and the residing experiments will provide a unique opportunity for exploitation.				
	KPI.D.2 (II-12)	**30% year-on-year increase in Big Data Value use cases supported in i-Spaces.** The number of use cases within the large-scale experiments will be an indicator of acceptance and will also prove the innovative capacity of the BDV partnership. An ever-expanding increase will guarantee a continuous value creation out of Big Data and will speed up the innovation process, thus also addressing the time to market. It will support market development in existing industries and potentially in establishing entirely new business models.				
Skills	KPI.D.3 (II-8)	**At least 50 training programs are established with participation of at least 100 participants per training session arising from the PPP.** Continuous development of skills and competences on the basis of the Big Data Value PPP will be supported by training and education activities. An appropriate environment (e.g. e-learning platform, contribution to University curricula) should be created to attract potential participants. This broadens the number of skilled people and serves as a unique opportunity to create new jobs and start-ups as a result of the PPP activities.				

	KPI.D.4 (II-8)	**At least 10 European training programs involving 3 different disciplines with the participation of at least 100 participants**. These interdisciplinary programs will contribute to knowledge and skills needed to deal with the complexity of Big Data. To broaden the number of students, Massive Open Online Courses (MOOC) would be proposed building on the diversity of skills and European multiculturalism				
Applications	KPI.D.5 (II-13)	**At least 10 major sectors and major domains are supported by Big Data technologies and applications developed in the PPP**. The usage of BDV technologies and applications developed in the PPP in different sectors will lead to increased value generation and finally to job growth in all the addressed sectors. The broad take-up of those technologies and applications across a number of sectors is also an indicator for efficient sharing of best practices and expertise leading to a build-up of a broad skills base. Furthermore, cross-sector activities should prove domain independent and cross-domain deployment leading to standards.				
Data	KPI.D.6 (II-14)	**Total amount of data made available to i-Spaces - including closed data – is in the 10x Exabyte range.** Experiments conducted in i-Spaces benefit from their scale, amount of different but integrated data sources, and especially on the value of data. This is key to accelerate Data Driven Innovation in Europe in liaison with Research and Education, with major advances expected in data management techniques, semantics, analytics, data learning, visualization. This includes both economic and societal objectives. These experiments will also contribute to the advances in data governance practices, such as encryption, pseudonymisation, anonymisation for better Intellectual Property and Privacy Protection.				
	KPI.D.7 (II-15)	**Availability of metrics for measuring the quality, diversity and value of data assets.** It is not only the amount of data made available to perform data analysis; of utmost importance are the quality, diversity and value of the data. The ultimate goal is to create value out of Big Data, to derive analytical findings on a minimal, yet most significant data set, allowing faster data processing and management of data for data analytics. During the PPP relevant metrics will be derived.				
Technical	KPI.D.8 (II-16)	**The speed of data throughput is increased by 100 times compared to 2014.** One of the main problems regarding today's data storage and processing techniques is the time required for accessing large datasets in order to analyse them. Techniques to be implemented in the scope of the Data Management priority will make data access for analysis much more efficient.				
	KPI.D.9 (II-6)	**The energy required to process the same amount of data is reduced by 10% per year.** One of the main problems today is the energy consumed processing data due to the huge amount of data and lack of algorithms coupled with new hardware designed devices that will reduce the energy required to process data. Beyond hardware optimization, new tools and algorithms will require fewer resources and time to provide the same quality of analytics.				
	KPI.D.10 (II-4)	**Enabling advanced privacy and security respecting mechanisms (including anonymisation) for data access, process and analysis.** The availability of suitable privacy and security respecting mechanisms will encourage data users to provide closed data for experiments and analyses in i-Spaces and lighthouses.				

Additional KPIs listed in the Contractual Arrangement

Besides the KPIs listed above the article 7 of the Contractual Arrangement of the PPP specifies further KPIs. For the sake of completeness the following table contains the full list of KPIs specified in article 7, and when applicable, we refer to the two KPI tables above.

KPIs in article 7 of the Contractual Arrangement		Societal	Competitiveness	Innovation	Operational
CA-KPI	Description				
KPI.CA.1	see corresponding strategic KPI				
KPI.CA.2 (I-4)	**PPP investments leveraged through sector investments by 4 times the PPP's total budget**				
KPI.CA.3 (I-5)	**SMEs participating in the PPP projects under this initiative represent at least 20% of participant organizations**				
KPI.CA.4 (I-3)	**Increased competitive European provision of big data value creation systems and technologies**				
KPI.CA.5	see corresponding strategic KPI				
KPI.CA.6	see corresponding strategic KPI				
KPI.CA.7	see KPI.D.10				
KPI.CA.8	see corresponding strategic KPI				
KPI.CA.9	see KPI.D.9				
KPI.CA.10 (II-7)	**New economically viable services of high societal value developed by PPP projects**				
KPI.CA.11	see KPI.D.3 and KPI.D.4				
KPI.CA.12 (II-9)	**Ensure efficiency, transparency and openness of the PPP's consultation process**				
KPI.CA.13 (II-10)	**Ensure that the technology is in line with the established multi-annual roadmap**				

6 Annexes

6.1 Acronyms and Terminology

Acronym/Term	Name/Description
General	
API	Application Programming Interface
BDV	Big Data Value
BDVA	Big Data Value Association
BPM	Business Process Management
CASD	Secure Remote Data Access Centre
PPP	Public Private Partnership
CSA	Coordination and Support Action
CEP	Complex Event Processing
DSMS	Data Stream Management Systems
EIP	European Innovation Partnership
EU	European Union
ETP	European Technology Platform
FI	Future Internet
FIRE	Future Internet Research & Experimentation
GDP	Gross Domestic Product
ICT	Information Communication Technologies
IoT	Internet of Things
IPR	Intellectual Property Rights
i-Space	(European) Innovation Space
KPI	Key Performance Indicators
MOU	Memorandum of Understanding
MPP	Massively Parallel architectures
NoSQL	Not only SQL (referred to databases)
SME	Small and Medium sized Enterprise
SRIA	Strategic Research & Innovation Agenda
SWOT	Strengths, Weaknesses, Opportunities and Threats
Data Orientated	
Open Data	Data available to everyone to use and republish
Private Data	Data which is generated by organisations, typically companies and in particular users, which and has not been made "open" and often is kept internally or has restricted conditions around it (e.g. NDAs)
Public Data	Freely reusable datasets from local, regional and national public bodies. Public Data is generally also Open Data
Closed Data	Data that has restrictions on its access or reuse (i.e. charges, technology, memberships, etc.). Typically Closed Data include Private Data
Free Data	Data that can be accesses or reused without a charge
Non-Free Data	Data which has a charge associated with use or reuse

6.2 Contributors

SRIA Version 3 in 2017

The following individuals and organisations are thanked for their involvement in creating this updated version of the SRIA document or other documents to which it heavily relates

SRIA Core Team		
Sonja Zillner	Co-Editor	Siemens AG
Edward Curry	Co-Editor	Insight @ NUI Galway
Andreas Metzger	Co-Editor	Paluno, Univ. Duisburg-Essen
Sören Auer	Co-Editor	Fraunhofer IAIS
Contribution		
Dirk Mayer	Contributor	Software AG
Nuria de Lama	Contributor	ATOS
Jim Keneally	Contributor	Intel
Souleiman Hasan	Contributor	Insight @ NUI Galway
Dumitru Roman	Contributor	SINTEF
Carlos A. Iglesias	Contributor	UPM
Meilof Veeningen	Contributor	Philips
Bjarne Kjær Ersbøll	Contributor	DTU
Giuseppa Caruso	Contributor	Engineering Ingegneria Informatica spa
Bas Kotterink	Contributor	TNO
Ernestina Menasalves	Contributor	UPM
Wolfgang Gerteis	Contributor	SAP
Other Participant involved in this update of the SRIA		
Numerous BDVA members participating in the various BDVA task forces working on state-of-the art research related to the SRIA priorities		
All BDVA members participating in the BDVA Member Expression of Interest as well as the participants of the BDVA Community Survey that was open from June to October 2016		
More than 200 participants of the BDVA Mini Summit in March 2016 in Den Hague and more than 350 participants of the BDVA summit in November 2016 in Valenica that have actively contributed to a large number of workshops dedicated to the various technical and non-technical topics. Workshop outputs that were appropriate were used as input for the SRIA update.		

SRIA Version 2 in 2016

The following individuals and organisations are thanked for their involvement in creating this updated version of the SRIA document or other documents to which it heavily relates

SRIA Core Team		
Sonja Zillner	Co-Editor	Siemens AG
Edward Curry	Co-Editor	Insight @ NUI Galway
Arne Berre	Co-Editor	SINTEF
Andreas Metzger	Co-Editor	Paluno, Univ. Duisburg-Essen
Colin Upstill	Co-Editor	IT Innovation
Contribution		
Wolfgang Gerteis	Contributor	SAP
Ernestina Menasalves	Contributor	UPM
Nuria de Lama	Contributor	ATOS
Pierre Pleven	Contributor	INSTITUT MINES TELECOM Paris
Corinna Schulze	Contributor	SAP
Freek Bomhof	Contributor	TNO
Carlos A. Iglesias	Contributor	UPM
Souleiman Hasan	Contributor	Insight @ NUI Galway

Dumitru Roman	Contributor	SINTEF
Aris Gkoulalas-Divanis	Contributor	IBM Research
Bjarne Kjær Ersbøll	Contributor	DTU
Other Participant involved in this update of the SRIA		
Numerous BDVA members participating in the various BDVA task forces working on state-of-the art research related to the SRIA priorities		
All BDVA members participating in the BDVA Member Expression of Interest as well as the participants of the BDVA Community Survey that was open from June to October 2015		
More than 300 participants of the BDVA Summit in June 2015 in Madrid that have actively contributed to a large number more than 60 workshops. Workshop outputs that were appropriate were used as input for the SRIA update.		
Other European Technology Platforms, for example ETP4HPC,that contributed through joint workshops and discussion in particular focusing on the alignment of related technical priorities and requirements		

SRIA Version 1 in 2015

The following individuals and organisations are thanked for their direct involvement in creating the first version of the SRIA document or other documents to which it heavily relates

SRIA Core Team		
Nuria de Lama	Co-Editor	ATOS
Julie Marguerite	Co-Editor	Thales
Klaus-Dieter Platte	Co-Editor	SAP
Josef Urban	Co-Editor	Nokia
Sonja Zillner	Co-Editor	Siemens AG
Edward Curry	Co-Editor	Insight @ NUI Galway
Primary Editing Team		
Antonio Alfaro	Contributor	Answare
Ernestina Menasalves	Contributor	UPM
Andreas Metzger	Contributor	Paluno, Univ. Duisburg-Essen
Robert Seidl	Contributor	Nokia
Colin Upstill	Contributor	IT Innovation
Walter Waterfeld	Contributor	Software AG
Stefan Wrobel	Contributor	Fraunhofer IAIS
Contribution		
Paolo Bellavista	Contributor	CINI
Stuart Campbell	Contributor	TIE Kinetix / BDVA SG
Thomas Delavallade, Yves Mabiala	Contributor	Thales
Nuria Gomez, Paolo Gonzales, Jesus Angel	Contributor	INDRA
Thierry Nagellen	Contributor	Orange
Dalit Naor, Elisa Molino	Contributor	IBM Research
Stefano de Panfilis, Stefano Scamuzzo	Contributor	Engineering
Nikos Sarris	Contributor	ATC
Bjørn Skjellaug, Arne Berre, Titi Roman	Contributor	SINTEF
Tonny Velin	Contributor	Answare
Alexandra Rosén, Francois Troussier	Contributor	NESSI Office
Other Participants involved in SRIA, Proposal or other related documents and discussions		
Volker Markl (TU Berlin), Burkhard Neidecker-Lutz (SAP), José María Cavanillas (ATOS), Roberto Martínez (UPM), Michael May (Siemens), Wolfgang Wahlster (DFKI) and Brigitte Cardinael (Orange).		
The 200 active participants of the NESSI Organised Big Data Value Workshops, organised between February and March 2014.		
All participants of the Public Consultation on the SRIA that was available at www.bigdatavalue.eu from 9 April to 15 May 2014.		

Numerous NESSI Partners and members as well as other contributors and their personnel
The partners of the EU project BIG.
Other European Technology Platforms, which through discussions contributed to defining the content of this SRIA, for example the ETPs NEM, ETP4HPC and Net!works.

6.3 SRIA Preparation Process and Update Process

SRIA Preparation Process

Within the SRIA preparation process, the proposers have heavily engaged with the wider community. Multiple workshops and consultations took place to ensure the widest representation of views and positions including the full range of public and private sector entities. These have been carried out in order to identify the main priorities with approximately 200 organisations and other relevant stakeholders physically participating and contributing. Extensive analysis reports were then produced which helped both formulate and construct this SRIA.

The series of workshops gathered views from different stakeholders in the existing value chains of different industrial sectors, including: energy, manufacturing, environment and geospatial, health, public sector, content and media. Additional workshops were organized to gather feedback on cross-sectorial aspects, for example, the view of SMEs. The selection of sectors was based on the criteria of their weight in the EU economy and potential impact of their data assets (source: demosEUROPA). The community involved in the Workshops included: Actors such as AGT International (DE), Hospital de la Hierro (ES), Press Association (UK), Reed Elsevier (NL); BIODONOSTIA (ES), Merck 8 (ES), Kongsberg Group (NO), and many more.

In addition, NESSI together with partners from the FP7 project BIG[34], ran an online public consultation on the BDV Strategic Research and Innovation Agenda between 9 April and 15 May 2014. The aim was to validate the main ideas put forward in the SRIA on how to advance Big Data Value in Europe in the next 5 to 10 years. 195 organisations from all over Europe participated in the consultation including companies such as Hitachi Data Systems, OKFN Belgium, TNO Innovation for Life, Euroalert, Tecnalia Research and Innovation, ESTeam AB, and CGI Nederland B.V. Furthermore, another ~20 organisations and companies such as Wolters Kluwer Germany, Reed Elsevier and LT-Innovate shared in more detail their views on the content of the SRIA.

Although the primary target is to create impact at a European-level, cooperation with stakeholders outside Europe will allow the transfer of knowledge and experiences around the globe. For future collaborations, NESSI has already set-up links to the following regions through NESSI partners: Mediterranean countries[35], LatAm countries[36], South East Asian countries[37] and the Russian speaking countries[38].

SRIA Update Process

The Big Data Value Association (BDVA)[39] is responsible for providing regular (yearly) updates of the SRIA defining and monitoring the priorities as well as metrics of the PPP.

[34] M. Cavanillas, E. Curry, W. Wahlster: New Horizons for a Data-Driven Economy – A Roadmap for Big Data in Europe, Springer International Publishing, 2016.
[35] MOSAIC (http://www.connect2sea.eu/) and MED-Dialogue (www.med-dialogue.eu)
[36] CONECTA 2020 (www.conecta2020.net)
[37] CONNECT2SEA (www.connect2sea.eu)
[38] EAST HORIZON (www.eeca-ict.eu/about/new_projects/easthorizon)
[39] In order to establish a contractual counterpart to the European Commission for the implementation of the PPP, the Big Data Value Association, a fully self-financed non-for-profit organisation under Belgian law, was founded by 24 organisations including large, SMEs and research organisations.

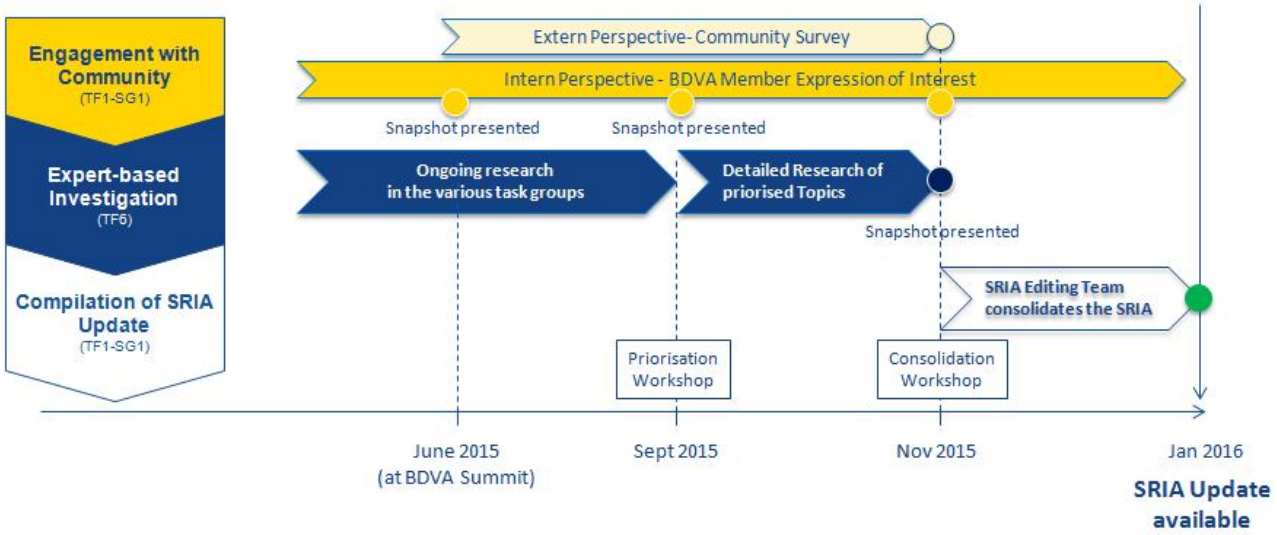

Figure 11: Overview of the annual SRIA Update Process in 2015

The purpose of update is to understand whether the SRIA – e.g. the technical & non-technical priorities -- needs to be update:

- How important are the priorities already covered in the SRIA?
- Are there other priorities relevant for BDV SRIA? How important are those priorities?

Engagement with the community: Within the update process, the BDVA engages with the a) BDVA members as well as b) the wider community to ensure a comprehensive perspective concerning the technical and business impact of the SRIA technical and non-technical priorities as well as to identify emerging priorities with high impact.

For engaging with the internal and wider community, two parallel interaction streams have been set up

- BDVA Member Expression of Interest: Each BDVA member is requested to express interest of research and innovation activities once a year (only one consolidated vote per member organization will be reflected). This interaction stream doesn't follow a fixed deadline. This is motivated by the fact that the BDVA is a fast growing community and it should be possible to include new members' opinion as soon as possible. In order to learn about the BDVA members' expressed interest snapshots of the survey results are taken on a regular basis
- Community Survey: The community survey is open to everybody for a fixed period of time. The consolidated votes of the wide community will establish an important outside perspective.

The results of both community engagement streams provide important insights about the relevance of covered SRIA priorities and highlight emerging topics that required detailed analysis.

Expert-based Investigations: The BDVA has established task groups for all technical and non-technical priorities. The task groups are continuously working in order to produce related state-of-the-art analysis or working papers.

In accordance to the outcome of the community engagement process, the particular task groups are consolidated in order to discuss in which detail and scope the SRIA needs to be updated. A dedicated *prioritisation workshop* was organized. In accordance to the outcome of the workshop, the task groups conduct a detailed state-of-the-art research on the agreed priorities which were presented in a dedicated SRIA Update *consolidation workshop*.

6.4 Big Data in Europe - Strengths, Weaknesses, Opportunities and Threats

The priorities identified in this Strategic Research and Innovation Agenda reflect the views of industry, research organizations and academia, representing providers and users of technologies and data assets in many sectors. A number of workshops were organised in order to ensure that the objectives set out in this SRIA are based on the real needs of both public and private entities in Europe.

The main task of each workshop was to identify the main priorities and a SWOT analysis for each of the sectors, including consideration of the benefits derived from cross-sector fertilization. The workshops addressed different industrial sectors, including energy, manufacturing, environment and geospatial, health, public sector, content and media. In addition to the sector workshops, additional workshops were organized to gather feedback on cross-sector aspects and the views from SMEs. A compilation of the workshop results is provided in the following pages as an integrated SWOT analysis for the European market.

These views form the basis for the strategic and specific objectives for the SRIA, set out in Section 1.6.

Strengths

European Aspects:

- Compared to the rest of the world, Europe has a strong medium-sized sector with regard to Big Data.
- Europe offers a stable environment in terms of life standards, currency, etc.

Market and Business:

- There is a specific European capacity that allows for companies to start in niches and then grow their business potential.
- There are many SMEs that are dynamic and flexible and can react quickly to market changes.
- There is an existing and strong content/data market in Europe.
- There are established cooperation networks between content providers in several domains.

Technical:

- Computer clusters and cloud resources are readily available.
- There is a growing interest in archiving, sensing, behavioural data, and personal data.

Data and Content:

- There is a large amount of content and data available – the issue is making use of it.
- There are already a number of existing ecosystems and portals (for example INSPIRE[40], Copernicus[41] and GEOSS[42]).
- Geospatial and environmental data sets and supporting infrastructure data are available.

Education and Skills:

- There is a broad and detailed domain know-how as well as process know-how available.
- Many domains have innovative technology and skilled people.
- There are many universities with high capacity where skills can be developed.
- Good engineering /domain specific education can be obtained.

Policy, Legal and Security:

- The European Union promotes free and open processes.

Weaknesses

European Aspects:

- Europe is decentralized which can lead to disparate policies.
- Some domains are characterized by conservatism and long innovation cycles.
- There is a lack of a solid start-up culture because of risk aversion and intolerance of failure.
- There are few European data analytics solution providers.

[40] http://inspire.ec.europa.eu/
[41] http://www.copernicus.eu/
[42] https://www.earthobservations.org/geoss.shtml

- There are few large companies to lead the market, and many small sized companies that need nurturing.

Market and Business:

- There is a lack of access to Big Data facilities that make data more easily accessible.
- There is no visibility of ecosystem service offerings.
- Is unclear what data should be preserved, and for how long, in all the different sectors and markets.

Technical:

- Lack of processable linked data, and of aggregated/combined data.
- Lack of seamless data access and inter-connectivity, and low levels of interoperability: data is often in silos and data sharing is difficult due to a lack of standards e.g. formats and semantics.
- Migration of data between systems, versions or partners is challenging.
- Access and processing of data sets that are too big to be given to the end user.

Data and Content:

- Public data in EU is not available to the extent it should be.
- The quality of data in open data portals is often very low.
- The different languages within Europe create a barrier (multilingualism) during data processing.
- Structural data sources often lack precise semantics e.g. labels from ontologies.
- Poor and inconsistent use or management of metadata.

Education and Skills:

- There is a lack of specialised education programs for data analysts.
- There are not enough skilled people to participate in training programmes.

Policy, Legal and Security:

- Legislative restrictions on data sharing decrease availability across Europe and makes European-focused initiatives that address these issues more difficult.
- Rules and regulations are fragmented across Europe.
- There are high security demands that can be difficult to address.

Usage:

- Europe is not good at analysing and changing consumer behaviour.
- Citizen science – how to qualify and use data from citizens.
- Providing Big Data (Value) for SME use.

Opportunities

European Aspects:

- Various cultures and various strengths can result in creative thinking if they are mixed.
- The existence of BDV topics and best practice examples in other initiatives can lead to synergies.
- Strengthening the European market, e.g. by fusing the emerging start-up nucleus.
- Create lots of SMEs for the low hanging fruits of Big Data for which agility is required.
- Investment in the entire innovation chain, beyond basic research.
- Investment support mechanisms for SMEs (e.g. European loans).

Market and Business:

- Opening up of private content to extend and complement existing assets.
- Increasing the use of analytics.
- Many opportunities exist for particular sectors, for example, environmental monitoring, social media, industrial processing.
- There is a potential for extending INSPIRE and Copernicus.

- Improve creativity to create cost-effective solutions.
- There is the opportunity to open up completely new and different business areas and services.
- New applications can be created throughout the Big Data ecosystem, ranging over acquisition, data extraction, analysis, visualization and utilisation.

Technical:

- Easier syndication of data and content.
- Micropayments for processed data or the results from analytics.
- Wearable sensors and sensor technologies become mainstream generating more data.
- The explosion of device types opens up access to any data from any device for greater and more varied usage.
- Development of APIs for access becoming standardised and available.
- Interoperability tools and standardised APIs to facilitate data exchange.
- Greater visibility and increased use of directory services for data sources.

Data and Content:

- Making use of European cultural and data assets.
- Use semantics to align content from various data sources.
- Providing facilities to better navigate and curate data.
- Contextualisation and personalisation of data.
- The evolution of different sectors and the increased volume of data enable innovative applications to be developed.

Education and Skills:

- Exploring new research areas.
- Training focussed on innovation in BDV.
- Use and exploration of Big Data to be ubiquitous in education and training.

Policy, Legal and Security:

- Address the safe and secure storage of data on a European basis.
- Develop uniform policies for data access in Europe to help build competitive capabilities.

Usage:

- User generated and crowd-sourced content increasingly available.
- Data-as-a-service can significantly lower the market entry barriers (in particular to new markets).
- Shift from technology push to end-user engagement.
- Create rich and complex data value chains.

Threats

European Aspects:

- Europe is lagging behind the US in the Big Data market.
- US players and their bottom-up ideas are dominating the market.
- Europe does not have a Big Data and data-sharing culture.

Market and Business:

- Dominant large corporations own important data.
- Consolidation of stakeholders and marketplaces are reducing competition.
- There are several barriers to market entry for SMEs, e.g. owning data.

Education and Skills:

- Many skilled professionals leave Europe to work in other regions; there is a risk of a "Brain Drain" in Europe.
- Continuous lack of skilled professionals and graduates.

Policy, Legal and Security:

- Policies are often too connected to the 'old data' world.

- Complete analysis of ethical and privacy issues are needed.
- There is a risk of over-regulation and protectionism in Europe; privacy regulations elsewhere are too permissive.
- Policies of data availability; for example companies are not willing to make data available 'just-in-case' it may cause harm in another territory.

Usage:

- Cross-border data flows.
- Data-driven services are not tied to a particular location, but are subject to different legislation in different countries.

6.5 History of document changes

HISTORY OF CHANGES		
Version	**Publication**	**Changes**
1.0	January 2015	▪ Initial version
2.0	January 2016	▪ The main changes compared to version 1 of BDV SRIA document are as follows ○ In Section 1: increase of structure by integrating more subheadings - Section 1.1: most relevant Big Data market numbers were consolidated and updated - Section 1.3: adjustment of argumentation to reflect the situation that the PPP is already running for one year - Section 1.4: a more condensed version of the original section - Section 1.5: a new section covering the objectives of the Contractual Arrangement was added - Section 1.6: a new section documenting the SRIA document history was added. ○ Section 2 - Section 2.1: the section about I-Spaces and Lighthouse projects were updated by incorporating respective material from BDVA task forces - Section 2.2: a new section describing the BDV methodology was added - Section 2.3: text from the original C-PPP document was re-used to improve the stakeholder platform description; a sub-section describing the ongoing cooperation with ETP4HPC was added ○ Section 3: - Section 3.2-3.6: several updates motivated by the SRIA survey results and proposed by the BDVA task forces have been incorporated, overlaps across technical priorities have been removed, and titles of sections have aligned to achieve consistency - Section 3.7 an innovation roadmap derived from SRIA survey results was added ○ Section 4 - Section 4.1, 4.3, and 4.4 have been improved by including information about already existing approaches, more precise definition, and relevant background information ○ Section 5: - The indicators represented in Section 5.2. have been updated to achieve a strong alignment with the KPIs specific in the Contractual Arrangement; KPI.D.6 was adjusted to "10x Exabyte range" in order to establish a sound basis for evaluation - To ensure completeness, the KPIs in Article 7 of the Contractual Arrangement were listed ○ The Appendix was extended by several sections - Appendix 6.3 provides a detailed description of the BDV SRIA Update process - Appendix 6.4. is covering Big Data in Europe SWOT analysis that was originally part of the introduction - Appendix 6.5 encompasses the history of document changes ▪ Other minor drafting changes and corrections of clerical mistakes have been accomplished across the document
3.0	January 2017	▪ The main changes compared to version 1 of BDV SRIA document are as follows ○ Section 1: the recent developments of the European data market have been reflected throughout the whole section - In Section 1.5: the objectives have been described in a more specific way ○ Section 2 - Section 2.1: definition of I-Spaces was updated in accordance to the discussion in the I-Spaces taskforce - Section 2.2. information about the future nature of Lighthouse projects has been added - Section 2.3. listing the projects that have been funded in the 2016 calls was added - Section 2.4. encompasses an update concerning the collaboration between ETP4HPC and BDVA ○ Section 3: - Section 3.2-3.6: The outcome descriptions have been consolidated in order to priorize technical

		aspects - Section 3.7. the innovation roadmap was updated and reflected based on a data analysis on tops of the survey results of 2015 and 2016. ○ Section 4 - Section 4.1: Activities related to skill development covered in the CSA BDVe have been incorporated - Section 4.2: Minor updates fostering alignment with ongoing activities in the Business taskforce. ○ Section 5: Minor updates to incorporate recent numbers from IDC report 2016. ▪ Other minor drafting changes and corrections of clerical mistakes have been accomplished across the document