BDV BIG DATA VALUE
ASSOCIATION

European
Big Data Value Strategic
Research & Innovation
Agenda

VERSION 1.0

January 2015

Big Data Value Europe

Rue de Trèves 49/5, B-1040 BRUSSELS

Email: info@bigdatavalue.eu – www.bigdatavalue.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 contractual Public Private Partnership (cPPP) on Big Data Value. The SRIA has been proposed by a partnership of European Big Data stakeholders that was initially led by NESSI, the European Technology Platform (ETP) for software, services and data, the partnership has been extended and formalised as non-profit organisation, the Big Data Value Association (BDVA).

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, analyses the associated strengths, weaknesses, opportunities and threats, and sets out the objectives and goals to be accomplished by the cPPP 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 cPPP 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 out into specific competitiveness objectives, innovation and technology objectives, societal objectives 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 for the development of skills, competence and best practices. Lighthouse projects are large-scale data-driven innovation and demonstration projects that will create high-level visibility, awareness and impact.

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: deep analysis, to improve data understanding; optimized architectures for analytics of data-at-rest and data-in-motion; mechanisms for managing privacy 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.

Contents

1 Introduction – The strategic importance of Big Data

The economic potential of Big Data

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

In parallel with the continuous and significant growth of data has come better data access, availability of powerful ICT[1] systems, and ubiquitous connectivity of both systems and people. This has led to intensified activities around Big Data and Big Data Value. **Powerful data tools** have been developed to collect, store, analyse, process, and visualize huge amounts of data. **Open data initiatives** have been launched to provide broad access to data from the public sector, business and science.

The volume of data is rapidly growing: it is expected that **by 2020 there will be more than 16 zettabytes** of **useful data** (16 Trillion GB)[2], 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.

It is clear that Data is now an asset that can create a significant competitive advantage and drive innovation, increase competitiveness, and create social impact. As EU Commissioner Kroes has stated on several occasions: "**Big Data is the new oil**". Big Data therefore has to be regarded as a **primary asset** for all sectors, organizations, countries and regions.

The following table provides some examples of how Big Data will impact different sectors:

Sectors/Domains	Big Data Value	Source
Public administration	EUR 150 billion to EUR 300 billion in new value (Considering EU 23 larger governments)	OECD[3], 2013
Healthcare & Social Care	EUR 90 billion considering only the reduction of national healthcare expenditure in the EU	McKinsey Global Institute[4], 2011
Utilities	Reduce CO2 emissions by more than 2 gigatonnes, equivalent to EUR 79 billion (Global figure)	OECD[6], 2013
Transport and logistics	USD 500 billion in value worldwide in the form of time and fuel savings, or 380 megatonnes of CO2 emissions saved	OECD[6], 2013
Retail & Trade	60% potential increase in retailers' operating margins possible with Big Data	McKinsey Global Institute[2], 2011
Geospatial	USD 800 billion in revenue to service providers and value to consumer and business end users	McKinsey Global Institute[2], 2011
Applications & Services	USD 51 billion worldwide directly associated to Big Data market (Services and applications)	Various[5,6]

The Big Data Value market measured by the revenue that vendors earn from sales of related hardware, software and ICT services is a fast growing multibillion-euro business. According to IDC[7] the Big Data market

[1] A full list of acronyms and terms is presented in 6.1

[2] "*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.

[3] "*Exploring Data-Driven Innovation as a New Source of Growth – mapping the policy issues raised by "Big Data""*, Report from OECD 18 June 2013.

[4] Applying assumptions from the McKinsey Global Institute report "Big Data: The next frontier for innovation, competition, and productivity", June 2011, to the European healthcare sector

[5] Big Data Market by Types (Hardware; Software; Services; BDaaS - HaaS; Analytics; Visualization as Service); By Software (Hadoop, Big Data Analytics and Databases, System Software (IMDB, IMC): Worldwide Forecasts & Analysis (2013 – 2018), available online at: www.marketsandmarkets.com, August 2013.

[6] "*Big Data Vendor Revenue and Market Forecast 2013-2017*", article, Wikibon, February 2014.

[7] "*Worldwide Big Data Technology and Services 2013–2017 Forecast*", report, IDC, December 2013

is growing six times faster than the overall ICT market. The compound annual growth rate (CAGR) of the Big Data market over the period 2013 – 2017 will be around 27%, reaching an overall total of $50 billion.

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.

A significant contribution to the European economy

As identified by demosEUROPA, "Overall, by 2020, big & open data can improve the European GDP by 1.9%, an equivalent of one full year of economic growth in the EU"[8]. The increased adoption of Big Data will have positive impact on employment, and is expected to result in 3.75 million jobs in the EU by 2017[9].

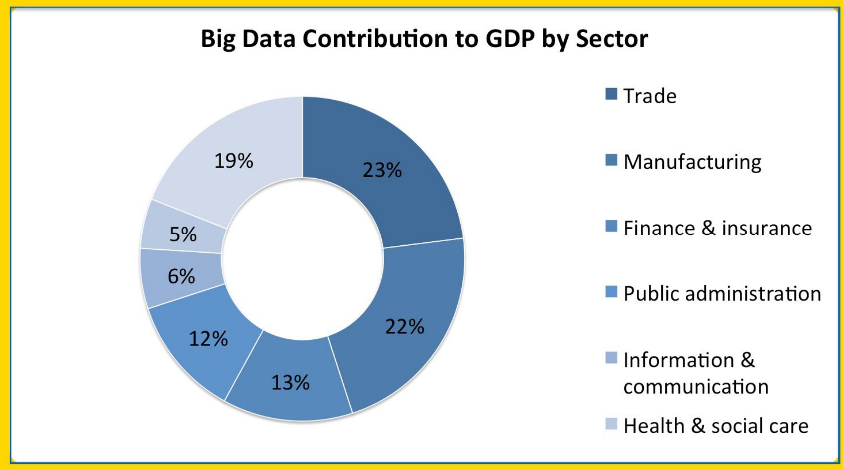


Figure 1: Economic potential of big and open data – source: demosEUROPA

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[10]. According to IDC[11], 30% of Western European companies will adopt Big Data by the end of 2015. For the other 70% of business actors, it is crucial to provide new tools and assets to propel them into the data-driven economy.

However, Europe is still at an early stage of adopting Big Data technologies and services; and is lagging behind North America[12], with the picture in third countries being less well determined. Companies intending to build and to rely on data-driven solutions will face challenges that go well beyond technology. Successful adoption of Big Data will require changes in business orientation and strategy, processes, procedures and the organizational setup. European enterprises will create new knowledge and hire new experts, enhancing a new ecosystem.

[8] "*Big and open data in Europe - A growth engine or a missed opportunity?",* Sonia Buchholtz, Maciej Bukowski, Aleksander Śniegocki (Warsaw Institute for Economic Studies), report commissioned by demosEUROPA, 2014.

[9] Big Data Value calculation based on http://www.eskillslandscape.eu/ict-workforce-in-europe/ (also footnote '7')

[10] "*The European Data Market*", Gabriella Catteneo, IDC, presentation given at the NESSI summit in Brussels on 27 May 2014, available online at: http://www.nessi-europe.eu/?Page=nessi_summit_2014

[11] IDC European Vertical Markets Survey, October 2013

[12] "*The European Data Market*", Gabriella Catteneo, IDC, presentation given at the NESSI summit in Brussels on 27 May 2014, available online at: http://www.nessi-europe.eu/?Page=nessi_summit_2014

The multiple dimensions of Big Data Value

In order to reduce the gap with other countries and regions and drive innovation and competitiveness, Europe needs to foster 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 the creation of 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 generated 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.

- **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 true 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.

- **Social:** 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 on the benefits and the Value that Big Data can create for business, the public sector, and the citizen.

Creating a favourable business environment for Big Data and pushing for its accelerated adoption requires an interdisciplinary approach addressing the dimensions of Big Data Value as described above.

The Data Value Chain and the central role of Ecosystems

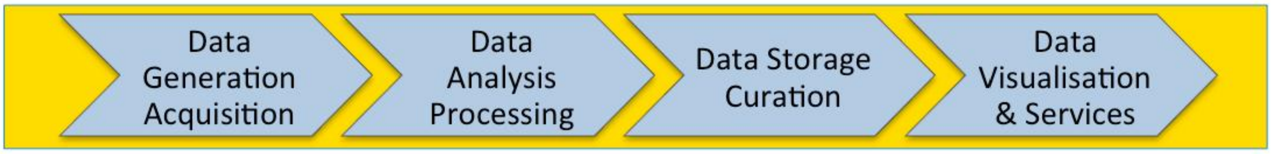


Figure 2: The Big Data Value chain

Europe needs strong players along the Big Data Value Chain[13] (Figure 2) 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.

There are already 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 analyzing data and recognizing correlations and patterns. Furthermore, there are companies that use these insights for predictions and decisions in various application domains.

Despite the growing number of companies active in the data business, an economic community supported by interacting organisations does not yet really exist for the Big Data Value Chain at the European-level. Data usage is growing, but in both businesses and science it is treated and handled in a fragmented way. In order 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 these many stakeholder and many datasets is a key element for advancing the Big Data economy in Europe.

Finally, it is vital that SMEs and web entrepreneurs participate in this ecosystem and become part of the Big Data Value chain. They are an essential part of the process to create value based on their specific and strong niche competences at the technical, application and business level.

Need for action

Big Data is one of the key economic assets of the future. Mastering the creation of Value from Big Data will impact the competitiveness of companies and will result in economic growth and jobs for Europe. Strategic investments by the industry, public sector and governments accompanied by forward-looking policies will enable Europe to take the lead in the global data-driven digital economy and to reap societal benefits from the unique opportunities offered by Big Data Value. The European Council highlighted Big Data in its conclusion of 24/25 October 2013[14] as a strategic technology and important enabler for productivity and better services. However, immediate action is required to ensure Europe does not miss these opportunities. Therefore, **Commissioner Kroes called for a European Public Private Partnership in Big Data** in her speech at the ICT 2013 event in Vilnius on 7 November 2013.

[13] "*Competitive Advantage –Creating and Sustaining Superior Performance",* Michael E. Porter, New York, 1998

[14] European Council Conclusion – 24/25 October 2013 – EUCO 169/13, available online at http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/139197.pdf

Europe must aim high and mobilise stakeholders in society, industry, academia and research to enable a European Big Data 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 environment in which value creation from Big Data flourishes.

In order to achieve these goals, a **European contractual Public Private Partnership (cPPP)** on Big Data Value has been proposed by a partnership of European Big Data stakeholders led by NESSI, the European Technology Platform (ETP) for software, services and data.

This **Strategic Research and Innovation Agenda (SRIA)** defines the overall goals, main technical and non-technical priorities, as well as a research and innovation roadmap for the cPPP. In order to establish a contractual counterpart to the European Commission for the implementation of the cPPP, 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 BDVA will provide regular updates of the SRIA.

A wide range of stakeholders have contributed to this version of the SRIA. It is built upon inputs and analysis from SMEs and Large Enterprises, public organisations, and research and academic institutions. They include suppliers and service providers, data owners, and early adopters of Big Data in many sectors.

The value that the intelligent use of Big Data can generate is already being recognised by some private and public organisations. There are relevant national initiatives in Germany[15], France[16], Ireland[17] and the UK[18]. It is essential that these be connected at the European-level, establishing knowledge sharing, and collaborating to advance the technology.

Discussions and workshops have clearly shown that, alongside vital research and innovation in technologies and applications, many economic, societal and legal challenges will have to be addressed in an interdisciplinary fashion. Underpinning successful exploitation will be the availability of skills and access to investment and capital. Citizens must also be involved, and provision has been made to emancipate them as stakeholders so that they can be the ethical integration point of their own data.

1.1 A Vision for Big Data

The vision for Big Data Value (BDV) in Europe is that in 2020:

- **Data:** Zettabytes of useful public and private data will be widely and openly available.
- **Skills:** Millions of jobs established for data engineers and scientists, and the Big Data discipline is integrated in technical and business degrees. The European workforce is more and more data-savvy seeing data as an asset.
- **Legal:** Privacy & Security can be guaranteed through the lifecycle of BDV exploitation. Data sharing and data privacy will be fully managed by end-users themselves in a trustworthy society – citizen emancipation.
- **Technology:** Real-time integration and interoperability among different multilingual, sensorial, and non-structured datasets and where content is automatically managed and visualised in real-time.
- **Application**: Applications using the BDV technologies can be built which will allow anyone to create, use, exploit and benefit from Big Data.
- **Business:** A true EU single data market will be established allowing EU companies to increase their competitiveness and become world leaders.

[15] http://www.sdil.de/

[16] http://www.teralab-datascience.fr/en/home

[17] http://insight-centre.org/content/launch-insight

[18] http://theodi.org/

- **Social:** Societal challenges are addressed through BDV systems addressing high data volume, high motion of data, 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 higher 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:** Citizen 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.

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.

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.

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[19]. 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.

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.

This foreseen evolution demands rethinking technologies around Big Data. Data collection, storage and processing must be improved in order 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 in order 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.

1.2 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.

[19]"*THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*", IDC report, December 2012

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 1.3.

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[20], Copernicus[21] and GEOSS[22]).
- 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.

[20] http://inspire.ec.europa.eu/

[21]http://www.copernicus.eu/

[22] https://www.earthobservations.org/geoss.shtml

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.
- 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.

1.3 Strategic and Specific Objectives

The BDV cPPP has its roots in the Industrial Leadership Priority in Horizon 2020, specifically the ICT industrial and technological leadership challenge "Content technologies and information management". This aims to strengthen Europe's position as a provider of innovative multilingual products and services based on digital content and data.

Crosscutting initiatives with other strategic areas, particularly societal challenges, will be facilitated in later Work Programmes of Horizon 2020. This will invariably widen the range of objectives of the cPPP and contribute to EU societal challenges through the implementation of BDV applications and solutions.

The following H2020 societal challenges have particular importance for the BDV cPPP due to Europe's current concerns and challenges in these areas (although the cPPP is domain neutral):

- **Health**:	Demographic change and wellbeing.
- **Energy**:	Secure, clean and efficient energy.
- **Transport**:	Smart, green and integrated transport.
- **Environment**:	Climate Action, Resource Efficiency and Raw Materials.

The Big Data Value cPPP is driven by the conviction that research and innovation focusing on a combination of business and usage needs is the best long-term strategy. This will bring many benefits and stimulate the creation of value from Big Data to reach the level that is needed to create jobs and prosperity.

The overarching **strategic objectives** are:

- **Data:** To access, compose and use data in an in a simple clearly defined manner that allows the transformation of data into information.
- **Skills:** To contribute to the conditions for skills development in industry and academia.
- **Legal & Policy:** To contribute to policy processes for finding favourable European regulatory environments, and address concerns of privacy and citizen inclusion.
- **Technology:** To foster European BDV technology leadership for job creation and prosperity by creating a European wide technology and application base and building up competence. In addition, enable research and innovation work, including the support of interoperability and standardisation, for the future basis of BDV creation in Europe.
- **Application:** To reinforce the European industrial leadership and capability to successfully compete on a global-level in the data value solution market by advancing applications transformed into new opportunities for business.
- **Business:** To facilitate the acceleration of business ecosystems and appropriate business models with particular focus on SMEs, enforced by Europe-wide benchmarking of usage, efficiency and benefits.
- **Social:** To provide successful solutions for the major societal challenges that Europe is facing such as: Health, Energy, Transport and the Environment. And to increase awareness about BDV benefits for businesses and the public sector, while engaging citizens as prosumers to accelerate acceptance and take-up.

Specific objectives are:

Competitiveness Objectives (Business):

- To use BDV technology for increased productivity optimised production, more efficient logistics (inbound and outbound), and effective service provision from public and private organisations.
- To create a Big Data Economy including new ecosystems and markets between data providers, knowledge providers and consumers that will profit from sectorial, organizational and individual collaboration.
- To reduce the gap between the traditional economy and new digital business models that will smooth changes in the economic value chain and stakeholders including End Users.
- To implement European-wide strategic projects (Lighthouse projects – See Section 2.2) for specific reference deployments of existing or near-to-market technologies that demonstrate the impact that can be achieved by BDV creation across sectors.

Innovation objectives (Technology):

- To develop and make available to industry and the public sector technology, applications and solutions for the creation of value from Big Data.
- To optimize architectures for real-time analytics of both data-at-rest and data-in-motion that enables data-driven decision-making on the fly with low latency, while improving scalability and processing of data validation and information discovery especially in heterogeneous data sets.
- To drive the integration of the BDV services into private and public decision making systems such as Enterprise Resource Planning and marketing systems for optimising the functioning of existing industries and potentially establishing entirely new business models.

- To validate technologies from a technical and a business perspective within cross-organisational, cross-sector, and cross-lingual innovation environments through early trials.
- To enable European industry, business, public sector and citizens to use and take advantage of value creation from Big Data that is validated with user involvement based on open and private data in secure and privacy respecting environments.
- To integrate advanced visualization of data and analytics for augmented user experience and prepare platforms, technologies and tools for disruptive changes in the management of data.

From the above objectives, the technical domains for the development of the BDV cPPP are detailed in Section 3 of this document: deep analytics to improve data understanding, optimized architectures of both data-at-rest and in motion, privacy and anonymisation mechanisms, advanced visualization, and data management engineering.

Societal objectives (Society):

- To support building extensive know-how, education and skills in Europe (e.g. by European curricula and sharing of best practices) for future systems in the research community and industry.
- To enable European industry, business, public sector and citizens to use and take advantage of value creation from Big Data; that is validated with user involvement based on open and private data in secure and privacy respecting environments.
- To develop and provide validated technology and tools for "deep data analysis" to improve data understanding, deep learning and increased meaningfulness of data for optimal information content.
- To create new personalized and enhanced product and services adapted to citizens and organizations need that will respect the security and privacy of individuals.
- To address European framework aspirations such as IPR, liability etc. within the Digital Single Market and the pan European innovation environments.
- To support the societal challenges that Europe faces through Lighthouse projects and i-Spaces in these challenge areas.

Operational Objectives:

- To create an environment for productive research and innovation activities by utilising proven approaches including clustering actors around key research and innovation areas.
- To establish a governance model that provides for efficient decision making at the level of concerned activity and at the same time following the target of openness and transparency.
- To maintain an effective flow of information between and amongst the cPPP projects to overcome barriers while respecting interests and rights of individual beneficiaries.
- To enable collaboration amongst the projects that support the common targets to positively impact European society and industry.

In order to achieve the strategic and specific objectives, the research and innovation strategy requires dedicated actions and mechanisms along its overarching strategic goals. In short, some of the main **technical** aspects are:

- **Data:** Data is placed at the centre of the Big Data Value activities. These will typically be based on their domain of operation and include industrial, private and open data sources whilst ensuring their availability, accessible, integrity, and confidentiality.
- **Technology:** Fostering research and innovation activities to develop Big Data Value technologies and tools. Focus will be put on those technologies and tools that are needed to support data-driven applications and business opportunities along the data value chain. They will be addressed in the technical projects (see Section 3).
- **Application:** Benchmarking and incubation of Big Data 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 these activities will provide inputs for standardization and regulation.

Complementary, further **socio-economic** aspects need to be addressed:

- **Skills:** Developing skills and federating best practices through linking with other similar initiatives at European and national-levels.
- **Business**: Exploring and stimulating new business models and ecosystems that will emerge from the exposure of new technologies and tools to both closed (industrial) and open data.
- **Legal, Policy and Privacy:** Providing insight into country, regional and European-wide legislation, regulation, and similar issues which impact the implementation or use of data-intensive technologies.
- **Social**: Acquiring 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.

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 of 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.

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.

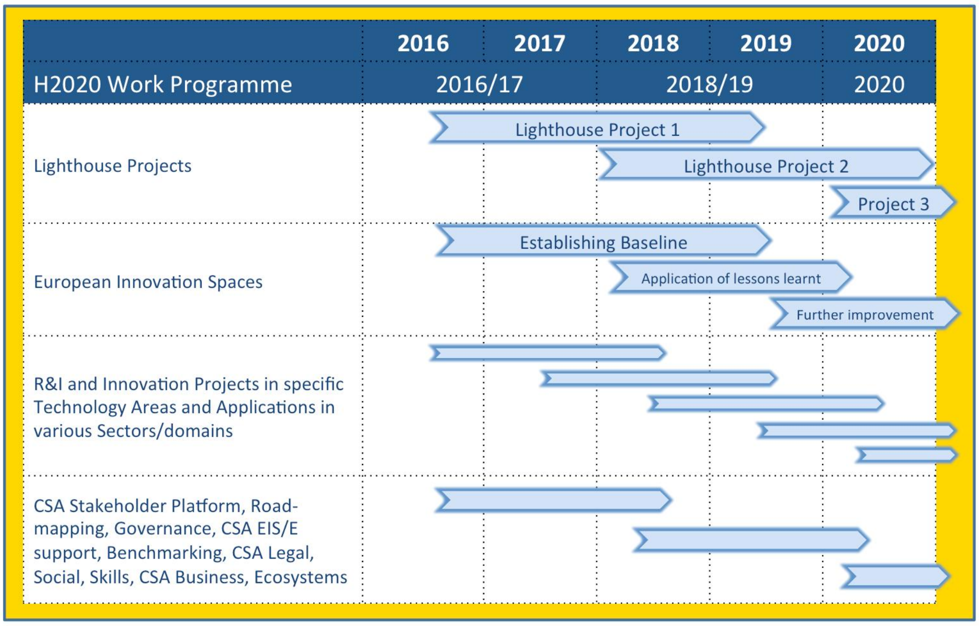

2016
2017
2018
2019
2020
H2020 Work Programme
2016/17
2018/19
2020
Lighthouse Projects
Lighthouse Project 1
Lighthouse Project 2
Project 3
European Innovation Spaces
Establishing Baseline
Application of lessons learnt
Further improvement
R&I and Innovation Projects in specific Technology Areas and Applications in various Sectors/domains
CSA Stakeholder Platform, Road-mapping, Governance, CSA EIS/E support, Benchmarking, CSA Legal, Social, Skills, CSA Business, Ecosystems

Figure 3: Timeline of activities

The lighthouse projects will start in the most mature sectors to play an early role as large-scale demonstrators for accelerating the take-up of BDV creation in industry and the public sector.

SRIA Preparation Process with Broader Community

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, 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[23], LatAm countries[24], South East Asian countries[25] and the Russian speaking countries[26].

SRIA Update Process

The technical and non-technical priorities reflected in this version of the SRIA reflect the consolidated results from the needs analysis performed by involving all relevant stakeholders of the Big Data environment. We are well aware that in a fast moving area such as Big Data, those priorities need to be regularly reflected and updated if needed. The Big Data Value Association (BDVA) will provide regular updates of this SRIA document defining and monitoring the priorities as well as metrics of the cPPP.

BDV Stakeholder Platform

The BDV Stakeholder Platform constitutes a permanent platform for BDV stakeholders to express their views, expectations and requirements. They will be captured in a continuous manner, thereby establishing a series of long-term SRIAs. BDV stakeholders that cannot commit to participate on a regular manner, but wish to participate and contribute, will be also facilitated by the platform.

The BDV Stakeholder Platform will take advantage of already established stakeholder groups and communities, such as those started in BIG and BYTE, and will take them into account wherever appropriate. Once set-up it will have the capacity to gather and coordinate BDV stakeholder recommendations along technology, application, skills, ecosystem and social dimensions. 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 ETP4 HPC, 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. The activity of the platform is at a more detailed, and typically domain-specific, level and will form key inputs to the BDVA in the development of its roadmaps and SRIAs.

Regular revision of the BDV Stakeholder Platform will create trust and ensure independency. Advisory sessions and hearings with international experts should be part of the activities. Steering committees for organizational and technology issues will be set-up to enable the projects to work as efficiently as possible and identify interdependencies and complementarities. The federation of Big Data initiatives from both the private and public sector in European Members States (and at regional-level) will make collaboration with other organizations vital. Examples are the Smart Data Innovation Lab in Germany or the Teralab in France.

[23] MOSAIC (http://www.connect2sea.eu/) and MED-Dialogue (www.med-dialogue.eu)

[24] CONECTA 2020 (www.conecta2020.net)

[25] CONNECT2SEA (www.connect2sea.eu)

[26] EAST HORIZON (www.eeca-ict.eu/about/new_projects/easthorizon)

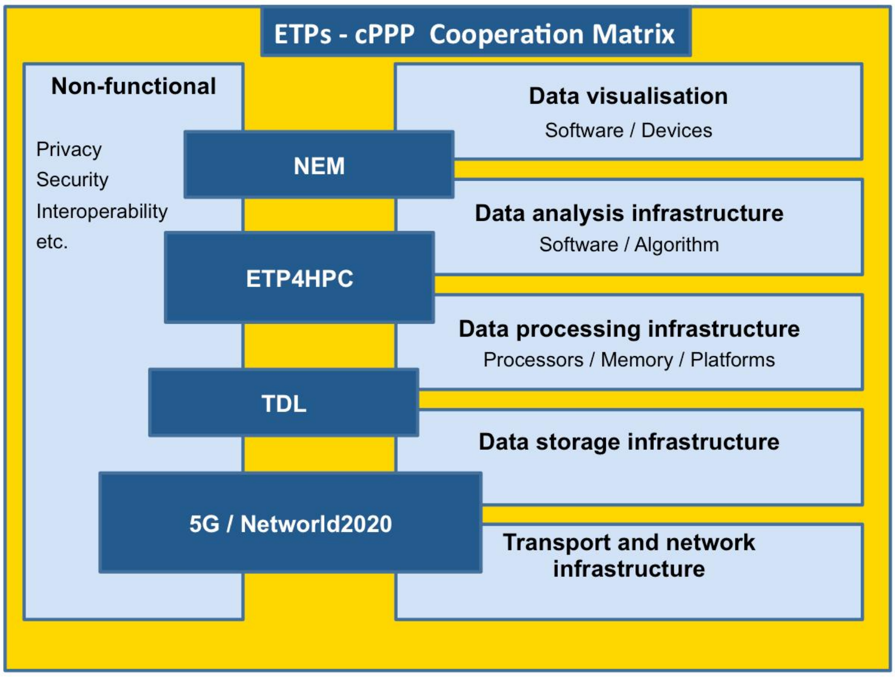

ETPs - cPPP Cooperation Matrix
Non-functional
Privacy
Security
Interoperability
etc.
NEM
ETP4HPC
TDL
5G / Networld2020
Data visualisation
Software / Devices
Data analysis infrastructure
Software / Algorithm
Data processing infrastructure
Processors / Memory / Platforms
Data storage infrastructure
Transport and network infrastructure

Figure 4: Examples of collaboration with other ETPs and PPPs [27] [28] [29] [30] [31]

2.1 European Innovation Spaces (i-Spaces)

Extensive consultation with many stakeholders of relevant areas related to Big Data Value (BDV) have 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 for 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.

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 an 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

[27] NEM: New European Media, http://nem-initiative.org/

[28] ETP4HPC: The European Technology Platform for High Performance Computing, http://www.etp4hpc.eu/

[29] TDL: Trust in the Digital World, http://www.trustindigitallife.eu/

[30] 5G: 5G PPP, http://5g-ppp.eu/

[31] NetWorld2020: The European Technology Platform for communications networks and services, http://networld2020.eu/

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.

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 in turn 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 of 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 interdisciplinary holistic approach, i-Spaces will be a key mechanism that targets BDV challenges along the dimensions as depicted in Figure 5. 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.

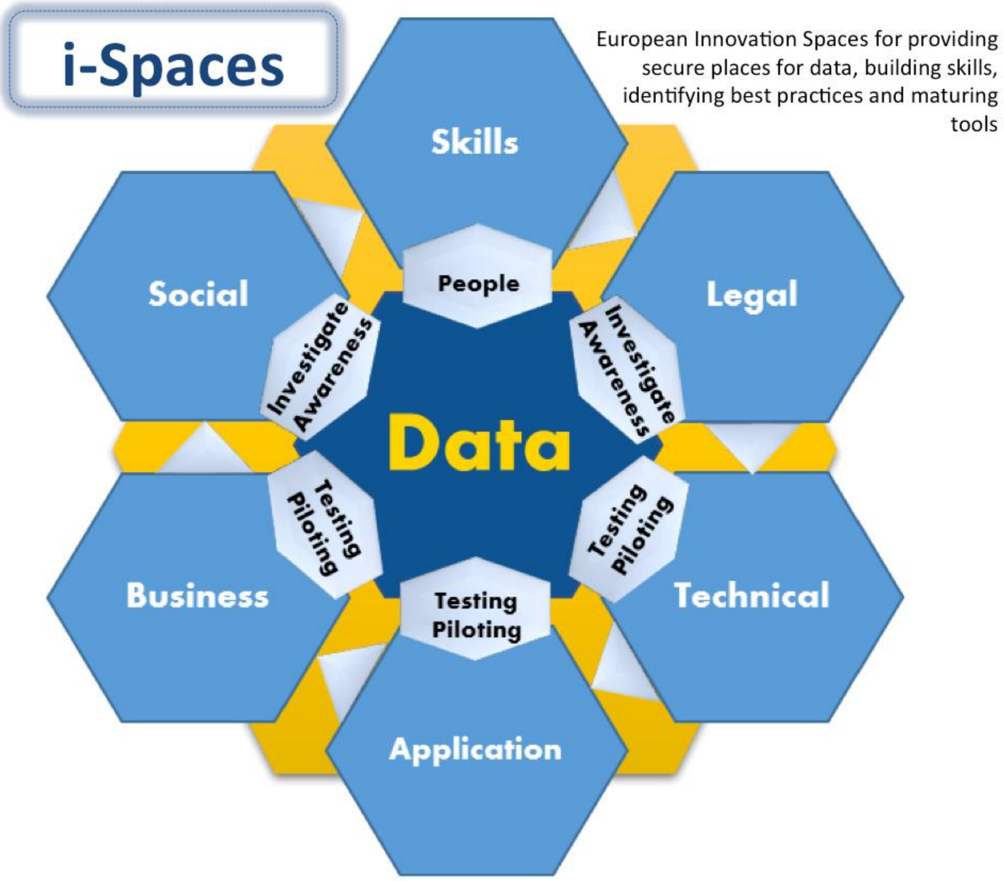


Figure 5: Interconnected Challenges of the cPPP within i-Spaces

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:

- **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.
- **Privacy:** Data protection and privacy 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.
- **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.

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.

2.1.1 Setup of i-Spaces

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

- A well-managed **IT infrastructure**, including remote access capabilities.
- **Secure and trustworthy data hosting and access.**
- Availability of a team providing basic IT **assistance**, as well as focused **support** by Big Data experts.

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

- **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.
- **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-Space 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.

2.1.2 Innovation spaces a tool for continuous benchmarking

An important provision by the i-Spaces is to facilitate benchmarking so that businesses, and in particular start-ups and SMEs, are able to evaluate whether their products and services will work in a real-world context. Hence, benchmarking is about comparing a specific product or service, for instance, with a peer product or service, respectively. Such comparison covers measures such as efficiency, effectiveness, cost, quality and return on investment.

Given the specific characteristics of BDV, i-Spaces will focus on four strands of benchmarking:

- **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.

For all the four strands mentioned above, i-Spaces will facilitate benchmarking processes. Since benchmarking is not a one size fits all and is highly context sensitive, the organisation planning a benchmark will have a leading role in the process. The following elements therefore will offered by i-Spaces to support such benchmarking:

- Services for assisting in identifying what is to be benchmarked, and how to benchmark it.
- Services for assisting in identifying businesses with, for instance, a peer product or a peer service useful for benchmark measures.
- Services for assisting in identifying best practices and measures.
- Services for assisting in running the benchmark.

2.2 Lighthouse projects

Lighthouse projects run data-driven large-scale demonstrations whose main objectives will be to create high-level impact, and broadcast visibility and awareness to drive faster uptake of Big Data Value applications and solutions. Lighthouse projects will propose replicable solutions by using existing technologies or very near to market technologies that could be integrated in an innovative way to demonstrate big data value.

Lighthouse projects will be the major, high-impact mechanism for Europe 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. The projects should prepare the path for other domains to replicate the experiences, leading to explicit business growth and job creation. To this end 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.

2.3 Technical projects

Technical projects[32] focus on one or few specific aspects of technical priorities, thereby providing the technical foundation for lighthouse projects and i-Spaces.

2.4 Cooperation and coordination projects

Cooperation and coordination projects[33] will work on detailed activities that ensure coordination and coherence along the cPPP implementation and will provide support to activities that fall under the skill, business, policy, regulatory, legal and social domains.

[32] for instance Research and Innovation actions (RIAs) and Innovation actions (IAs)

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

3 Technical Priorities

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 to 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

3.1.1 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 very 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), 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 real added value of Big Data in terms of innovations lies in applications driven by the data analysis.

3.1.2 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 6). For each of these roles the following main technical needs are identified in the following Table:

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	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	Need for 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 services	Applies decisions and insights derived from analysis to optimization of own business	Privacy and anonymisation, Advanced visualization

Table 1: Roles and activities of ecosystem actors

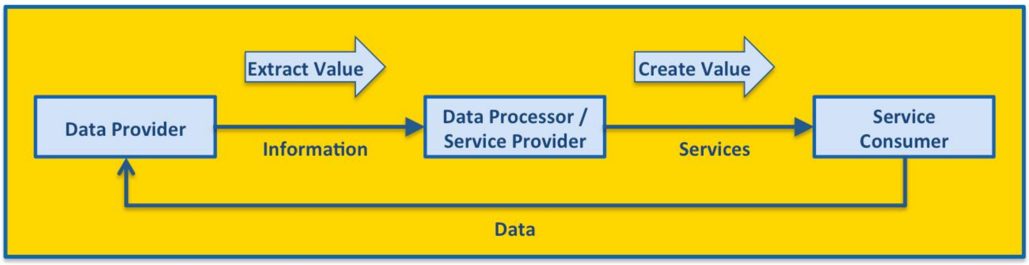


Figure 6: Various technical needs and concerns according to the role in the ecosystems

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, deep 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.
- Deep 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 the “**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, and multimedia. 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), 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:** 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.
- **Data provenance:** Data, information and content needs to be syndicated from data providers to data consumers whilst maintaining provenance, control and source information including IPR considerations.
- **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.

Outcome

- 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), as well as, standardized annotation frameworks for different sectors supporting the technical integration of different annotation technologies and data formats.
- 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.
- 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.

- Methods and tools for a complete **data management lifecycle** ranging from data curation and cleaning (including pre-processing veracity, integrity, and quality of the data), 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.
- Languages and tools for **data provenance,** control and IPR.
- Methods and Tools for the sound **integration of analytics results** from **data and business** processes.
- Principles for a clear **data-as-a-service 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.

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 started by applying the existing architectures and technologies of Big Data, which we categorize as data-at-rest. In the mean time the need for processing immense amounts of sensor data streams has increased. This type of data-in-motion 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.

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 architectural solutions will leverage the true potential. Optimized 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. 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 Kafka).

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:

- **Integrated Processing of data-in-motion and data-at-rest:** Real-time Analytics and Stream Processing: which span across the areas of inductive reasoning (machine learning), deductive reasoning (inference), high performance computing (data centre optimization, efficient resource allocation, quality of service provisioning) and statistical analysis, adapted to allow continuous querying over streams (i.e., on-line processing). New Big Data-specific parallelization techniques and (at least partially) automated distribution of tasks over clusters are crucial elements for effective stream processing. Most of these

processing techniques have only been applied to data-at-rest and in some cases to data-in-motion. The challenge here is to have an integrated processing for both, at the same time.
- **Scalable Analytics for data-in-motion and data-at-rest:** Being able to apply complex analytics techniques at scale and for data-in-motion and data-at-rest 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 increasing low latency demands of streams and the volume demands of Big Data sets.
- **Performance and Scalability:** 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 utilize existing and emerging hardware-oriented developments like main memory technology with different type of caches, like software-defined storage with built-in functionality for computation near the data (e.g. Storlets) and like data reduction to support storing, sharing, and efficient in-place processing of the data.

Outcome

- **Architectures for data-in-motion and data-at-rest:** Architectures, frameworks and tools for the integration of mostly existing components to new types of platforms, which address the orthogonal challenges in completely new ways, by widening and generalizing known data processing capabilities for data-at-rest and data-in-motion. Furthermore, there is a need to dynamically reconfigure such architectures and data processing capabilities on the fly for example to cope with context changes, changing requirements and optimization in various dimensions (e.g., performance, energy consumption and security).
- **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.
- **Scalable algorithms and techniques for real-time analytics:** that is capable of analysing large amounts of data-in-motion and data-at-rest by updating the analysis results as the information content changes. Algorithms and techniques are needed for the demanding low latency requirements.

3.4 Priority “Deep Analytics”

Background

The progress of deep analytics of Big Data is not only key to turn Big Data into value, but also to make it accessible to the wider public. Deep 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.

Deep 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[34] 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[35].

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. Deep 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-

[34] “*Big Data: The next frontier for innovation, competition and productivity*”, McKinsey Global Institute, June 2011

[35] “*Game changers: Five opportunities for US growth and renewal*”, McKinsey Global Institute, 2013

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 Deep Data Analytics:

- **Semantic Analysis:** Improvement to the analysis of data to provide a near-real-time interpretation of the data (i.e. sentiment, semantics, etc.) enable the sharing of data semantics without dependency.
- **Content Validation:** Implementation of veracity (source reliability / information credibility) models for validating content and exploiting content recommendations from unknown users.
- **Analytics Frameworks:** New analytics frameworks and open APIs for the quality-aware distribution of stream processing analytics with minimal development effort from application developers and domain experts.
- **Processing:** 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**: Deep learning techniques and graph mining techniques applied on extremely large graphs. 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 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 and Pseudonymisation Mechanisms”

Background

The security and data protection issue, important in the development of any information system, becomes crucial in the context of the development of cloud computing and massive data processing, along with an increase in sensitive data processing.

Privacy and data anonymisation is one of the major issues in the area of Big Data and data analytics involving all stakeholders in the value chain and is even reflected in legal privacy regulations such as the EU Data Protection Directive. Data privacy and security is indeed often perceived as a hurdle that would prevent data owners from joining Big Data innovation environments. A similar statement can be made for citizens who are more seriously taking into account privacy guarantees.

Recent studies have demonstrated that naïve anonymisation (sometimes referred as de-identification) strategies can be easily circumvented by attackers to re-identify individuals, simply combining the anonymised data with other publicly available data. Due to the critical importance of privacy issues in lots of business domains, optimizing data anonymisation has emerged as an independent and very active research domain. Very promising techniques such as differential privacy, private information retrieval, k-anonymisation, homomorphic encryption, secure search encryption, secure indexes and multi-party computation, have been developed but are, unfortunately not yet suitable for commercial, large-scale processing tasks.

Privacy-enhancing technologies can be an asset for Europe as this is currently an underdeveloped market. Indeed, there is a risk of a ‘race to the bottom’ of privacy protection, where failure to comply with data protection rules and the acquisition of data through anti-competitive means may have become symptomatic of market power, with externality costs borne by users. Users may not yet perceive the extent to which their personal data is marketed but this is starting to change. Hence firms operating in the digital economy may realize that privacy is an opportunity for competitive advantage.

Challenges

In this perspective the following main challenges have been identified:

- A more **generic and easy to use data protection approach** suitable for commercial large-scale processing is needed. For instance, the framework CASD[36], which was already successfully implemented at the French Innovation Space TERALAB, can be used as a starting point and extended in accordance to the requirements of BDV projects. 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 control access, storage and usage of their data 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.
- **Anonymity** is an important challenge, which also implies sub-challenges, such as the need for data analytics to cope with anonymised datasets. 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 irreversibility of the anonymisation of Big Data assets is a key Big Data issue. 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

[36] *A French acronym which stands for Secure Remote Data Access Centre*

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.
- **Risk based approaches** calibrating controllers' obligations regarding privacy and personal data protection must be considered especially when dealing with 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 by, for instance, generalizing the CASD framework. This includes protecting the cloud infrastructure, analytics applications, and the data from leakage and threats, but also provides easy to use privacy mechanisms.
- **Data minimisation**: Methods for secure deletion of data and personal data minimization.
- **Mining Algorithms:** Developed privacy-preserving data mining algorithms.
- **Robust anonymisation algorithms:** Scalable algorithms that guarantee anonymity even when other public data is integrated. In addition, algorithms allow the generation of reliable insights by crossing data from a particular user in multiple databases while protecting the identity of the user.
- **Protection against reversibility:** Risk assessment tools to evaluate the reversibility of the anonymisation mechanisms.
- **Pattern Hiding:** Design of mechanisms for pattern hiding so data is transformed in a way that certain patterns cannot be derived (via mining), while others can.
- **Multiparty Mining:** 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.

3.6 Priority "Advanced Visualisation and User Experience"

Background

Data visualisation is vital if people are to effectively consume Big Data. 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.

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[37]. 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 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 capabilities for the exploration of unknown and unpredictable data within the visualisation layer. The following list briefly details the research areas identified for Deep Data Analytics:

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

- **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.
- **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.
- **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 Big Data can deliver.
- **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.
- **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).
- **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.
- **3D Visualization Techniques and Tools:** Real-time and collaborative 3-D visualization techniques and tools.
- **Personalized End-User Centric Data Visualization Mechanisms:** Adaptation to the needs of end users (user adaptation and personalization but 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 time[38]; 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).
- **Techniques and Tools for Visualization of Interrelated/Linked Data:** Rather than data islands, visual interfaces must take account of semantic relationships, relying on interactive graph visualization techniques to allow easy exploration of network structures.
- **Plug-and-Play Reusable Components for Data Visualization:** User adaptable interactive visualization components that support the combination of any visualization asset in a real-time plug-and play manner – for instance: maps, graph visualization, dashboards.

[38] "*Space, Time, and Visual Analytics* ", G.Andrienko, N.Andrienko, U.Demšar, D.Dransch, J.Dykes, S.Fabrikant, M.Jern, M.-J.Kraak, H.Schumann, C.Tominski , International Journal Geographical Information Science, 2010, v.24 (10), pp. 1577-1600

3.7 Roadmap and Timeframe

In order to achieve the overall, strategic technical goal laid out in Section 1.3 and to address the aforementioned technical priorities, a roadmap defining expected outcomes will be developed.

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 cPPP[39], 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

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.

At this early stage of the Big Data discipline, we see two sub-disciplines emerging that require two distinct breeds of skills and expertise: **Data Scientists** and **Data Engineers** (as will be elaborated below). In fact, this is very similar to what happened to the software discipline in the years since the seminal NATO conference on Software Engineering in 1968. In software engineering there are now two principle, complementary types of specialists: (1) computer scientists, who are concerned with theoretical foundations and basic technology for creating software; (2) software engineers, who are concerned with establishing principles, tools, methods and sound engineering principles to efficiently and effectively develop, maintain and evolve software.

Drawing on this similarity with the software field, we currently, see a first trend of two important Big Data Value related skill sets[40]:

Data Scientist*:* 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. 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. Hence, innovation spaces could be used to a certain extent to build such networks.

Data Engineers*:* These are the specialists that develop and exploit techniques, processes, tools and methods for developing applications that actually turn data into value. In addition to technical expertise, Data Engineers 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

[39] Which are described in detail in the Big Data Value cPPP proposal.

[40] Please note that, as always in novel fields, there are many different, even contrary definitions out there; e.g., some further consider a data analysts being a specific additional type of specialist (in our case we subsumed the competencies in our definition of data engineer); some flip the definitions of data engineer and scientists altogether.

and business. In a sense they thereby constitute the link between technology experts and the business analysts. Data Engineers will foster the development of Big Data applications from an "art" into a disciplined engineering approach. Data Engineers 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 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 in data science and data engineering based on an 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 skillsets needed to be Data Engineers and Data Scientists.
- 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

The Big Data Value ecosystem (See Figure 7) will comprise many new stakeholders. New concepts of data collecting, processing, storing, analysing, handling, visualisation and most importantly the usage will be found and business models created around it. Identifying valid and sustainable business models and ecosystems around sectors or platforms will be challenging. In particular the many SMEs involved in specific, if not niche, roles will need support to align and adapt to the new value chain opportunities.

Dedicated projects for investigating and evaluating business models will be connected to the innovation spaces where suppliers and users will meet. Those projects will:

- Establish a mapping of technology providers and their value contribution.
- Identify mechanisms by which data value is determined and value is established.
- Provide a platform for entrepreneurs and financial actors including venture capitalists to identify appropriate levels of value chain understanding.
- Describe and validate business models that can be successful and sustainable in the future data-driven economy.

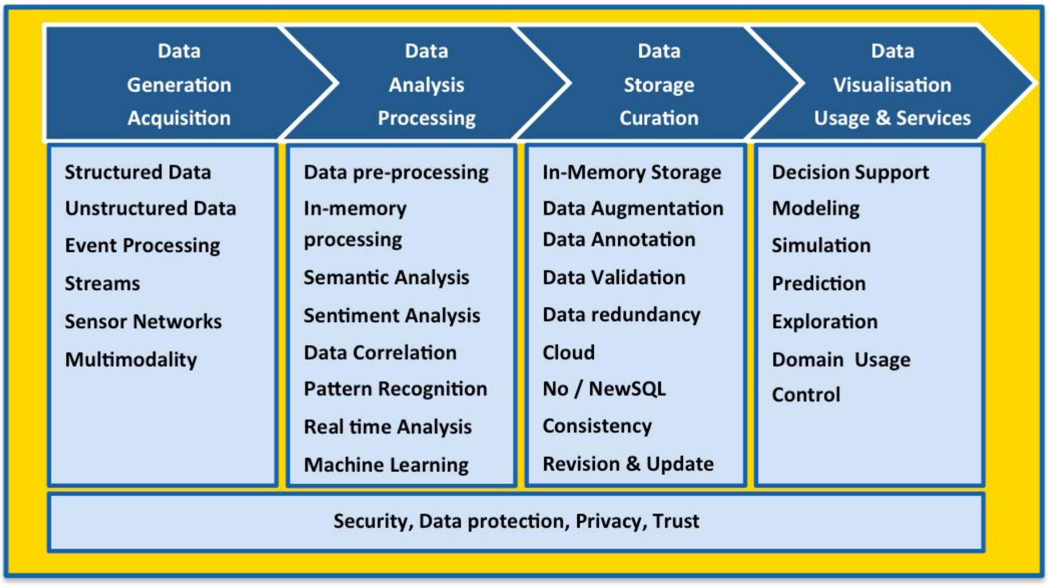


Figure 7: Big Data Ecosystem along the Value Chain

The outcomes of these projects will foster the creation of a more stable business environment that enables business, particularly web entrepreneurs and SMEs, to enter Big Data markets and ecosystems.

Europe needs to establish strong players in order to make the whole Big Data Value ecosystem, and consequently Europe's economy, strong, vibrant and valuable. The following **key stakeholders** are seen as actors along the Big Data Value chain:

- **User Enterprises**, e.g. enterprises in all sectors and of all sizes 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 that are 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.

Furthermore, the known stakeholders in H2020 along the phases of research, innovation, exploitation, and usage will play one or more roles of the Big Data Value chain:

- **Industry-Large** as providers of technologies and services who can also become users.
- **Industry-SMEs** to provide particular know-how and innovative solutions for specific concepts.
- **Universities** that research new algorithms and technologies to be applied in tools & platforms.
- **Research centres** which investigate new algorithms, methodologies and which define new business cases.

4.3 Policy, Regulation and Standardisation

4.3.1 Input to policy making and legal support

The cPPP has no mandate or competence to be involved directly in policy making for legal or regulatory framework conditions. However, the cPPP 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 new data ownership 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.
- Create a catalogue of legal practice in European Member States and other OECD countries and critical issues for BDV actors.
- Establish a Big Data helpdesk and prepare for legal clinics to give advice on legal issues.

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.

4.3.2 Standardisation

Standardisation is essential to the creation of a Data Economy and the cPPP will support establishing and augmenting both formal and de facto standards. The cPPP 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 cPPP 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 is 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 cPPP 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 how trust can be built up, and privacy concerned built into the solutions and services. Societal challenges will be addressed in dedicated projects:

- Investigate the lack of trust in Big Data Value technology, applications and solutions.
- Establish a competence on the impact of trust that can be made by technology changes.
- Address privacy-by-design principles and create a common understanding amongst the technical community.
- Identify key privacy concerns and develop answers based on new solutions.
- Work towards a clearer profile of the social benefits that Big Data Value technology can provide.

These projects will assure that the citizen's views and perception is taken into account so that technology and applications are not developed without a chance to be widely accepted.

5 Expected Impact

5.1 Expected Impact of strategic objectives

The expected impact of the cPPP 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 cPPP 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 cPPP 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 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 establish 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 own data or annotations to existing data. The environment will move from a few state-orientated broadcasters to a prosumer approach, where data and content is 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 business. 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.

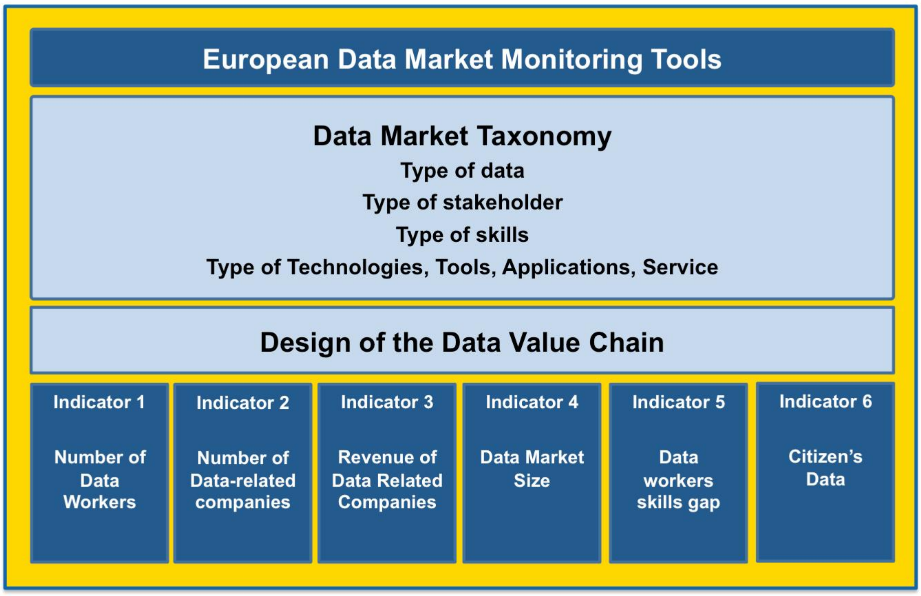
5.2 Monitoring of objectives

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 measured 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 1.3, an interdisciplinary and holistic approach will be followed. Consequently, the indicators to be used for measuring the impact of the SRIA have to address strategic, social, competitiveness, and innovation aspects.



European Data Market Monitoring Tools

Data Market Taxonomy

Type of data

Type of stakeholder

Type of skills

Type of Technologies, Tools, Applications, Service

Design of the Data Value Chain

Indicator 1 Number of Data Workers

Indicator 2 Number of Data-related companies

Indicator 3 Revenue of Data Related Companies

Indicator 4 Data Market Size

Indicator 5 Data workers skills gap

Indicator 6 Citizen's Data

Figure 8: 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[41] has proposed some preliminary European Data Market Indicators that are shown in Figure 8. The SRIA proposes to use this kind of market metrics as indirect indicators for monitoring its strategic impact. This monitoring will need to be adapted in alignment with current on-going work such as that in IDC on establishing metrics for the BDV market.

Strategic Indicators		Societal	Competitiveness	Innovation	Operational
KPI.S.1	Number of Data Workers in different sectors, domains and sub-professions	■			
KPI.S.2	Economic impact on productivity within the EU and comparison to other geographical areas	■	■		
KPI.S.3	Development of the positioning of European companies in the ranking of leading global BDV companies		■		
KPI.S.4	Number of SMEs and web entrepreneurs dealing with Big Data services and products		■	■	
KPI.S.5	Number of SMEs and web entrepreneurs in the cPPP and BDVA				■

[41] "*The European Data Market*", Gabriella Catteneo, IDC, presentation given at the NESSI summit in Brussels on 27 May 2014, available online at: www.nessi-europe.eu /?Page=nessi_summit_2014

KPI.S.6	Year-on-Year increase of number and % of SMEs and web entrepreneurs in the BDV Implementation at R&I as well as user-level		■	■	■
KPI.S.7	Big Data market revenues in Europe and globally including market share of EU industry		■		
KPI.S.8	Deployment of Big Data technology in industry, public sector and its use by citizen			■	
KPI.S.9	Skills gap of workers, graduates and scientists including comparisons with other geographies	■			
KPI.S.10	Inclusion of Citizens in BDV ecosystem as well as their contribution of data (where approved)	■			

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	**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	**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	**At least 50 training programs are established with participation of at least 100 participants per training session arising from the cPPP.** Continuous development of skills and competences on the basis of the Big Data Value cPPP 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 cPPP activities.	■	■		
	KPI.D.4	**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	**At least 10 major sectors and major domains are supported by Big Data technologies and applications developed in the cPPP**. The usage of BDV technologies and applications developed in the cPPP 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	**Total amount of data made available to i-Spaces - including closed data – is in the Zettabyte 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 performing deep analytics to improve data understanding, deep learning and the meaningfulness of data. Combined with advanced visualization techniques it will guide a unique user experience. In order to assure privacy protection, and to respect a user's privacy, relevant measures and anonymisation mechanisms need to be applied.			🟦	🟦
	KPI.D.7	**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 deep analytics. During the cPPP relevant metrics will be derived.		🟦	🟦	🟦
Technical	KPI.D.8	**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	**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	**Enabling advanced privacy and security respecting mechanisms (including anonymisation) for data access, process and analysis.** 10% year-to-year increase of closed data sets available in i-Spaces.	🟦	🟦	🟦	

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
cPPP	(Contractual) 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

The following individuals and organisations are thanked for their direct involvement it creating this specific proposal 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.		