Update of the European Data Market SMART 2016/0063

Story 9 – Scaling up data-driven innovation: European industry requirements and the role of European data spaces

11th June 2020

Author(s)	Gabriella Cattaneo (IDC) , Chiara Francalanci (Politecnico of Milano)
Deliverable	D 3.8 Quarterly Stories (Story 9)
Date of delivery	11th June 2020
Version	1.0
Addressee officer	Katalin IMREI Policy Officer European Commission - DG CONNECT Unit G1 – Data Policy & Innovation EUFO 1/178, L-2557 Luxembourg/Gasperich katalin.imrei@ec.europa.eu
Contract ref.	N. 30-CE-0835309/00-96

Table of contents

Table of Figures

Executive summary

Common European Data Spaces are one of the fundamental pillars of the new European Strategy for Data (COM (2020) 66 final, 02/19/20), whose aim to make Europe a global leader in the data-agile economy, To become operational, and exert a systemic impact in Europe's strategic economic sectors, data spaces will need to develop data governance mechanisms while accessing high value datasets to enable data-driven innovation within vertical ecosystems and foster their developments.

This research looks precisely at common data spaces from the viewpoint of those European industries that have already invested in data-driven innovation, have achieved measurable business benefits and are engaged in scaling-up these efforts. This provides an evidence-based and industry-specific view about the pragmatic requirements of Common European data spaces, with a focus on the requirements for data governance, access to data, access to infrastructures. The results show that the path towards data spaces effectively supporting data sharing at ecosystem level will not be easy. Enterprises are still concentrating most of their efforts and investments on developing data-driven innovation within their organization, at best sharing some data with a few trusted sub-suppliers. The most relevant barrier to scalability comes from the cost of cloud infrastructures and the dependency on a few global suppliers resulting in potential customer lock-in effects.

The story features a panoply of case studies developed by Politecnico of Milano (POLIMI) and IDC across seven different industries. The paper highlights how remarkable levels of business benefits are already being achieved through data-driven innovation in Europe. This is the case of cost reduction (with 80% reduction of operational expenditures for fraud detection in the financial services industry, for example or 30% reduction of maintenance costs in manufacturing thanks to predictive manufacturing) and customer benefits (for example, a 110% improvement of customer retention in manufacturing and 85% improvement of conversion rates from potential to actual customers thanks to data-driven targeting in retail).

The case studies included in this research revolve around services and applications, although most of them are still confined to individual departments or branches of the company. When scaling up these services, the following issues would emerge:

- Even when the technology solution has been well selected and designed, there is no guarantee that scalability to the whole organization is feasible and cost-effective until it is actually implemented;
- Business intelligence remains central to ensure the strategic use of data, and this is still to be found in human resources rather than in machines (the so-called "business superficiality" of data-driven technologies). For example, automated recommendation systems built on standard solutions will tend to recommend the products with highest sales, which are likely to be the products with lower prices and lower margins. A business manager must intervene to provide an intelligent strategy, for example finding ways to nudge customers towards products with higher prices and margins. Integrating business intelligence with the use of data analytics is still immature in many industries.
- Most of these solutions rely on public cloud technologies. This creates a lock-in risk, since migrating to other cloud providers requires considerable costs and time investments for the redesign of software. The high concentration of the cloud providers market reduces the potential choices of business users and constraints the scaling up of data-driven innovation.
- When scaling up, particularly if the solution requires real-time data processing, the cloud computing costs tend to rise very quickly and cross a threshold where technology costs are higher than business benefits (edge vs. cloud decisions). This is a sensitive aspect, particularly for solutions combining BDA (Big Data Analytics) and Artificial Intelligence such as machine learning.
- The cloud computing costs are still relevant. Basic cloud services are quite convenient, but prices increase very fast as soon as more sophisticated services are needed. For example, in the retail industry, a leading supermarket chain found that leveraging AI for sales prediction in one shop led to a 5% increase of margins (equivalent to roughly 5 million €/year); but applying the same machine-learning application to all shops in the chain would wipe out the benefits and cost more than the margin increase. These issues

are behind the decision of the European Commission to promote federated European cloud infrastructures. The European data strategy in fact plans to fund a High Impact Project on European data spaces and federated cloud infrastructures.

- Access to data processing and computing capabilities needs to be improved. These case studies confirm the need for enhanced access to data processing and computing capacities, as foreseen in the Data Strategy in terms of support for data spaces.

In conclusion, there has been considerable progress in the use of data-driven innovation by European industries over the past few years, including an increasing use of AI techniques such as machine learning leveraging the power of data. Nevertheless, there is still a high level of immaturity in the capability to merge datasets within a company, and relevant barriers against data sharing even in advanced sectors such as manufacturing. A relevant issue which emerged from most case studies is the availability of affordable and efficient cloud computing infrastructures, allowing the scaling up of successful pilots and individual company-site experiences to the whole company domain. Even if potentially Common data spaces could provide a valuable answer to the need for greater access to high quality datasets and computing infrastructures, this will require solving practical and technology challenges, not simply providing a favourable environment for encouraging stakeholder collaboration.

1. Introduction

The strategic objective of the new European Strategy for Data (COM (2020) 66 final, 02/19/20) is to make Europe a global leader in the data-agile economy, by creating a favorable policy environment and a genuine single data market for data. A pillar of this strategy will be the creation of common European data spaces in strategic economic sectors and domains of public interest, where data driven innovation will have system impact on the entire ecosystem and on citizens. To become operational, data spaces will need to develop data governance mechanisms and access to high value datasets to enable data-driven innovation within vertical ecosystems and foster their developments.

The European Commission has investigated the potential needs and requirements for common data spaces in a series of workshops with stakeholders[1] as well as several other initiatives. This story looks at common data spaces from the viewpoint of European industries who have already invested in data-driven innovation, have achieved measurable business benefits and are engaged in scaling-up these efforts.

1.1 Main objectives and scope

This story leverages a series of case studies of innovative big data applications by European industries to investigate the potential lessons learned for the establishment of European common data spaces. The story aims at:

- Providing an evidence-based and industry-specific view about the pragmatic requirements of Common European data spaces, with a focus on the requirements for data governance, access to data, access to infrastructures;
- Investigating the insights from industrial stakeholders engaged in scaling up data-driven innovation about drivers and barriers of data sharing;
- Comparing such insights with the results of the stakeholders' discussion collected by the EC in the workshops about data-driven innovation in several industries;
- Draw final considerations about the potential requirements of common European data spaces in three main domains: manufacturing, finance and agriculture.

1.2 Methodology and structure

The main research question is: from the viewpoint of industrial stakeholders already engaged in data-driven innovation, what are the potential advantages and disadvantages of participating to common data spaces in their industry? What are the pragmatic requirements and incentives of common data spaces able to attract data owners to confer their high-quality datasets? Because we are all very aware that successful innovative companies do not necessarily need data sharing and European platforms, which must be implemented on behalf of SMEs and less advanced companies to maximise growth for all. There is a potential interest for leading companies to contribute to common data platforms, as they will also profit from faster growth and data sharing on a large scale. But this requires that common data spaces be able to provide a truly efficient and effective environment for data innovation providing advantages for all stakeholders.

To investigate these research question, we have combined desk and field research as follows:

- We leverage the rich database of 18 case studies of big data innovation developed by IDC and the Politecnico of Milano in the context of the H2020 DataBench project in 2019, and the desk research by Politecnico which collected approximately 700 big data case studies from literature, last updated at the end of 2019;
- We have carried out 3 case studies in-depth, Whirlpool in manufacturing, a leading Spanish bank for finance e-Geos for agriculture.

The report is structured as follows:

[1] Report on the European Commission's Workshops on Common European Data Spaces https://ec.europa.eu/digital-single-market/en/news/report-european-commissions-workshops-common-european-data-spaces

- Chapter 1 introduction, objectives and scope;
- Chapter 2 policy context and lessons learned from the DataBench case studies;
- Chapter 3 description and analysis of the 3 in-depth case studies
- Chapter 4 final considerations

2. Policy context

2.1 Rationale of Common Data Spaces

The digital policy package presented in February 2020 by the new Commission led by Ursula Von der Leyen widens considerably the scope and breadth of data policies, reflecting the new policy awareness about the critical role of data for the competitiveness of the European economy. In the "Political guidelines of the Commission 2019-2024"[2], the development of a Digital Europe is indicated as one of the six main priorities of the new Commission, recognizing its capability to support the achievement of other high-profile policy goals, such as a climate-neutral continent and a fair and open society. The transformation potential of data technologies is clearly recognized in von der Leyen's political agenda, as well as the need to understand and manage the whole range of data positive and negative impacts, with a specific attention to data ethics and the development of trustworthy Artificial Intelligence (AI).

The Communication "Shaping Europe's digital future"[3] is articulated in four main action areas, covering all the framework and enabling conditions to develop "a digital society based on European values and rules" (Technology for people, A fair and competitive economy, An open, democratic and sustainable society and Europe as a global leader). The new European Data Strategy is a cornerstone of the new Europe's Digital Strategy and underline the ambition for Europe to become a leading role model for a society empowered by data to make better decisions in business and the public sector.

A renewed ambition for Europe to become a global leader in the data-agile economy is accompanied by the need to achieve "technological sovereignty" based on a resilient and independent data infrastructure. The White Paper on Artificial Intelligence (AI) presented the same day suggests several policy options to develop AI ecosystems of excellence and trust, reflecting the understanding of the systemic impacts of technological innovation. European competitiveness in AI is clearly seen as a component of technological sovereignty, enabled and supported by the data infrastructure.

The European Strategy for Data' is articulated in four, mutually enhancing pillars. The Pillar A on the development of "A cross-sectoral governance framework for data access and re-use" brings together several initiatives to enable and stimulate data sharing, but at the same time ensure a fair playing field for all organizations. This recognizes the need to deal with data as a strategic asset influencing power dynamics in the socio-economic system. The Data Strategy's Pillar B on Enablers (Investments in data and strengthening Europe's capabilities and infrastructures for hosting, processing and using data, interoperability) and Pillar C on Competences (Empowering individuals, investing in skills and in SMEs) cover the needs for investments, interoperability, standardization and infrastructures as well as skills.

Finally, the Data Strategy Pillar D for "Common European data spaces in strategic sectors and domains of public interest" complements the three other pillars and responds to the need to improve access to data in Europe, facilitate and promote data sharing and therefore support the development of a critical component of data infrastructure, access to data.

Insuring access to high quality and real-time datasets has emerged as a critical competitive factor for the digital economy. There is clear understanding in Europe that global OTT (over the top) platforms such as Google, Facebook and their Chinese counterparts have accumulated a massive advantage in the hosting and control of massive consumer datasets, including those from European citizens. These datasets are also needed for machine learning and developing AI tools and services. Because of network effects, there is no way that a European Google can emerge and therefore European enterprises must work with these platforms as gatekeepers of access to data. The only way for Europe to even the competitive field is fighting for fair access to data and respect of privacy rights and citizens' control of their own data. The European Union is recognized as a world leader in the development of digital regulation in this field (thanks to GDPR the Global Data Protection Regulation, but not only) and plans to do the same with AI, as indicated in the White Paper on AI.

---

[2] https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf
[3] https://ec.europa.eu/info/sites/info/files/communication-shaping-europes-digital-future-feb2020_en_4.pdf

But there are many other sources of data other than consumer data, and these are still mostly unregulated and untapped. This means data from public and private organizations (data owners), B2B and B2G data (dataflows from interactions between businesses and businesses and governments), data generated from the huge number of sensors from the Internet of Things (IoT). These data flows are still mostly unregulated and unexploited, and access to these data in Europe is fragmented across borders even though specific policy initiatives are trying to prevent this (namely the Free Flow of non-personal Data regulation, May 2019).

The European vision is to create "*a single European data space – a genuine single market for data, open to data from across the world – where personal as well as non-personal data, including sensitive business data, are secure and businesses also have easy access to an almost infinite amount of high-quality industrial data, boosting growth and creating value, while minimising the human carbon and environmental footprint*"[4].

The development of Common data spaces will contribute to achieve this vision, with the specific task to help making B2B data sets actionable for data-driven innovation in Europe, across borders and within industrial value chains, enabling disruptive digital transformation.

To do so, data spaces are meant to foster an ecosystem of companies, civil society and individuals providing new products and services based on data. Enabling conditions for the data spaces are:

- EU investments in next-generation technologies and infrastructures to be applied in the data spaces, to support the creation of European data pools enabling Big Data analytics and machine learning;
- The public sector leading by example, both employing more data (and therefore increasing demand for data-enabled products and services) and facilitating the use of data for public good;
- Favourable data regulation also at sector level, removing barriers and incentivizing data sharing and exploitation;
- Investing in advanced ICT and data skills.

The potential business model for these common data spaces relies in providing incentives for organizations to contribute data, by providing them in return increased access to aggregated data from other data owners, analytical results or services based on data (e.g.predictive analytics) or license fees. Because of the specificities of different industries, the EC proposes to organize 9 common data spaces in the following domains of strategic value for Europe:

- A Common European industrial (manufacturing) data space
- A Common European Green Deal data space
- A Common European mobility data space
- A Common European health data space
- A Common European financial data space
- A Common European energy data space
- A Common European agriculture data space,
- Common European data spaces for public administration
- A Common European skills data space

The Common data spaces will aim at overcoming legal and technical barriers to data sharing across organisations, by combining the necessary tools and infrastructures and addressing issues of trust, for example by way of common rules developed for the space. They will include:

- the deployment of data-sharing tools and platforms;
- the creation of data governance frameworks;
- improving the availability, quality and interoperability of data – both in domain-specific settings and across sectors

Public support at the EC and/or national and regional level can include:

- Funding to support authorities in the Member States in making high value data sets available for re-use in the different common data spaces;

4 European Data Strategy, page 4

- Supporting data processing and computing capacities that comply with essential requirements in terms of environmental performance, security, data protection, interoperability and scalability.

The EC plans several actions to promote the development of the European common data space and of vertical data spaces. This includes the launch of a High Impact Project on European data spaces and federated cloud infrastructures (under Pillar B of the Data strategy) to fund infrastructures, data-sharing tools, architectures and governance mechanisms for thriving data-sharing and Artificial Intelligence ecosystems.

This report focuses on 3 Common data spaces, whose specific plans are the following.

A Common European industrial (manufacturing) data space

Goal: to support the competitiveness and performance of the EU's industry, allowing to capture the potential value of use of non-personal data in manufacturing (estimated at € 1,5 trillion by 2027).

The manufacturing industry is already quite advanced in the implementation of data-driven innovation but needs to keep pace with international competition and technology advances. This data space will focus on issues related to the usage rights on co-generated industrial data (IoT data created in industrial settings), as part of a wider Data Act foreseen by the Data strategy. This is critical since the combined adoption of Big data-IoT-AI is driving manufacturers' innovation investments and IoT is generally seen as a critical driver for a wide range of products and services. The Commission will also continue to work with the key players from the sector to overcome barriers to data sharing and find ways to further boost data generation notably from smart connected products which tend to go across traditional industries barriers.

A Common European financial data space

Goal: to stimulate, through enhanced data sharing, innovation, market transparency, sustainable finance, as well as access to finance for European businesses and a more integrated market.

This data space will build on the revised Payment Services Directive which has provisions on promoting open banking and innovative payment services based on leveraging data from bank accounts and their access. The objective is also to support access to credit for SMEs and individuals by helping financial actors use different sources of data, which may help making better and non-discriminatory credit decisions. In the credit crunch and difficulties following the recession caused by the Covid-19 pandemic, this aspect is likely to be very important. The Commission will further facilitate access to public disclosures of financial data or supervisory reporting data, currently mandated by law, for example by promoting the use of common pro-competitive technical standards.

A Common European agriculture data space

Goal: to enhance the sustainability, performance and competitiveness of the agricultural sector through the processing and analysis of production and other data, allowing for precise and tailored application of production approaches at farm level.

The use of digital technologies in agriculture until now has been limited, but the potential is huge for data-driven innovation to drive a jump forward increasing productivity and social welfare. The data space aims at accelerating strongly the pooling of public and private data through a neutral platform and ecosystem based on fair contractual relationships. Satellite data (the focus of our case study) is one of the categories of data which can bring strong advantages to the common agricultural platform. Europe's agricultural sector has many small farmers with weak contractual power towards technical providers and distributors, who can be helped by better access to information and greater transparency in the value chain. In addition agriculture is a heavily regulated sector and better monitoring capacity can help to reduce the administrative burden on governments and beneficiaries to check compliance and control the appropriate use of public funding. Also, as the traditional common agricultural funds are planned to gradually decrease in the next years, farmers may find requesting support for innovation investments as a productive alternative.

This agricultural data space will build on several already existing collaborative initiatives, including the declaration of cooperation signed by MS in 2019 on "A smart and sustainable digital future for

European agriculture and rural areas"[5], the Stakeholder code of conduct on agricultural data sharing[6], and the data spaces initiatives created by H2020 projects.

2.2 Lessons Learned from Big Data case studies

In this paragraph we analyse the lessons learned from the set of 18 case studies developed by Politecnico of Milano (Polimi) and IDC across 7 industries within the context of the H2020 DataBench project[7], and the desk research carried out by Polimi in the same period on 700 big data use cases from literature. We provide an overview of the main business benefits achieved by these stakeholders and the main use cases they implemented (Figure 1). We focus on the factors affecting the scalability and large-scale implementation of the investigated use cases, in order to understand the main industrial needs for the successful deployment of data-driven innovation.

These case studies show a good level of business benefits achieved from data-driven innovation, with a high level of cost reduction (such as 80% reduction of operational expenditures for fraud detection in the financial services industry, 30% reduction of maintenance costs in manufacturing thanks to predictive manufacturing) and customer benefits (for example, a 110% improvement of customer retention in manufacturing and 85% improvement of conversion rates from potential to actual customers thanks to data-driven targeting in retail). These case studies represent operational services and applications, but most of them are still confined to individual departments or branches of the company, in the process of being scaled-up to the whole organization. When scaling up these services, the lessons learned from the case studies highlight the following problems and risks:

- Innovative technology solutions are always tested first in a limited environment. Even when the technology solution has been well selected and designed, scaling-up always brings new challenges particularly in terms of cost effectiveness.
- Business superficiality: business intelligence is always needed to lead the strategic use of data, and this is still found in human resources rather than in machines. For example, automated recommendation systems built on standard solutions will tend to recommend the products with highest sales, which are likely to be the products with lower prices and lower margins. A business manager must intervene to provide an intelligent strategy, for example finding ways to nudge customers towards products with higher prices and margins. Integrating business intelligence with the use of data analytics is still immature in many industries.
- Most of these solutions rely on public cloud technologies. This creates a lock-in risk, since migrating to other cloud providers requires considerable costs and time investments for the redesign of software. The high concentration of the cloud providers market reduces the potential choices of business users and constraints the scaling up of data-driven innovation.
- When scaling up, particularly if the solution requires real-time data processing, the cloud computing costs tend to rise very quickly and cross a threshold where technology costs are higher than business benefits (edge vs. cloud decisions). This is a sensitive aspect, particularly for solutions combining BDA (Big Data Analytics) and Artificial Intelligence such as machine learning.

5 https://ec.europa.eu/digital-single-market/en/news/eu-member-states-join-forces-digitalisation-european-agriculture-and-rural-areas

6 EU Code of conduct on agricultural data sharing by contractual agreement, 2018

https://copa-cogeca.eu/img/user/files/EU%20CODE/EU_Code_2018_web_version.pdf

7 Evidence-Based Big Data Benchmarking to Improve Business Performance, www.databench.eu

The cloud computing costs issue is one of the relevant results of the case study analysis. Basic cloud services are quite convenient, but prices increase very fast as soon as more sophisticated services are needed. For example, in the retail industry, a leading supermarket chain found that leveraging AI for sales prediction in one shop led to a 5% increase of margins (equivalent to roughly 5 million €/year); but applying the same machine-learning application to all shops in the chain would wipe out the benefits and cost more than the margin increase. These issues are behind the decision of the European Commission to promote federated European cloud infrastructures. The European data strategy in fact plans to fund a High Impact Project on European data spaces and federated cloud infrastructures.

These case studies confirm the need for improved access to data processing and computing capacities, as foreseen in the Data Strategy in terms of support for data spaces[8].

Figure 1 – Big Data success stories: main use cases and business impacts, 2019

Agriculture	*Crops monitoring*: Costs = -10%	*Equipment optimization*	*Precision agriculture*
Automotive	*Predictive maintenance*	*Self driving*	*Smart services*: Costs = -80%
Financial Services	*Fraud detection*: Operational Ex. = -80%	*Risk assessment*	*Targeting*: Marketing costs = -35% TCO costs = -80% Conversion rate = 10x
Healthcare	*Diagnostic*	*Patient monitoring*	*Preventive systems*
Manufacturing	*Predictive maintenance*: Maintenance costs = -30%	*Smart manufacturing*: Utilities costs = -20% Cust. retention = +110%	*R&D optimization/ Smart design*
Retail	*Assortment optimization/ Intelligent fulfilment*	*Price optimization/ Promotions*: Conversion rate = 50% Cust. retention = +14%	*Targeting*: Conversion rate = +85% TCO costs = -15%
Telecommunication	*Churn prediction/ Promotions*	*Network capacity optimization*	*Targeting*: Conversion rate = +130%
Transport & logistics	*Churn prediction/ Promotions*	*Fleet management*	*Network capacity optimization*: TCO costs = -90%
Utilities	*Churn prediction/ Promotions*	*Network capacity optimization*: Costs = -20% Cust. Expenses = -30%	Personalized fares: Marketing costs = -50% TCO costs = -50%

Source: Chiara Francalanci, Politecnico of Milano, "Virtual BenchLearning: Success Stories on Big Data & Analytics", DataBench Webinar 28 May 2020 Note: the sample includes the 18 case studies based on direct interviews and the 700 case studies from desk research.

[8] European Commission, "A data strategy for Europe", February 2020, page 17

3. The Case Studies

3.1 Whirlpool

This case study analyses a case of predictive analytics in the manufacturing industry and the current problematics concerning data sharing even between and innovative manufacturer and its suppliers of factory plants.

Case study Description

Whirlpool Corporation is a leading global home appliance company, with approximately $20 billion in annual sales, 77,000 employees and 59 manufacturing and technology research centers in 2019. The company results from a merger between the American corporation of the same name and Indesit, an Italian multinational. In the EMEA region the group employs approximately 19,700 employees with operations in 13 manufacturing and technology research centers in 6 countries (Italy, UK, Poland, Slovakia, Russia and Turkey). The EMEA headquarter is in Italy where 6 manufacturing plants are located. Global strategies are decided in the US, but decentralized region headquarters have a high degree of freedom. The case study is based on interviews carried out in Italy which is the leading reference point for the region for technology innovation strategies including on Big Data. The company defines itself as a smart follower, rather than a pioneer in terms of technology.

The focus of this case study concerns a pilot carried out within the context of the H 2020 project Boost 4.0[9] for the application of Big Data to forecasting spare parts demand. The pilot will be concluded in December 2020 and even though the final phase has been slowed down by the Covid-19 pandemic, the preliminary results are available.

The company produces home appliances, so spare parts for repair must be sent out to the final customers' location. Spare part production and distribution is one of the most relevant challenges for after sales services, requiring careful and timely managing of central warehouses of spare parts (so customers don't have to wait too long for reparations), with a large variety of product families and product codes. There is also an issue of data silos within the company: for example, the logistics dataset on distribution of products by geography is separate from the maintenance dataset by customer services. The headquarter does not know from where the demand of spare parts is likely to come, by type of product and part.

This is changing with the diffusion of smart appliances which can be connected to the Internet and provide data directly to the manufacturer about intensity of use and potential problems. However, at the moment they are still a small part of the installed base (4-5% at the end of 2019) and customers must give consensus (according to GDPR) for the use of data; apparently not many actually do. The company is thinking about incentives for customers to share data, for example with a gamification approach through mobile apps (as is being pioneered in the automotive industry). At the moment, the data flow from smart appliances is not sufficient and the process to manage the spare parts stock and inventory is managed by human resources.

The pilot was designed to forecast the spare parts consumption based on a Big Data lake including historical spare parts demand data, product quality data from manufacturing plants (tests from the production line), customer feedback data (service orders), and connected devices real time data (current status of the installed base). Currently, the forecast is based on historical spare parts demand data only and several manual corrections of the predictions and the planned stock levels are needed in order to achieve a satisfactory service level for customers. The pilot developed new forecast models based on big data technology and analytical techniques from SAS, in order to improve the accuracy of the forecast estimations and optimize the subsequent Demand Planning Process.

Driven by the need to improve the exploitation of all the data they generate, the company is gradually merging the different datasets into a data lake and investing in new ways to access and use the data. The company implemented a new platform for self-service analytics for the internal users (based on Tableau), so that they can access the relevant data in a more flexible and

9 Boost 4.0 Big Data for Factories, Lighthouse project 2017-2020. https://boost40.eu/

personalized way. They started implementing few services and gradually added new ones, so the platform is growing. The aim is to enable users to develop new applications on top of the company data, for example develop new graphs or reports without requiring the IT department support as previously done. The business areas included in the platform are: customer services, finance, connectivity (data collected from smart appliances), digital area.

Main Big Data Business Benefits

The spare parts pilot preliminary results show that the company is on track to achieve a 15% reduction of the spare part stock (which means a cost reduction), an increase of inventory turnover by 35%, a reduction of 25% of the lead time to consumer and 1% improvement of the overall service quality of the plant. Once implemented at scale, the company expects business benefits in terms of improvement of the quality of customer service for repairs and maintenance with lower costs. In the long term (5 years) the company wants to integrate the data from smart appliances at the customer premises and continue increasing the efficiency and effectiveness of the spare part management process.

In the long term they also hope to exploit the data stored in the data lake with advanced analytics applied to use cases such as customer satisfaction and targeting, customer services and marketing analysis, for example assessing customer reviews about appliances.

Main Challenges

The business case for the extensive use of big data is still an issue. Creating a data lake is very expensive and performing only exploratory analytics without clear business goals is not going to provide a sufficient return on investment. It is a question of culture and skills; business expertise as well as specific data analytics skills are needed to develop models to extract value from these data.

The volume of data collected on production processes is approximately 50 GB/year on a product category per plant. Smart appliances could generate much higher data flows requiring analytics to add value (simply knowing how many times a customer uses a washing machine does not mean much).

There is also a storage and processing cost issue. The data volume and processing requirements for advanced big data applications require high computing capacity. The alternative is between investing in a local IT infrastructure or entrusting a cloud provider (for example Google), whose costs are low for storage but high for data queries, constraining scalability and real time applications.

Considerations on data sharing

The pilot on spare parts management does not foresee data sharing with third parties, but it could in principle. In the future it could once the data from smart appliances on customers' premises as well as data from distributors was integrated. The company is still working on overcoming data silos and developing analytics with a business impact. Apart from investment and processing costs, the main barrier is knowledge and combined business and data analytics expertise.

The company is also developing a "digital twin" experimentation in Poland but is struggling to merge different datasets, particularly to negotiate an agreement with the plant suppliers. According to the interviewee, there are cultural and technological barriers preventing data sharing between the manufacturer and his sub-suppliers. The suppliers are focused on collecting data from the production cycle for maintenance and improving efficiency but are not willing to share the data. In addition, the different typologies of data and methods of collection create technological barriers.

3.2 A Spanish Financial group

This case study analyses the sharing of customer data between a bank and specialized service suppliers through a synthetic dataset, considering the specific barriers to data sharing in the financial industry.

Description of the case study

This case study concerns a leading large Spanish Financial group, including retail banking and insurance activities. The focus of the case study is on a pilot developing a dataset about the customers' interactions with the bank and using analytics to improve the detection and avoidance of misbehaviour and illicit uses of the group financial services. Most interesting, the plan is to develop a synthetic dataset (thereby respecting the anonymization and privacy protection requirements of GDPR, the General Data Protection Regulation) which can be shared with the group's external providers to deploy and validate proofs-of-concept of different use cases (e.g., potential fraud based on customers' relationships). The ultimate objective is to enable the group to outsource fraud detection to external third parties, without needing to disclose sensitive personal data and maintaining control of the relationship. Therefore, this is a case of data sharing in the value chain between a data owner and service providers.

The synthetic dataset is generated based on real data from various sources (such as customers' operations on the online banking platform, interactions with automated cashiers (ATM) and so on), in the form of a set of restricted tables (relational database).

The volume of the real data used to generate the synthetic dataset is truly big data, representing approximately 12 Petabytes and growing approximately 7Terabyte per day. The original data collected is structured (basically customers connections) and unstructured (SIEM logs). From this information, and depending on the case objective, the generated data can be structured (transactional) or unstructured (graph) containing part of the social graph.

Main Big Data Business Benefits

The group has not yet measured the business impact of the pilot, but they expect to achieve time efficiency and accuracy improvements in the identification of potential misbehavior by customers and control of fraud; cost reduction in the processing of the huge amount of data about customer behavior and potentially develop better services and improve customer satisfaction.

To provide an example, one of the use cases focuses on the detection of anomalies related with bank transfers and cash deposit. The bank has detected, for instance, cash income at unusual time windows using business rules, however, discovering a generalized pattern is a difficult and time-consuming task, even though it has a high added value. To be able to automate and delegate the discovery of such anomalies to third parties would free up resources in the bank, improve productivity and of course reduce risk and costs by improving the accuracy of detection of potential frauds. Also, it would improve customer satisfaction.

Main challenges

The main challenge is to analyze the enormous quantity of data. The social graph of customer data is stored in a relational database, but they are looking for a more scalable and performant solution. In the medium term the focus is on improving the speed of data analytics which is particularly important for financial fraud detection.

Considerations on data sharing

The financial group interviewed is one of the leaders in data-driven innovation in the industry with multiple experimentations ongoing. Nevertheless, they are very cautious concerning data sharing. The financial industry is a particularly hard case for data sharing because of the sensitiveness of the personal data, second only to healthcare in the public opinion. In this case the approach chosen has been to develop a synthetic dataset, which is complex and quite expensive, but in the field of fraud detection in the financial services it can be acceptable. The motivation for data sharing is interesting because it is driven by the need to outsource a highly specialized task requiring an expertise different from the range of skills typical of the financial sector.

3.3 E-Geos

This case study concerns the use of satellite data in the agriculture domain, with a focus on crops yield prediction. The specific typology of satellite data and its potential role for data sharing in agriculture are examined.

Description of the case study

E-Geos, an ASI (20%) / Telespazio (80%) company, is a leading international player in the Earth Observation and Geo-Spatial Information business. e-GEOS is the global distributor for the COSMO-SkyMed data, a constellation of four radar satellites for Earth Observation, founded by the Italian Space Agency and the Italian Ministry of Defense. e-Geos own data is private because it is generated by a national satellite constellation, with Italian military having the first priority on the use of data. The company provides big data analytics application services covering the whole value chain, from data acquisition to the generation of analytics reports, based on the integration of different sources. Thanks to proprietary assets and algorithms, e-Geos integrates data from all satellites with the IoT information gathered over different sources, creating a big data lake from which all the company platforms are able to extract significant indicators dedicated to different markets. This approach is one of the key assets of the new services and products offered by the company and it can be run both on premises and on cloud. E-Geos participates to the European Copernicus Program, the European Union's Earth Observation Programme.

e-GEOS application platforms include services for environmental protection, rush mapping in support to natural disaster management, specialized products for defense and intelligence, oil spill and ship detection for maritime surveillance, interferometric measurements for landslides and ground subsidence analysis, thematic mapping for agriculture and forestry.

The USA and particularly NASA were the pioneers of the use of satellite data and until recently the dominant players with the Landsat constellation. Since 2015, when the European satellite Sentinel 2 was launched, Europe has gained greater autonomy and control on the data. This is part of a process set in motion by the European Union to place a constellation of almost 20 more satellites in orbit before 2030, designed to support the Copernicus Programme.

Satellite data is different from other typologies of data and generates a different type of market. It is classified based on the level of imaging resolution (detail of the image on the ground) which influences also the type of application of the data. Spatial resolution refers to the size of one pixel on the ground. A pixel is that smallest ‘dot’ that makes up an optical satellite image and basically determines how detailed a picture is. Landsat (the NASA satellite constellation for Earth observation), for example, has a 30m resolution, meaning each pixel stands for a 30m x 30m area on the ground. It is considered a medium-resolution image, which can cover an entire city area alone, but the level of detail isn’t fine enough to distinguish individual objects like houses or cars[10]. Today, satellite data resolution is classified in three main categories (whose boundaries are constantly shifting since imaging technology is in continuous progress):

- Low resolution: over 60m/pixel
- Medium resolution: 10 – 30m/pixel
- High to very high resolution: 30cm – 5m/pixel

The finest resolution as of now is 30cm provided by very high-resolution commercial satellites. Low and medium resolution satellite data is useful for monitoring the environment at a large scale, for example for monitoring the level of growth of crops in the Po plain in Northern Italy or covering very large farms, but is not adequate to monitor small farms.

Satellite data is mostly open, normally excluding very high resolution data. Nevertheless, satellite data is basically raw material which needs to be processed and treated with sophisticated tools and or merged with other data before it can be used. In addition, most satellites are partially funded by defense ministries and if there are national security needs, they can be re-directed causing the suspension of contracts with non-military customers. This happened for example in Europe during the war in the Balkans.

10 EGeos Blog https://eos.com/blog/satellite-data-what-spatial-resolution-is-enough-for-you/

This case study concerns a commercial application of satellite data for the agriculture sector. E-Geos developed an innovative yield prediction machine learning algorithm based on Sentinel and Landsat open data high-resolution satellite information. The algorithm was used to predict the production of soy beans and corn in the US on behalf of a company operating in the financial industry, needing predictions as a support to trading decisions.

This approach was based on the analysis of optical Multi-temporal Time Series to detect temporal trends and behaviours of different targets of interest. A critical success factor was the use of a consistent dataset, created through the harmonization of different EO (Earth Observation) data sources such as Landsat, Sentinel-2 and MODIS and the application of systematic data analysis and classification tools in order to detect the targets and behaviours of interest.

The integration of data from different types of satellites was used to solve the following problems:

- Problem #1: Detection of emerging crops with a confidence level higher than 90% over a user defined large area;
- Problem #2: Detection of crops at the end of the phenological cycle with a confidence level higher than 95% over a user defined large area.
- Problem #3: Detection of grassland mowing over countries such as Italy, the Netherlands, Lithuania, Romania, Spain

In terms of the use of AI, e-Geos used multiple families of supervised or unsupervised classification algorithms applied to multiple combinations/subsets of the provided input data.

The company is still working on developing the algorithm. In the short term (1 year) the plan is to use feedback from the user in order to self-learn and improve subsequent analysis/classification activities. In the long term (5 years) the plan is to refine further the recommendation engine, providing to the user a limited number of possible solutions, asking for an additional feedback or training in order to further improve the classification/analysis results.

Main Big Data Business Benefits

The goal was to predict crops yield for a financial company active in the trading market. The machine learning algorithms demonstrated roughly 10% more accurate than previous forecasts approaches, supporting better investment decisions, which helped the financial company to gain from trading around 0.32% on traded volumes. This is a very valuable gain for the competitive trading market.

Further improvement of the service quality (quality of forecasting) is expected in the near term, with the improvement of the forecasting capability of the system thanks to additional feedback and development of interpretation of the data.

Main challenges

One of the key challenges is screening the quality of the input data which affects the machine learning algorithm. This step is currently semi-automated, as it is required to focus on framing the problem and understanding the quality of the output with the goal of designing an innovative and problem-specific effective approach to data screening. The project is able to manage different quality levels associated with different portions of input data (e.g. data from Sensor 1 are considered more reliable than data from Sensor 2).

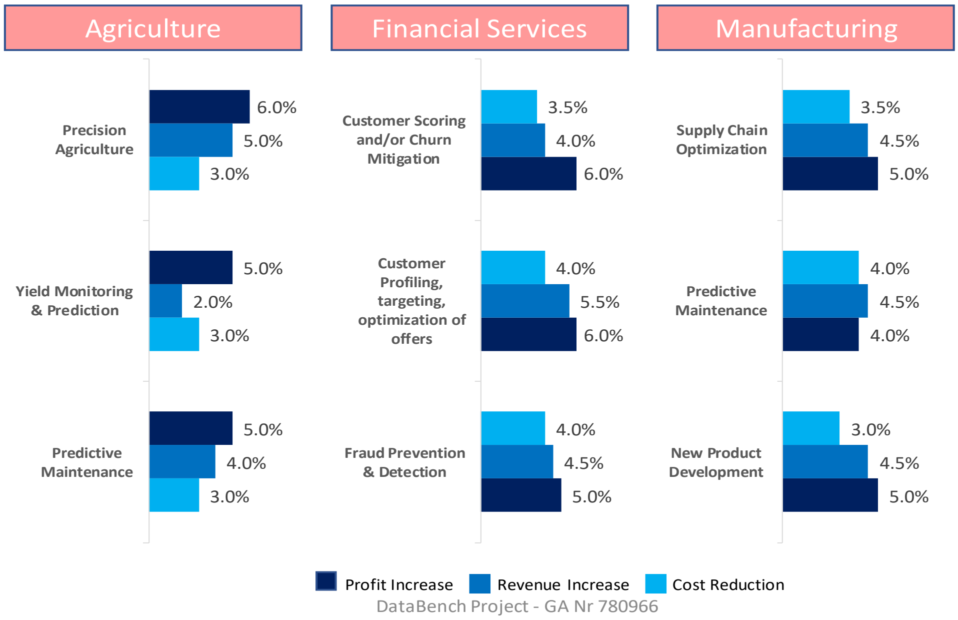
Considerations on data sharing

The yield prediction machine learning algorithm developed by e-Geos is designed to be portable and therefore re-usable: for this first iteration it was deployed on AWS (Amazon Web Services) but it can be deployed theoretically on any cloud environment, being it public or private, with minimal adjustments of the interfaces and of the IaaS basic services. The dataset is open, so it can also be re-used. However, its applicability is valuable only for large-scale monitoring of crops, because of the level of resolution of the data, rather than for individual farms. It could be used for example by national agencies interested to monitor crops or livestock breeding, their compliance with national programs or initiatives in order to provide funding. For example, e-Geos collaborates with AGEA (Agenzia per le Erogazioni in Agricoltura) the Italian agency distributing European agriculture funds.

4. Final Considerations

The research and case studies presented in this report provide insights from European industrial stakeholders engaged in experimenting and scaling up data-driven innovation. They show that there has been considerable progress in the use of data-driven innovation by European industries in the last years, including an increasing use of AI techniques such as machine learning leveraging the power of data (Figure 4). Nevertheless, there is still a high level of immaturity in the capability to merge datasets within a company, and relevant barriers against data sharing even in advanced sectors such as manufacturing. Enterprises are still concentrating most of their efforts and investments on developing data-driven innovation within their organization, at best sharing some data with a few trusted sub-suppliers.

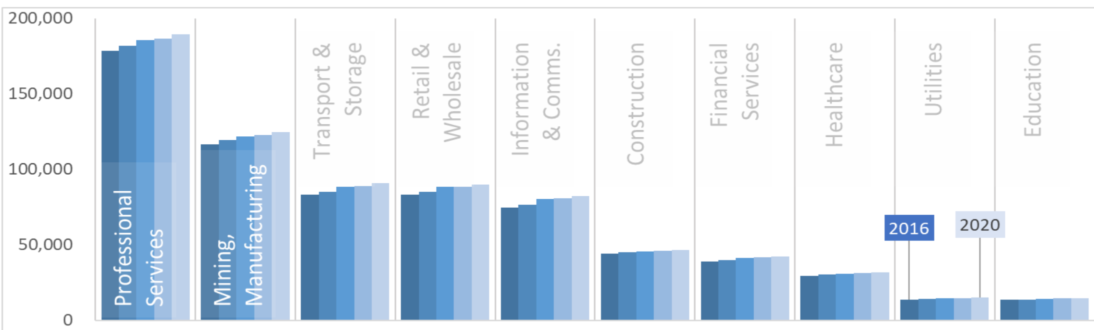
Figure 2 Benchmark KPIs by Industry, 2019 (% increase thanks to adoption of Big Data)



Source: DataBench Project -D2.4 – Benchmarks of European and Industrial Significance (December 2019) – www.databench.eu

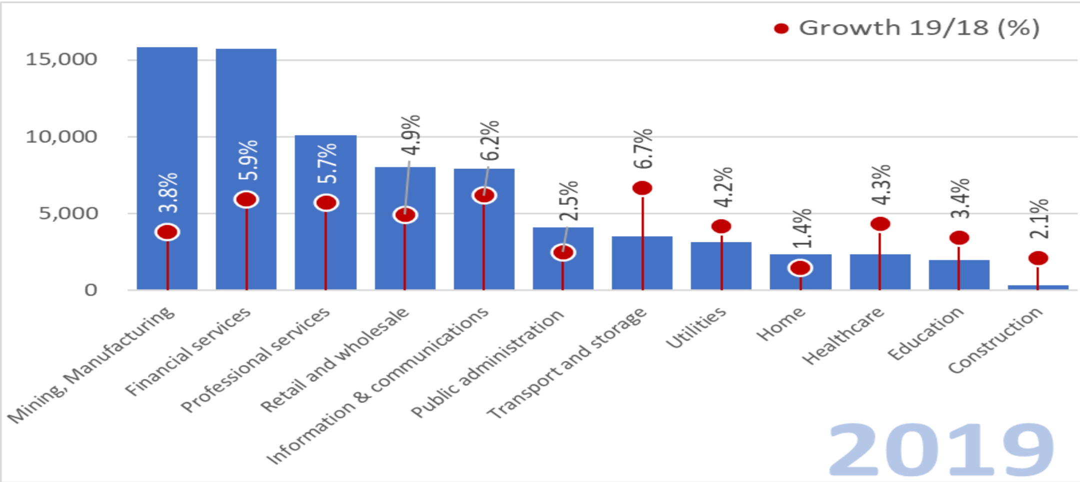
The goal of Common data spaces is to overcome legal and technical barriers to data sharing across organization, by combining the necessary tools and infrastructures and addressing issues of trust, for example by way of common rules developed for the space. For this reason, common data spaces will need to have a pragmatic approach and respond to the requirements of stakeholders. Before drawing common conclusions, it is important to consider the specificities of the three vertical domains considered by this report, also leveraging the data from the European Data Market monitoring tool (Figures below).

Figure 3: Number of Data Users by Industry, 2016-2020 (Units)



Source: European Data Market Monitoring Tool, IDC 2020 – www.datalandscape.eu

Figure 4: Data Market Size, Growth by Industry 2019 (€, Million; %)



Source: European Data Market Monitoring Tool, IDC 2020

Manufacturing

From the point of view of the economy, the common European industrial data space is probably the one with the highest potential impact. As shown by the EDM monitoring tool (Figure 2), manufacturing is the sector with the second highest number of data users (organizations that use data intensively and strategically), over 124,000 enterprises. Manufacturing is the first sector for the value of the data market (17B€ corresponding to 21% of total for the EU27) with good forecast growth perspectives. The top 3 use cases of data innovation, sourced from the benchmarking survey of the DataBench project, are supply chain optimization, predictive maintenance (similar to the use case analysed in this report) and new product development. According to the DataBench research, the median improvements from the use cases are approximately 5% for profits, 4.5% for revenues and 3% for cost reduction, showing that manufacturing enterprises focus more on increasing their business performance than cutting costs. All these three use cases would achieve better results from data sharing across the value chain and access to multiple high-quality data sources.

Manufacturing is in the second-best group of sectors in terms of the level of business impacts, better than most but following after profit leaders such as finance or telecom/media. There seems to be still a large potential of improvement. These business improvements are median values of a large sample, the variability of individual results in case studies is quite large and we have seen several cases of much higher benefits.

Manufacturing enterprises have been engaged in deep digital transformation of their industrial processes for years, having to deal with subsequent waves of innovation. Now they are engaged in moving towards widespread automation and deepening the use of technologies such as IoT. IDC believes that there are strong demand trends toward supply chain integration and stronger interaction with partners and customers, for example IDC expects that by 2022, 70% of manufacturers across industry will use cloud-based innovation platforms and marketplaces for cross-industry and customer co-development that creates 50% of new products and service ideas[11].

These general considerations are coherent with the results of our case study. Whirlpool is not a pioneer but a close follower in the industry and is engaged in multiple innovation investments for operational technologies, digital technologies and specifically big data. Nevertheless, they are still struggling with internal data silos and lack of skills for exploiting data analytics for the sake of business intelligence. Their considerations about the implementation of digital twins are spot on, compared to IDC's assessment of the level of maturity of the sector. IDC believes that 60% of manufacturing companies will engage in digital twins experimentation by 2021, but that only 15% will reap tangible business benefits – for lack of full integration into the main business processes. Concerning digital twins, Whirlpool also complained about the difficulty of data sharing with

11 IDC Manufacturing Insights, Manufacturing Futurescape, 2020

supplier partners to collect the data flows from production machinery for technical barriers (different typologies of data) and mistrust by the suppliers.

In another DataBench case study, a producer of factory lines for manufacturers made the opposite complaint, that customers did not want to share their operational and maintenance data and it was difficult to collect maintenance data across different factories in order to implement predictive maintenance forecasting models. Currently, they are interesting to experiment homomorphic encryption which allows to elaborate the data for forecasting purposes without decrypting it (sending the algorithm to the data rather than collecting the data together).

Therefore, based on the analysis presented here, manufacturing stakeholders could be attracted to Common European data spaces providing good quality data analytics services and business intelligence about the best ways to exploit data analytics. The simple availability of datasets will not be sufficient, since many companies are still struggling simply to make sense of their own data. Clearly a sound business model for the data common space to offer not only data governance frameworks, but also ways to monetize data, would be necessary for companies to agree to share their data.

The report on the workshop on “Common European Data Spaces in Smart Manufacturing” held in Luxemburg in September 2019 is coherent with the issues emerging from our field research. The suggestion to provide “sandboxes” for experimentation of data sharing to show that new business models are possible and how they can be adapted to B2B value networks is particularly interesting and fits with the industrial needs observed here. However, it is important to proceed in practice and implement such initiatives. For example, the BDVA paper “Towards a European data sharing space” provides useful guidelines to create conditions for development of a trusted European data sharing framework, and to support EU businesses to safely embrace new technologies, practices and policies. Also, the IDS (Industrial Data Space) association is proposing a standard enabling data interoperability while ensuring data sovereignty and is building a community of business users to implement it.

Finance

The financial sector and particularly retail banking and insurance are in the middle of a disruptive innovation wave driven by fintech companies experimenting new online payment method and different financial relationships. New competitors are challenging retail banking, such as consumer good companies providing fidelity cards with credit and payment services to their customers. In the finance sector digital technologies represent the core operational technology and is not a case if the value of their data market is 16 B€ in 2019, corresponding to 21% of the European data market, same value as manufacturing even though the sector is much smaller in terms of overall revenues and employees. Looking at business impacts, finance reaps the highest profit and revenue increases from big data compared to other industries, according to the DataBench analysis. The leading use cases are customer scoring and/or churn mitigation, customer profiling and fraud prevention and detection (which is the focus of our case study).

The prevalence of personal data in this industry creates a very high barrier to data sharing. Financial companies are very scrupulous generally in respecting GDPR. As shown by our case study from a leading banking group with multiple innovation investments, they are now experimenting with synthetic datasets in order to guarantee the anonymization of data.

Stakeholders though are moving in a different direction, for example at the financial industry workshop on data-driven innovation organized by the European Commission last July 2019[12] the trend for data sharing was “moving the algorithm to the data”, in other words find ways to leave the data in possession of the data owners and find ways to carry out analytics without disclosing sensitive information. According to the workshop participants, “Data lakes are not strictly necessary for most application scenarios” also because they may lead to risks of security breaches.

Agriculture

Unfortunately, the EDM monitoring tool does not provide specific data for agriculture. However, another survey carried out in 2019 on a representative sample of European enterprises for the

[12] Report on the workshops on common European data spaces, https://ec.europa.eu/digital-single-market/en/news/report-european-commissions-workshops-common-european-data-spaces

ATI (Advanced technologies for Industry) study[13] found that approximately 18% of agricultural companies with more than 10 employees are already using Big Data, another 7% plans to do so, but over 60% are not using and have no plans. This is compared to 39% already using for the whole sample. The only technology with high take-up in the agriculture sector is public cloud with 75%.

The leading use cases for Agriculture are precision agriculture, yield monitoring and predictive maintenance. Even if DataBench found few organizations already implementing big data innovation, their return is impressive. This is a sector with a history of low usage of digital technologies but a high potential.

Our use case looks at crops yield prediction based on satellite data and confirms such positive perspectives. Looking in a broader sense at the potential of leveraging satellite data sharing for agriculture, it is clear that much more could be done to exploit such data. The EU plans for a common data space for agricultural data foresee the establishment of a neutral platform for sharing and pooling agricultural data, and already includes Earth observation data sourced from satellites as one of the main data sources. However, as anticipated satellite data requires sophisticated analytics tools whose development and execution is expensive. To provide access to farmers to this type of data in a common data space requires a governance framework where the costs of data processing can be compensated and specialized intermediaries like e-Geos can provide the necessary tools.

Satellite data is not well suited for many precision farming use cases, for example optimizing the use of water or of pesticides, which require detailed data for the specific field or cultivation. These data are better sourced from IoT sensors close to the ground or used by machines such as tractors. However, one application area of satellite data of high interest for farmers not yet fully developed concerns the assessment of environmental risks for insurance contracts on farms. Given the different priorities and potential conflicts of interest between the two categories, this application would benefit from being developed in a common European data space insuring that all parties' interests are balanced.

Final remarks

The results of our analysis show that the path towards data spaces effectively supporting data sharing at ecosystem level will not be easy. Common data space plan to overcome technical and regulatory barriers with deep roots. The research on industrial requirements for such initiatives highlight that the key success factor for common data spaces will be the ability to provide concrete business value added for stakeholders in order to convince them to share their data. The approach designed for common data spaces, focusing on the deployment of data-sharing tools and platforms, the creation of data governance frameworks and improving the availability, quality and interoperability of data, goes in the right direction.

If we consider the main barriers to data sharing that are well known these are the considerations relevant for common data spaces:

- Economic barriers: extracting, curating and making datasets available is expensive and organizations are only slowly understanding the business case for making these investments. It is important to invest in building business intelligence and libraries of business cases in the vertical domains.
- Technical barriers: this concerns merging different typologies of data, lack of interoperability and standards, data anonymization, security and data protection challenges. While research challenges will be dealt with by the new FP Programs Digital Europe and Horizon Europe, it would be useful for Common data spaces to leverage concrete standard and collaboration proposals guaranteeing data sovereignty such as the IDS proposal for manufacturing.
- Organizational barriers: data silos due to separate business processes prevent effective data exploitation, lack of data analytics and data processing skills is a constraint. The common data spaces should lead the way to raise awareness and fight a cultural barrier on this front, helping organizations to re-train their human resources in this area.
- Lack of trust, fear to lose control of proprietary datasets representing competitive secrets, uncertainty about compliance with regulation: again, these barriers which appear in all

[13] Add source

case studies considered and are deeply rooted can be fought but by supporting cultural change and providing protected environments for experimentation, as well as governance models ensuring data control by stakeholders and data protection as well as security.

Finally, a relevant issue which emerged from most case studies is the availability of affordable and efficient cloud computing infrastructures, allowing the scaling up of successful pilots and individual company-site experiences to the whole company domain. Even if potentially Common data spaces could provide a valuable answer to the need for greater access to high quality datasets and computing infrastructures, this will require solving practical and technology challenges, not simply providing a favourable environment for encouraging stakeholder collaboration.

References

Databench project, deliverable D4.2 “Data collection results” https://www.databench.eu/public-deliverables/#

Databench project, deliverable D2.4 “Benchmarks of European and industrial significance” https://www.databench.eu/public-deliverables/#

Chiara Francalanci, Politecnico of Milano, “Virtual BenchLearning: Success Stories on Big Data & Analytics”, DataBench Webinar 28 May 2020

BDVA and EU Robotics, SRIDA-Strategic Research, Innovation and Deployment Agenda, September 2019 http://www.bdva.eu/node/1359

Member States **Declaration of cooperation on smart and sustainable digital future for European agriculture and rural areas**, April 2019, https://ec.europa.eu/digital-single-market/en/news/eu-member-states-join-forces-digitalisation-european-agriculture-and-rural-areas

European Commission “Shaping Europe’s Digital Future”, February 2020, Luxembourg: Publications Office of the European Union, 2020, https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/shaping-europe-digital-future_en

European Commission Communication, “A European strategy for data”, Brussels, 19.2.2020 COM(2020) 66 final https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

European Commission, Digital Europe Programme: a proposed €9.2 Billion of funding for 2021-2027, June 2019 - https://ec.europa.eu/digital-single-market/en/news/digital-europe-programme-proposed-eu92-billion-funding-2021-2027

European Union Regulation on the Free Flow of non-personal data, May 2019, https://ec.europa.eu/digital-single-market/en/free-flow-non-personal-data

FAO (Food and Agriculture Organization) “Harnessing Artificial Intelligence's transformative power to make food systems more efficient, sustainable and inclusive”, February 2020 http://www.fao.org/news/story/en/item/1263916/icode/