European Big Data Research for Industry

3 Projects
7 Sectors
9 Applications
41 Software components

Now what?

Introduction

On Tuesday November 3rd, 2020, within the framework of the European Big Data Value Forum, Big Data EU H2020 research projects I-BiDaaS (Industrial-Driven Big Data as a Self-Service Solution), BigDataStack (High performance data-centric stack for Big Data applications and operations) and Track & Know (Big Data for Mobility Tracking Knowledge Extraction in Urban Areas) hosted a joint session.

The collaboration between the three EU projects was initiated in the beginning of 2020, when during the BDV PPP Summit 2020, they decided to join forces in a series of 9 online demonstrations of innovative Big Data Technologies in the pilot studies and their applicability to an ever wider scope contributing to Europe's digital future: the Big Data PIlot Demo Days.

After 3 years of research and innovation, I-BiDaaS, Track & Know and BigDataStack joined forces again in the expert-led discussion on the impact and uptake of Big Data research results. The purpose of the discussion was to (i) Identify shared barriers to adoption of Big Data research in different sectors, and mechanisms to overcome them, (ii) map the current and future impact and sustainability of their Big Data research, (iii) share best practices on the concrete business questions that have been answered in the project pilots.

This joint report aims to summarise the key findings derived from the session with recommendations on **European Big Data Research for Industry.**
3 projects.
7 sectors.
9 applications.
41 software components.

Now what?

Who joined BigDataStack, I-BiDaaS and Track & Know in the discussion?

In total, 117 attendees from more than 14 countries attended the session (see figure 1), a number that significantly increased the visibility of the projects.

EBDVF 2020 - Parallel session on European Big Data Research for Industry. 3 projects. 7 sectors. 9 applications. 41 software components. Now what?
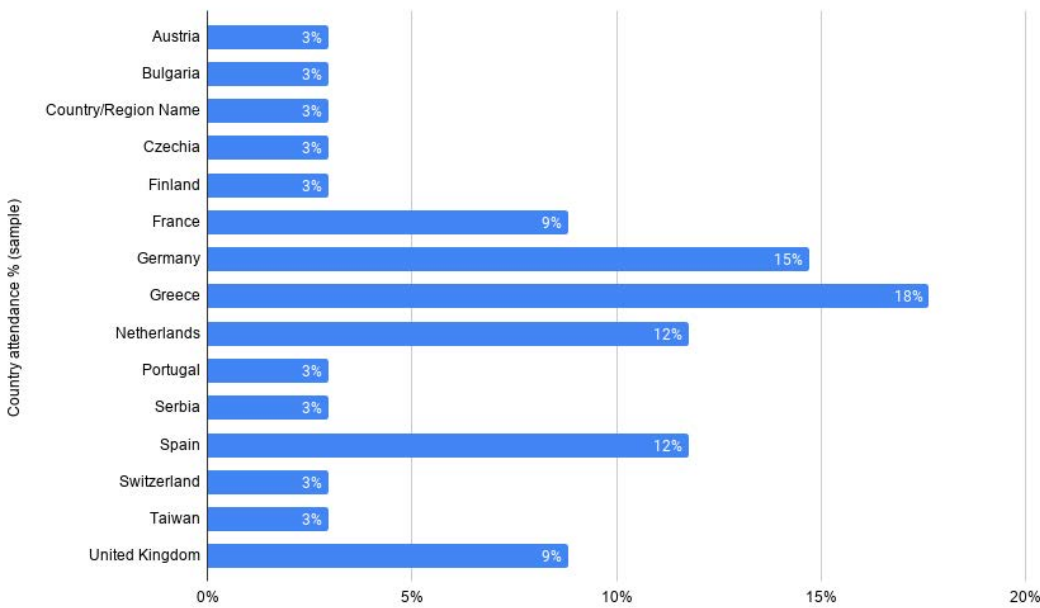

Figure 1: Geographical Spread

Fostering engagement, the attendees were asked a few questions in order to understand their background and interests. From the responses received, the majority of the attendees were from Research and Academia (62%), followed by Big Data Technology providers (23%) as shown in figure 2. An overwhelming 92% worked with Big Data and 85% were interested in Big Data technologies to improve their customer experiences, the main barrier to achieve this was considered the lack of skills (54%) as shown in figure 3.
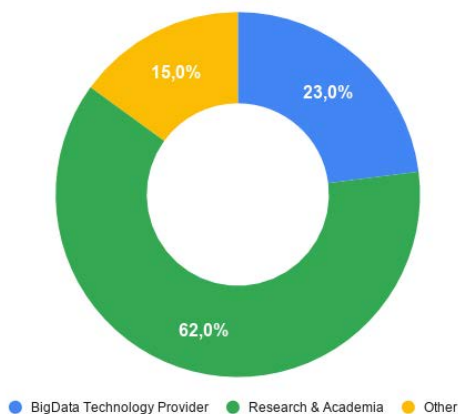

Figure 2: To which of our stakeholder types do you belong?
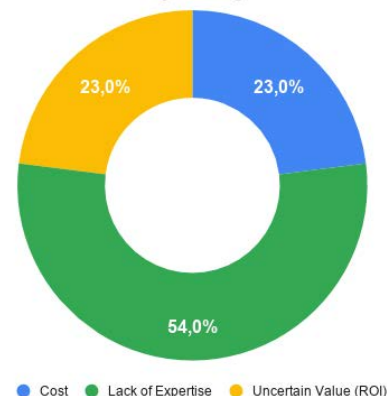

Figure 3: What is the main barrier or risk preventing you from implementing Big Data Analytical solutions in your organisation?

European Big Data Research for Industry.
3 Projects, 7 Sectors, 9 Applications, 41 Software components.
Now what?

BigDataStack, I-BiDaaS and Track & Know developed three aligned infographics, showcasing the main results ready for further adoption. The infographics were shown during the session and highlighted here below.

BigDataStack
Holistic stack for big data applications and operations

BigDataStack is a unique high-powered stack of solutions with a focus on providing fully efficient and optimised processes across the complete set of technologies required by data operations and data-intensive applications.

Check out Website & Social
- www.bigdatastack.eu
- @BigDataStackEU
- company/bigdatastack
- ZENODO Search: bigdatastack

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under grant agreement **No 779747**.

3 We made applications for **3 SECTORS**
- Retail
- Insurance,
- Shipping

14 PARTNERS

3 APPLICATIONS

17 SOFTWARE COMPONENTS | 14 Open Source 3 Proprietary

I-BiDaaS
Industrial-Driven Big Data as a Self-Service Solution

I-BiDaaS aims to empower IT and non-IT big data experts to easily utilize and interact with big data technologies.

Check out Website & Social
- www.ibidaas.eu
- twitter.com/ibidaas
- linkedin.com/in/i-bidaas
- zenodo.org/communities/i-bidaas

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under grant agreement **No 780787**.

3 We made applications for **3 SECTORS**
- Financial
- Telecommunications
- Manufacturing

13 PARTNERS

10 USE CASES

11 SOFTWARE COMPONENTS | 5 Open Source 6 Proprietary

Track & Know

Track & Know - Big Data for Mobility Tracking Knowledge Extraction in Urban Areas - researches, develops and exploits a new software framework that aims at increasing the efficiency of Big Data.

Check out Website & Social
- www.trackandknowproject.eu
- TrackandKnow
- @TrackandKnow
- linkedin.com/groups/12122105

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under grant agreement **No 780754**.

3 We made applications for **3 SECTORS**
- Health Care
- Fleet management
- Motor insurance

14 PARTNERS

9 APPLICATIONS

13 SOFTWARE COMPONENTS | All Open Source

About the Speakers

Dr Toni Staykova

Dr Toni Staykova is a specialist physician and geriatrician who has been actively involved in clinical innovation projects in Cambridge and across the EU over the past 10 years. She is passionate about improving health care provision across the globe in an innovative way. Toni Staykova is the Vice President of Cambridge Medical Academy, which delivers the prestigious UKeMED international knowledge sharing program. She leads the knowledge & experience sharing activities between health care professionals to improve patient care, bridging the gaps between nations and cultures. Toni's ambition is to harness the collective intelligence of our collaboration partners to foster clinical innovation and to impact on global health.

Alon Rozen

Alon Rozen is the Dean of Ecole des Ponts Business School and Associate Professor of Innovation and Entrepreneurship. His research interest includes business modelling, circular economy, social impact and social entrepreneurship, as well as the impact that technology has on the (business) world around us. Actively developing, practising and applying his own dynamic business modelling framework and value proposition tools to help start-ups and innovation efforts at all levels.

Richard McCreadie

Richard McCreadie is a Lecturer in Information Retrieval and Data Systems at the University of Glasgow, UK. His core research topics include real-time information retrieval, machine learning, big data stream processing and evaluation methodologies over streaming data.

How are we contributing to the European Big Data Ecosystem?

Alon Rozen

I-BiDaaS as an actor and enabler of an EU big data ecosystem: The I-BiDaaS project is in itself contributing to the ecosystem as it is both an enabler, a component, and a catalyst while looking for a solution to enlarge the ecosystem. I-BiDaaS has 13 different partners from small, medium and large companies, multinationals, several institutions of higher learning, different kind of forms of academia, universities, business schools from eight different countries and even more nationalities. This ecosystem is part of a broader ecosystem, like the three projects that collaborated and organized this session, is already another part of the ecosystem. It's also helping to bring the ecosystem closer together. Even further, it's a catalyst for developing the ecosystem as what we're trying to do is bring in more companies and organizations into this EU data ecosystem by doing the capacity building. It feels like we're both acting and enabling, and we're trying to do this with an agile lean, and a bottom-up approach, which is more European in style than anything else.

I-BiDaaS as an Actor and enabler of EU big data innovations: In terms of enabling big data innovations, which is a key outcome of the I-BiDaaS project, we have developed 11 software components, five of which are open source. We have developed 10 European use cases across three different sectors - finance, telecom and manufacturing. Five innovations that were deemed mature by the EU Innovation Radar. Creating this much innovation despite having so many partners, time zones, nationalities, and accents in English to manage, required a lean innovation approach. I'm proud to say that we adopted a rapid prototyping approach and framework, which is called dynamic business modelling. During our workshops, we collapsed customer product and business model development into a one rapid prototyping phase. It's different from the usual business model canvases and helped us to ideate multiple innovations rapidly. It allowed us to use lean and agile principles to identify which were the most promising. The fact that five were identified as mature was a testimony to the ability of the consortium to act in a lean innovation manner.

One exciting challenge to helping an EU Big Data ecosystem emerge came up in the project - how do you get around restrictions in data sharing and data availability? We have partners from the banking sector, so compliance issues, KYC issues and GDPR, so we initially only used synthetic data. We quickly discovered that this had too many practical limitations. The solution that the consortium found early in the project that removed this significant barrier was to move from synthetic to either encrypted or anonymized data, which allowed everyone to share data effectively and go for example from 1-month delay to a couple of days. And it let everyone to go from sharing only synthetic data to be able to share real anonymized (or encrypted) data. This is a best practice we could apply at the EU level as we advance.

I-BiDaaS as an Actor and enabler of democratization of Big Data solutions to EU SMEs: Two innovations were deemed business ready by the EU Innovation Radar, and one as market-ready, which is the Holy grail in the EU innovation radar system. The key to ecosystem expansion is moving data beyond the scope of Fortune 1000 companies and giving much broader access to SMEs. MuSES developed by Qbeast, a spinoff of the Barcelona Supercomputer Center, is an excellent example of a Big Data cloud platform based on an innovative data storage technology that allows a simple, future-proof architecture and blazing-fast data analytics and thus democratizes big data in the EU. Qbeast is just an example, but several other innovations are moving in that direction too. The entire I-BiDaaS project has this democratization of big data analytics as its mission.

How are we contributing to the European Big Data Ecosystem?

Richard McCreadie

Ease discoverability of services, data and models for commercialization

Before answering how BigDataStack contributes to the European Big Data Ecosystem, I think it is important to highlight where such ecosystems are heading in the near term, to give some context for where BigDataStack and other projects fit.

In general, we are seeing movement towards the consolidation of both **data and software services into centralized marketplaces**, where they can be bought and sold. The core goal is to ease discoverability of services, data and models for commercialization, while simultaneously reducing barriers to entry for using those datasets and services. By doing so, the EU hopes to improve competitiveness of European businesses and more generally obtain the benefits of a data-driven economy.

BigDataStack fits into this wider vision as a supporting technology stack that aims to make it easier for companies to take the services, data and models obtained from such marketplaces or developed in-house and put them into production systems. To illustrate, consider a company, say a grocery retailer, who is opening an online store-front. The business developer decides that they should provide product recommendations for their customers, based on their shopping history. But how might this be achieved?

For companies like this there is both a **technology and knowledge gap** that needs to be overcome, which is not trivial especially for smaller companies that are less likely to have in-house expertise. Such a company might be able to go to a central marketplace and obtain large grocery recommendation datasets and/or services that use deep learning technologies to produce high-quality recommendations. But how do they put these into production? BigDataStack defines a series of processes and supporting technological solutions that the company can use to make the realization of their desired service easier, while also guiding them into producing an efficient and maintainable solution.

BigDataStack is at its core a set of containerized services that aim to simplify or automate as much of the configuration, deployment and management of user-services as possible, focusing particularly on those requiring Big Data and Hyper-Compute capacity within private clusters or public clouds. While possible to deploy as a single holistic stack or product, BigDataStack is comprised of a range of modular specialized components targeted at particular components of the process. A good way to break these components down is based on what process they target: pre-deployment, initial deployment or post-deployment.

Pre-deployment, business analysts want to map their business goals into services, with the help of application engineers. BigDataStack provides three main services here, process modelling (for defining the business logic), process mapping (that converts this to a service list) and the data toolkit (configures each service). For deployment, BigDataStack builds and extends the Openshift platform, as well as providing an additional management layer (the Realization Engine) to make it easier to manage complex applications. Finally, post deployment, BigDataStack has been developing new automated application management tooling, with the triple monitoring framework, quality of service evaluator, and dynamic orchestrator. These enable automated decision making and alteration of the application deployment at run-time to account for changing environmental factors (like traffic load). These form the core of the BigDataStack offering.

BigDataStack, contributing to open source, supporting open innovation

However, research and development from BigDataStack also supports innovations used outside of BigDataStack itself. For instance, **BigDataStack, already has made open source contributions** to Openshift itself, enabling data storage to be shared across virtualized environments. Meanwhile, underlying data-skipping technologies developed within BigDataStack have been integrated into IBM's object storage service. It is also worth noting that BigDataStack's use-cases have also resulted in some valuable technologies to the European software ecosystem. For instance, as part of development for the grocery-recommendation use-case, a new product recommendation platform has been released as open source, that provides out-of-the-box support for a wide range of state of the art deep learning models for recommending items to users.

How can we apply Big Data in businesses? What barriers did you encounter with the industries in your project?

Dr Toni Staykova

Distinguishing between barriers to implementation and barriers to exploitation

Firstly, it is necessary to distinguish between barriers to implementation and barriers to exploitation. In the Track and Know Healthcare business case the focus has been on gathering novel intelligence regarding our service operations across a geographical area, in addition the goal was to use this intelligence alongside the simulation tools to make decisions regarding service positioning and optimization.

The first barriers to implementation that was encountered is the **gap in understanding between technical and clinical partners**. The clinical partners required some time in order to process and understand the project terminology from the big data technical domain and it's application to the business domain.

Technical project partners are very knowledgeable about the State of the Art with regards to ICT however, healthcare ICT is generally less advanced and faces various limitations such as data protection ethical regulations, patient safety concerns, etc. As well as being limited by management decisions for example purchasing decisions for tools.

This was evident at workshop events but through dialogue between healthcare professionals and technical experts through demonstration and discussion strengthened the Healthcare business case and the potential impact.

Requirement to anonymise healthcare data

Another important barrier typically encountered in research involving healthcare is the requirement to anonymise healthcare data. In the Track and Know project early recognition and involvement of the right decision makers enabled access to the correct data without delays and indeed through this engagement Track and Know was able to obtain additional data, which pertained to the second business case relating to driving behavior. This data is being made available to external researchers.

The barrier of the COVID-19 crisis

The third barrier to implementation was the challenge arising from the COVID-19 crisis. This impacted the availability of healthcare professionals to participate in the project over a period of time. It also necessitated the modification of the pilot. This was overcome partly with a shift of pilot validation to available services.

Bridging gaps through communication and collaboration

To summarise the key message gained from experience in conducting the healthcare pilot within Track and Know is that the communication gap between the different domains is the most difficult part to bridge, but through working together this can be reduced and can lead to further and continued implementation of big data technologies once understanding has been achieved.

How can we apply Big Data in businesses? What barriers did you encounter with the industries in your project?

Richard McCreadie

Containers, the future in a box

One of the most important enabling technologies that has made it possible for applications to be shared, bought and sold is the wide-scale adoption of containerized services. As a quick introduction for those unfamiliar, a container is a mechanism via which user-application code along with an associated environment can be packaged together into a single unit. Once a container has been built, it can be run on almost any platform (subject to available resources and if configured correctly). This allows developers to construct reusable application components, as well as enabling the marketplaces I mentioned earlier. The analyst firm Gartner predicts that by 2023, 70% of organizations will be running three or more containerized applications in production.

From our observations from both BigDataStack, and also from the new Flagship Infinitech project which is using containers as the standard for next-generation financial services, containers are the future of big data compute. While there is some initial training needed to get companies up-to-speed with the technology, our end-users were impressed by the ease of use and portability of their applications once containerized. Additionally, containers provide a standardized mechanism for both data services and user application scaling on distributed infrastructures, enabling big data processing at scale to even small companies.

Furthermore, there are a wide range of future benefits that are only just starting to be realized. From a business perspective, as container service marketplaces become more common and widely used, it will become much easier for businesses to access state-of-the-art solutions and deploy them in production using platforms like BigDataStack. Meanwhile, as containerized solutions increasingly implement common standards (e.g. as Infinitech is defining for the financial domain), run-time orchestration as well as roll-out of incremental improvements (such as improved effectiveness or security features) will be automatable, dramatically reducing the cost of maintenance.

Rise of Deep Learning (AI)

The second main new disruptive technology that emerged during BigDataStack and is having a strong impact on Big Data applications is Deep Learning, colloquially known as AI. While there is excessive hype around 'AI' applications, deep learning when paired with Big Data is resulting in better systems than were thought possible only a few years ago. Deep learning is at its core simply a more effective means for a computer to identify and record patterns found when analysing very large datasets. For example, the current state-of-the-art deep learning model for natural language processing tasks (e.g. language translation or question answering) known as GPT-3, was trained on a filtered crawl of the entire Web, in addition to various book corpora and Wikipedia, comprising half a trillion words. This is the next frontier for Big Data, as training a single model would take 350 years on a single machine, and cost in excess of 4.5 million U.S. dollars. These types of models have only recently become feasible due to significant advances in cheap GPU compute.

These types of big data application require massive amounts of both data throughput and hyper-compute, and hence are a natural use-case for BigDataStack and containerized infrastructures in general. However, we encountered significant barriers to enabling such applications in BigDataStack. In particular, container-support for GPUs is not currently standardized, and is dependent on custom driver support by individual GPU providers. Furthermore, the space is currently dominated by a single hardware provider (Nvidia), but with a competitor only just emerging (AMD), each using different APIs, leading to software fragmentation. Over the next couple of years I expect the current issues to be resolved, which will enable companies both large and small easier and more efficient access to the new wave of deep learning technologies.

Now What?

Alon Rozen

A little bit of background and context is necessary. I-BiDaaS solution offers three different modes, reflecting the different types of knowledge that a user has.

1. **Self-Service mode** related to users that have some data science and statistic knowledge so they can select a pre-defined algorithm from a pool of algorithms for a given domain problem.

2. **Expert mode** for the big data developers where they can upload their code to the platform.

3. **Co-develop mode** for in-house industrial employees that will develop their application with the help of I-BiDaaS team.

Project-level: one of the key outcomes, in my opinion, is something we are calling the I-BiDaaS Cookbook/ User Guide, which explains how to do big data in both expert and self-service mode, which means that any organization with an IT team in-house could quickly and economically start running I-BiDaaS big data solution effectively. I think this will spread quite a bit across the EU SME landscape and even beyond that, even the larger organizations that don't have the in-house skills that are presently using consultants for that.

Consortium-level: One exciting outcome is that at the end of the project (end of 2020), the consortium will continue to live and work together as an agile ecosystem. The idea is to also share the cookbook with small to big organizations to give companies an idea of what is possible. Then depending on the use case, different members of the consortium will come together to support the customer and tailor the solution to their needs.
The idea is to take the innovations to the market, help them mature, climb the experience and learning curve

by training and upskilling in-house talent. Thus, they can use the cookbook for the I-BiDaaS Self-Service Solution without becoming addicted to consulting or outside IT partners which is more expensive or hard to find talented people in the market with the necessary skills. The approach is both Agile in action and the EU big data ecosystem in practice.

Geopolitics: The last level, which I think is very important, is the geopolitical level. What role is Europe going to be playing in this very key strategic sector in the years to come? I-BiDaaS consortium is building a bottom-up emergent big data ecosystem, a fractal of the bigger EU ecosystem.

If we look at the three projects today (I-BiDaaS, BigDataStack, Track&Know), it is an extra fractal of that as well. Even if the politicians can't quite unite on the subject, it is clear that we are moving to a polarized tech world in which the US and China will try to establish their supremacy. This could be the premises of a truly European approach to establishing things not by dictating but by collaborating on the ground.

For me, I-BiDaaS is a potential future. We have a unique EU approach to establish a robust ecosystem by doing something that is European in its DNA. Showing how formal and informal collaboration across sectors and countries can create powerful synergies with widely distributed benefits instead of narrowly held profits, which seems to be the model until now. It is a value economy approach rather than a winner-take-all approach. I-BiDaaS made me hopeful that this is a possibility, not a utopian dream, but it is actually something that is already happening at this scale.

Now What?

Dr Toni Staykova

Track & Know aimed to develop the tools for commercialisation and were focused on looking for additional funding and related social services that would use the same tool until critical mass could be achieved. The COVID-19 crisis impacted the UK[1] National Health Service's (NHS) ability to deliver services and which has resulted in a severe backlog of patient appointments. There are few tools to help NHS managers decide how to clear the backlog, or indeed what actions could be taken for example, additional locations of home visits, instead of clinic visits, to reduce these delays.

Towards the end of the Track & Know project (8 Months to project end) Innovate UK launched a funding call for COVID solutions that could address problems related to the crisis. A core group of project partners were able to prepare and submit an application to develop business intelligence and decision support tools extending the Track & Know tools for patient referral allocation specifically addressing the COVID crisis needs.

This application was successful and the team were able to take this initiative forward. The key learning from this activity is to engage in early planning of exploitation and definition of partners and IP and to be aware and actively seek funding opportunities in the original and related domains. It is important to carry out ethics and data protection assessments which could otherwise cause delays and lastly it is very important to promote your results to external stakeholders

1. UKRI – UK Research and Innovation

Recordings of the session can be viewed here:
https://youtu.be/gMDpZtKzQQ4

Slides of the session can be reused here:
https://zenodo.org/record/4312914#.X9DTqbPSLV8

Stay tuned

About the Projects

BigDataStack - Holistic Stack for Big Data Applications and Operations

BigDataStack delivers a complete high-performance stack of technologies addressing the emerging needs of data operations and applications. BigDataStack promotes automation and quality and ensures that the provided data are meaningful, of value, and fit-for-purpose through its Data as a Service offering that addresses the complete data path with approaches for data cleaning, modelling, semantic interoperability, and distributed storage. To enable data operations and data-intensive applications to fully exploit the sustainability of BigDataStack and take full advantage of the developed technologies, the consortium has brought on board three use cases that will exhibit their applicability through Smart Insurance, Connected Consumer and Real-Time Shipping.

I-BiDaaS - Industrial-Driven Big Data as a Self-Service Solution

I-BiDaaS aims to empower users to easily utilize and interact with big data technologies, by designing, building, and demonstrating, a unified framework that: significantly increases the speed of data analysis while coping with the rate of data asset growth, and facilitates cross-domain data-flow towards a thriving data-driven EU economy. I-BiDaaS is being tangibly validated by real-word, industry-lead experiments in the domains of banking, manufacturing, and telecommunication.

Track & Know – Big Data for Mobility Tracking Knowledge Extraction in Urban Areas

The project has developed a software framework and demonstrated increased efficiency of Big Data applications. The project has developed a scalable, fault tolerant platform to manage big data coming from a variety of sources and has created efficient, interoperable and scalable toolboxes integrated into a Track and Know software platform. The 'Big Data Processing', 'Big Data Analytics', 'Complex Event Recognition' and 'Visual Analytics' toolboxes were successfully applied to mobility data from three domains area, fleet management, motor insurance and health sectors and delivered insights and efficiency improvements in the target domains.

www.bigdatastack.eu

www.ibidaas.eu

www.trackandknowproject.eu