Big Data Applications

Valentina Janev

Institute Mihajlo Pupin

LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS

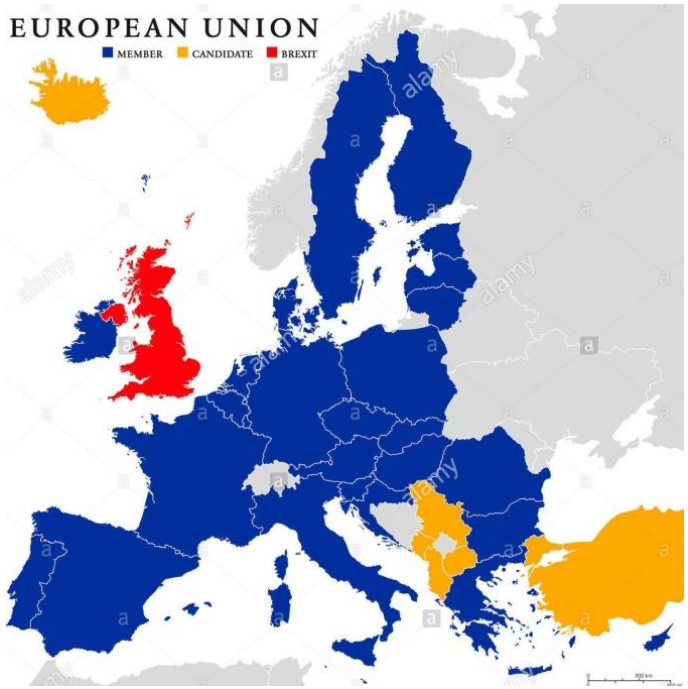
This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 809965.

Overview

- Objectives & Motivation
  - Big Data Ecosystem
  - Big Data Analytics
- Survey of Applications
- Challenges & Research Trends
- Conclusions

Vision and Primary Objectives

Strengthening the Human capital and Education, Research and Development capacities of "Mihajlo Pupin" Institute the leading Serbian R&D institution in information and communication technologies in order to serve as a Big Data & Analytics HUB that connects and integrates scientists and professionals from the West Balkans and the entire region into the European Research Area.



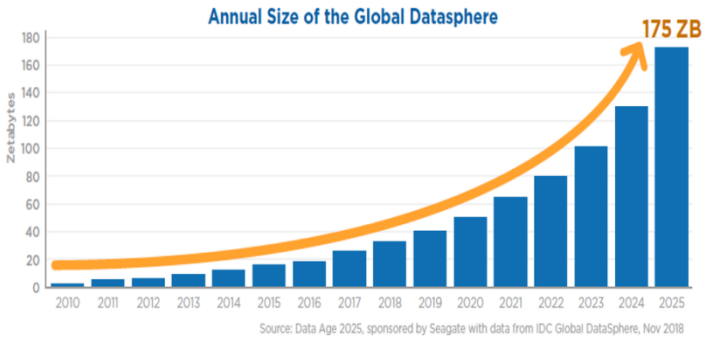
Decreasing the existing European regional R&I disparity by Fostering excellence in the Big Data Ecosystem areas unlocking and raising the scientific profile of academic institutions from Serbia and the region while contributing to European progress beyond the state-of-the-art of related research and technology, as well as establishing productive and fruitful long-term cooperation.

Motivation

- Demystify the term Big Data Ecosystem
  - "Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.
  - How big a dataset needs to be in order to be considered big data ?
  - Big data to Amazon or Google is quite different from big data to a medium-sized insurance or telecommunications organization

Big Data Statistics

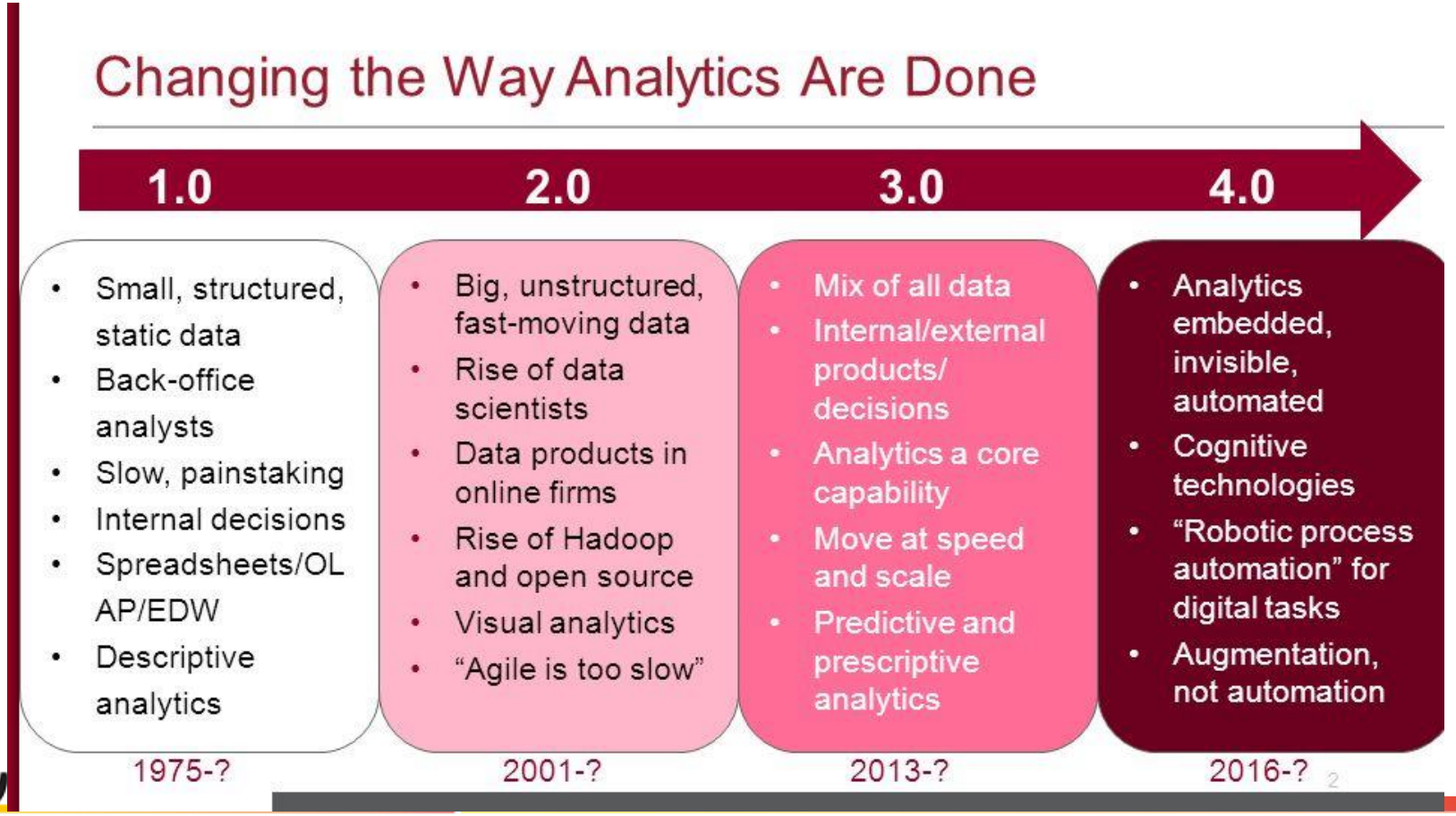
- In 2010, Big data size was standing at 1.2 zettabytes **(IDC)**
- In 2019, there are 2.3 billion active Facebook users, and they generate a lot of data **(Data Never Sleeps**
- 90% of all data has been created in the last two years **(IBM)**
- Twitter users send over half a million tweets every minute **(Internet Live Stats, Domo)**
- 40 zettabytes of data will be created in 2020 (40) **(EMC)**
- 175 zettabytes by 2025 **(IDC)**



Motivation

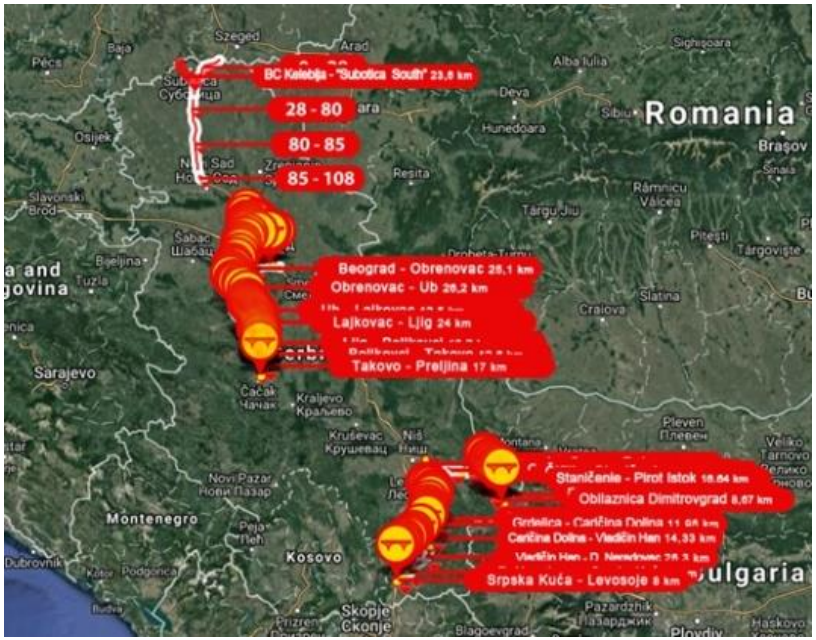
- Demystify the term Big Data Analytics

Four Eras of Analytics Thomas H. Davenport (2016)

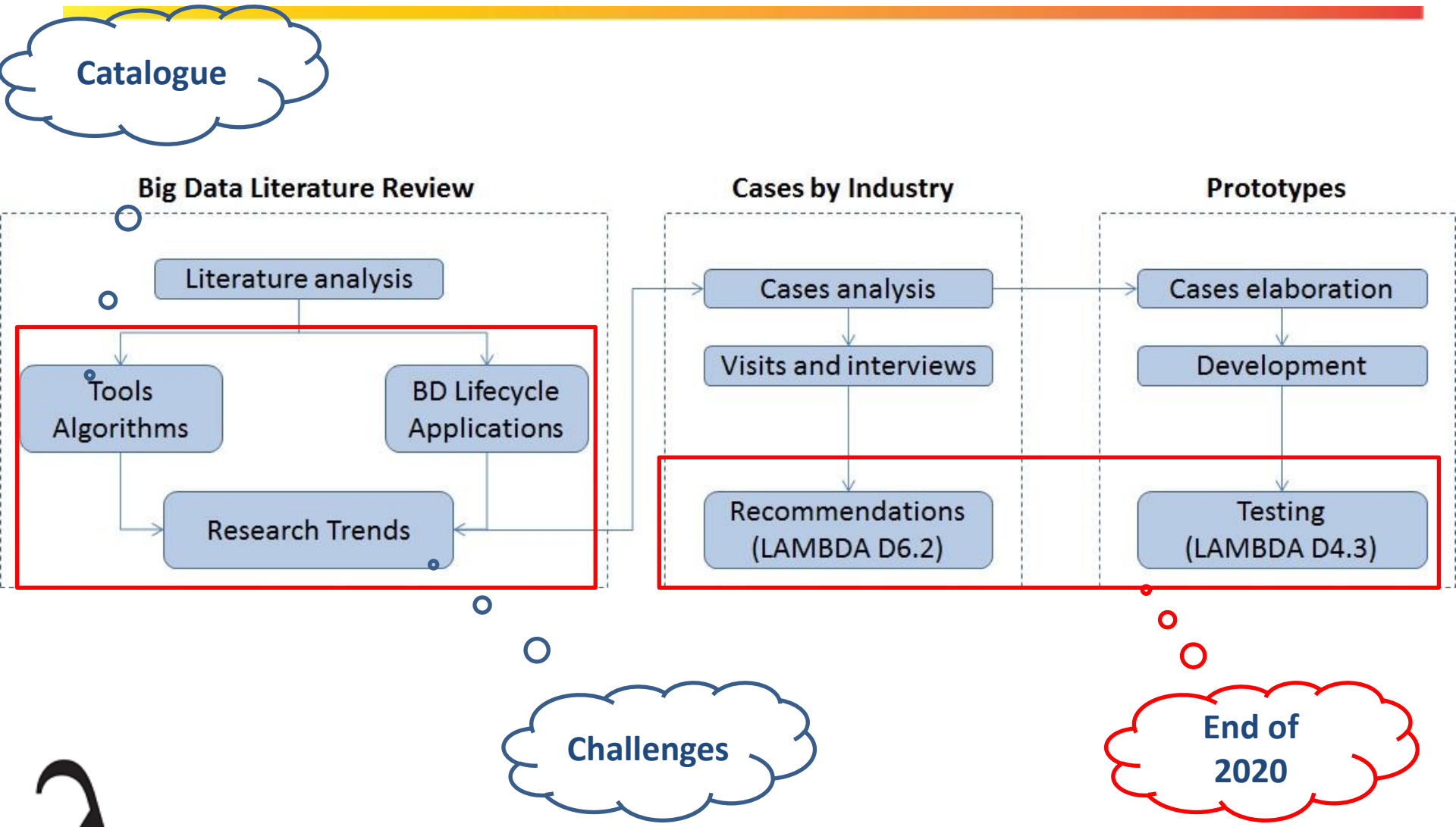


Motivation

- Survey - Identify future trends in Big Data Analytics, Tools and Technologies
- Analysis of applicability & Challenges for Exploiting the Potential of Big Data



Research Methodology



Survey of Big Data Applications

LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS

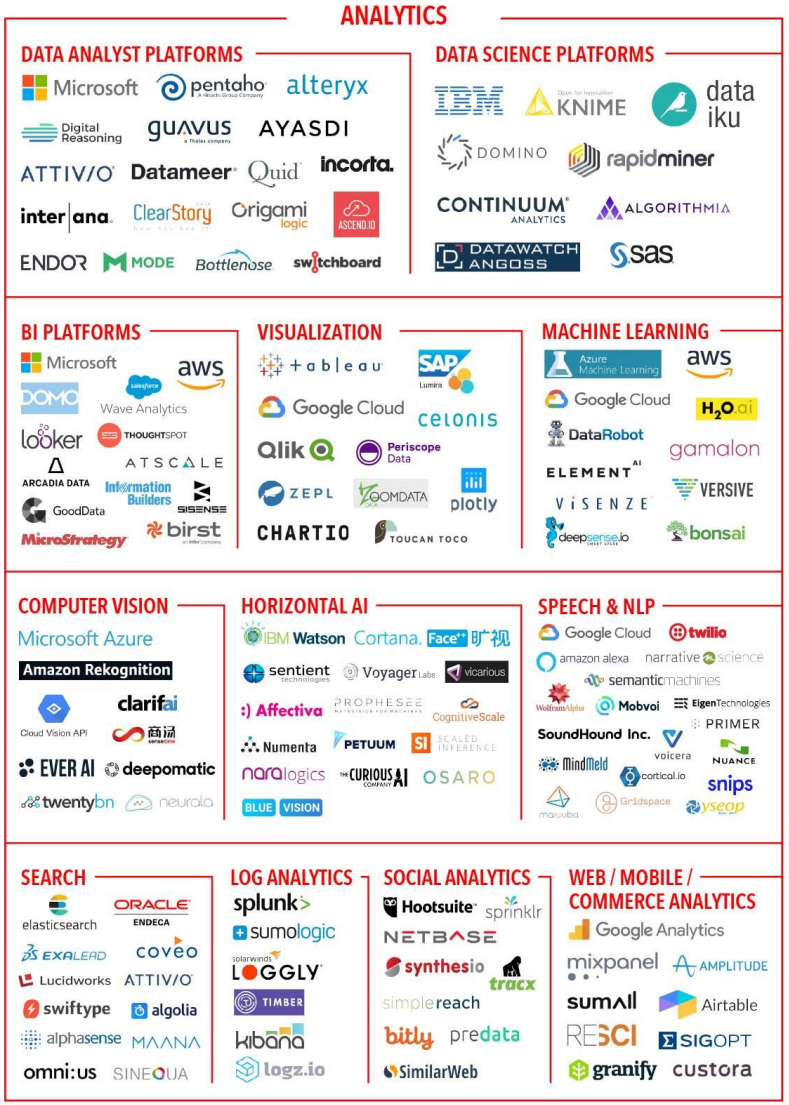
This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 809965.

Big Data Analytics

Literature Review – Main Sources

- **Elsevier ScienceDirect** is a website which provides subscription-based access to a large database of scientic and medical research. It hosts over 12 million pieces of content from 3,500 academic journals and 34,000 e-books.
- **SpringerLink** is the world's most comprehensive online collection of scientic, technological and medical journals, books and reference works printed from Springer-Verlag.

- **Additional Sources**
  - Transforming Data with Intelligence (TDWI)
  - Big Data - Forrester
  - …. Learning Big Data Tools…



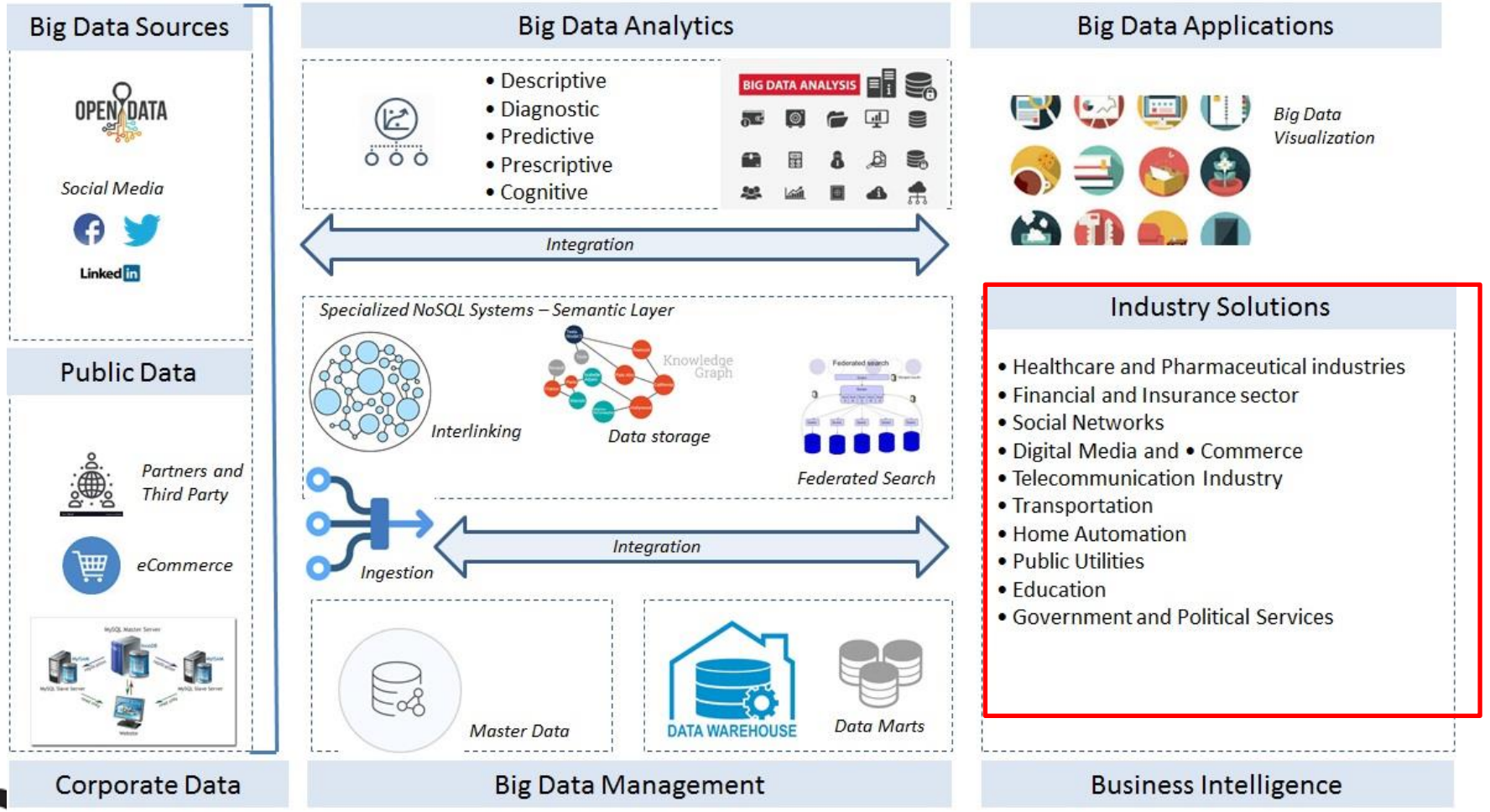
Big Data Ecosystem

- The term **Ecosystem** is defined in scientific literature as a complex network or interconnected systems.
- While in the past corporations used to deal with static, centrally stored data collected from various sources, with the birth of cloud services, cloud computing is rapidly overtaking the traditional in-house system as a reliable, scalable and cost-effective IT solution.
- Thus, large datasets – log files, social media sentiments, click-streams – are no longer expected to reside within a central server or within a fixed place in the cloud.
- To handle the copious amounts of data, **advanced analytical tools are needed which can process and store billions of bytes of real-time data, with hundreds of thousands of transactions per second**.

Examples of Big Data Ecosystems

Company	Year	Infrastructure
Facebook	2017	has more than **two billion users on millions of servers**, running thousands of configuration changes every day involving trillions of configuration checks. [310]
LinkedIn	2017	..**467 million members worldwide** (in 2017), … …**100,000 servers spread across multiple data centers**. [215]
Alibaba	2017	Hosts 402,000 web-facing computers from China-allocated IP addresses **the second largest hosting company in the world today**. [321]
Google	2016	Gartner estimated in a July 2016 report that Google at the time had **2.5 million servers…**Google data centers process an average of 40 million searches per second, resulting in 3.5 billion searches per day and 1.2 trillion searches per year, Internet Live Stats reports. [390]
Amazon	2014	... an estimate of 87 AWS datacenters in total and a range of somewhere between **2.8 and 5.6 million servers in Amazon's cloud (2014)**. [301]

From Data to Applications (Enterprise)



Deferent Types of Data Analytics

- **Descriptive analytics** focus on analyzing historic data for the purpose of identifying patterns (*hindsights*) or trends.

- **Diagnostic analytics** [364] discloses the root causes of a problem and gives *insight*. The methods are treated as an extension to descriptive analytics that provide an explanation to the question "Why did it happen?"

- **Predictive analytics**-based services apply forecasting and statistical modeling to give insight into "what is likely to happen" in the future (*foresight*) based on supervised, unsupervised, and semi-supervised learning models.

Deferent Types of Data Analytics

- **Prescriptive analytics**-based services answers the question “What should I do?”. ..software tools utilize artificial intelligence, optimization algorithms and expert systems approaches.

- **Cognitive analytics** is a term introduced recently in the context of cognitive computing (see also *Deloitte Tech Trends 2019*). The goal is to develop ”AI-based services that are able to interact with humans like a fellow human, interpret the contextual meaning, analyze the past record of the user and draw deductions based on that interactive session” [174], [176].

Big Data Analytics - Literature Review

Keyword based query on term Big Data Analytics returns (April 2020):

- **180,736 results in ScienceDirect** (or 3 percent more than in December 2019, 174,470 results), 10,042 of them review articles, where the oldest 2 papers are from 1989 and discuss the use of supercomputers in atmospheric science, astronomy, materials science, molecular biology, aerodynamics, ..
- **40,317 results in SpringerLink** (or 7 percent more than in December 2019, 33,249 results), where the oldest publications dating from 1950s are related to mathematics.

Big Data Analytics

- Big Data Analytics (BDA) is a broad topic that, depending on the objectives of the research, can be linked on the one hand to data science and machine learning, and on the other to data / software engineering and cloud computing.

Table 1: Number of Review articles in ScienceDirect database

Keywords	**1995-1999**	**2000-2005**	**2006-2009**	**2010-2015**	**2016-2020**	**Total**
BDA	388	718	1349	2190	4,605	10,042
BDA and BI	12	15	45	80	437	615
BDA and BI and NoSQL				3	31	35
BDA and Apps and NoSQL				8	46	54

Business Intelligence (BI)

Journals

Table 2: Journals that match the search criteria

'Big Data' and 'Application' (128,033)	Neurocomputing, Journal of Cleaner Production, Procedia Computer Science, IFAC Proceedings Volumes, Expert Systems with Applications, Physica A: Statistical Mechanics and its Applications, Sensors and Actuators B: Chemical, Journal of Chromatography A, Nuclear Physics B, European Journal of Operational Research
'Big Data' and 'Industry' (59,734)	Journal of Cleaner Production, Future Generation Computer Systems, Energy Policy, Journal of Membrane Science, Expert Systems with Applications, Procedia Computer Science, Journal of Banking and Finance, Research Policy, European Journal of Operational Research
'Big Data Analytics' and 'Applications' (41,031)	Journal of Cleaner Production, Future Generation Computer Systems, Neurocomputing, Journal of Chromatography A, IFAC Proceedings Volumes, Physica A: Statistical Mechanics and its Applications, Sensors and Actuators B: Chemical, Analytica Chimica Acta, Journal of Membrane Science, Nuclear Physics B
'Big Data Analytics' and 'Business Intelligence' (3,539)	Future Generation Computer Systems, Procedia Computer Science, Technological Forecasting and Social Change, Expert Systems with Applications, Decision Support Systems, IFAC Proceedings Volumes, Accounting, Organizations and Society

Applications

In Scope

- eGovernment
- Healthcare and Pharma
- Transportation and Smart Cities
- Energy Production and Smart Grids
- Energy Consumption and Home Automation
- Banking and Insurance
- Social Networks and e-Commerce
- Environment Monitoring
- Natural Disasters, Safety and Security
- Telecommunications
- Manufacturing

Out of Scope

- Nuclear Physics and Astrophysics, Materials Science, Construction and Architecture, Chemistry and Chromatography

eGovernment

- The Data Strategy and the White Paper on Artificial Intelligence are the first pillars of the new digital strategy of the Commission.
- **Drivers of change toward a data-driven society**
  - Digitization has massively increased the quantity of management information available, the resolution and frequency at which it is captured, and the speed at which it can be processed.
  - Connectivity has led to network effects, which enable the integration and sharing of data. [sectoral data spaces]
  - The application of intelligence on top of data and networks.

  Big data shows great promise in public services to personalized e-government service delivery.

Transportation and Smart Cities

- Smart transportation is one of the key big data vertical applications, and refers to the integrated application of modern technologies and management strategies in transportation systems.
- Nowadays, an increasing number of cities around the world struggle with traffic congestion, optimizing public transport, planning parking spaces, and planning cycling routes.
- These issues call for new approaches for studying human mobility by exploiting machine learning techniques [406], forecasting models or through the application of complex event processing tools [135].

Transportation and Smart Cities (Examples)

- **Highways and motorways control systems** generate a high volume of data that is relevant for a number of stakeholders from traffic and environmental departments to transport providers, citizens and the police.
  - Interoperability of tolling services on the entire European Union road network because the ones introduced at local and national levels from the early 1990s onwards are still generically non-interoperable;
- **Self-driving cars** rely on vast amounts of data that are constantly being provided by its users and used for training the algorithms governing the vehicle in auto-pilot mode.
- Big data processing in the transportation domain could even be used to **govern traffic light scheduling**, which would have a significant impact on this sector, at least until all vehicles become autonomous and traffic lights are no longer required.

Transportation and Smart Cities Scenarios

- **Predicting and preventing road traffic congestion** analytics is used to improve congestion diagnosis and to enable traffic managers to proactively manage traffic and to organize the activities at toll collection stations before congestion is reached.
- **Strategic environmental impact assessment** analytics is used to study the environmental impact and the effect of highways on adjacent flora, fauna, air, soil, water, humans, landscape, cultural heritage, etc. based on historical and real-time analysis.
- **Passive pollution monitoring** involves collecting data about the diffusion of air pollutants, e.g. emission estimates based on traffic counting. Passive pollution monitoring has been used to determine trends in long-term pollution levels. Road traffic pollution monitoring and visualization requires the integration of high volumes of (historical) traffic data with other parameters such as vehicle emission factors, background pollution data, meteorology data, and road topography.

Energy Management - Smart Grids

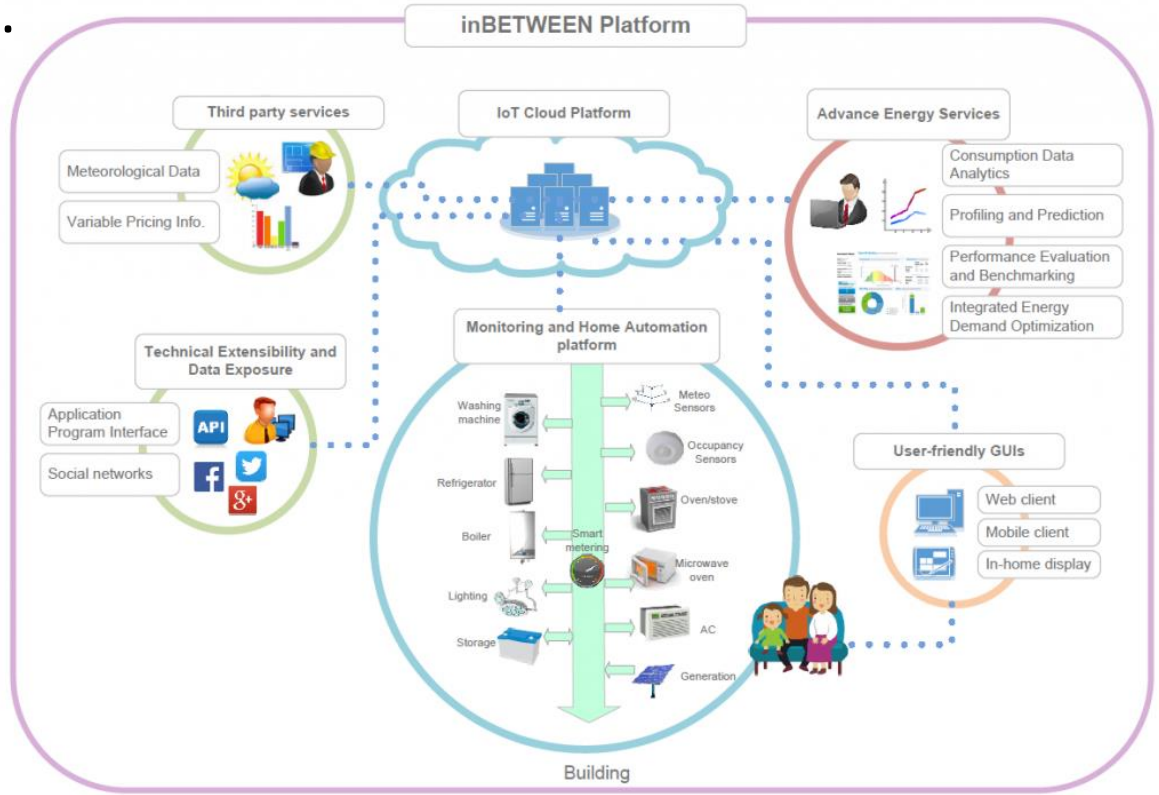
- The smart grid (SG) is the next-generation power grid, which uses two-way flows of electricity and information to create a widely distributed automated energy delivery network [155]. The goal is to optimize the generation, distribution and consumption of electricity. In general, there are three main areas where data analytics have been applied:
  - Ensuring smart grid stability, load forecast and prediction of energy demand for planning and managing energy network resources;
  - Improving malfunction diagnosis, either on the production side (in plant facilities) or health state estimation, and identifying locations and forecasting future line outages in order to decrease the outage costs and improve system reliability;
  - Profiling user behaviours to adjust individual consumption patterns and to design policies for specific users.

Energy Management - Smart Grids

- Long list of various open issues and challenges in the future for smart grids such as
  - lack of comprehensive and general standard, specifically concentrated on big data management in SGs;
  - interoperability of smart devices dealing with massive data used in the SGs;
  - the constraint to work with approximate analytics and data uncertainty due to the increasing size of datasets and real-time necessity of processing [354];
  - security and privacy issues and the balance between easier data processing and data access control for big data analytics, etc.

Energy Consumption and Home Automation

- The Internet of Things plays a crucial role in home automation solutions that based on this data are capable of processing and providing accurate predictions, and energy saving recommendations.



Energy Consumption and Home Automation

- Challenges:
  - Provide optimal device scheduling to maximize comfort and minimize costs.
  - Planning and offering possible home adjustments or suggesting investments in renewable sources.
  - Predictive maintenance and automatic fault detection from sensor data for both basic household appliances and larger mechanical systems like cars, motors, generators, etc.

Social Networks and e-Commerce

- Social networks provide a source of personalized big data suitable for data mining with several hundreds of thousands of new posts being published every minute.
- The wide variety of on-line shopping websites also presents a continuous source of huge volumes of data that can be stored, processed, analyzed and inferred to create recommendation engines with predictive analytics.

- Challenges:
  - Optimize product presence in the media fed to the user.
  - Examining user behaviour patterns and tendencies allows for offer categorization in the best possible way so that the right offer is presented precisely when it needs to be, thus maximizing sale conversions
  - Development of new architectures in the big data domain.

e-Commerce

Amazon Product Graph – 100 K product types, 5K attributes

- Amazon uses knowledge graphs to represent the hierarchical relationships between product types on amazon.com; the relationships between creators and content on Amazon Music and Prime Video; and general information for Alexa's question-answering service

10 Challenges in Building KGs

1. We do not know what we really want to model ?
2. Too little structured data
3. Heterogeneous data
4. How to extract the data ?
5. Non-trustworthy data
6. Scalability of solution (100K types require 100K models)
7. Train the ML models with limited data
8. Understand user needs
9. How to get insights from the facts
10. Can we explain the recommendations ?

Environment Monitoring

- Advances in remote sensing using satellite and radar technologies have created new possibilities in oceanography, meteorology, forestry, agriculture and construction (urban planning).
- Satellite-based measurement systems
  - The most valuable source of data from this category is the Landsat, a joint satellite program of the USGS and NASA, that has been observing the Earth continuously from 1972 through to the present day. More than 8 million images [207] are available via the NASA website and Google Earth Engine Data Catalog. Additionally, the Earth observation mission from the EU Copernicus Programme produces 12 terabytes of daily observations (optical imagery at high spatial resolution over land and coastal waters) each day that can be freely accessed and analysed with DIAS, or Data and Information Access Services.

Environment Monitoring (Examples)

Smart farming.	Big data research in Smart Farming is still in an early development stage. Challenges foreseen are related both to technical and organizational issues. Technical challenges include the automation of the data acquisition process, the availability and quality of the data, and the semantic integration of these data from a diversity of sources (information on planting, spraying, materials, yields, in-season imagery, soil types, weather, and other practices). Although, from a business perspective, farmers are seeking ways to improve profitability and efficiency, there are challenges related to the governance (incl. data ownership, privacy, security) and business models for integration of the farms in the entire food supply chain [469].
Rainforest monitoring.	The contribution of the world's rainforests to to the reduction of the impact of climate change is well-known to environment scientists, therefore projects have been started to integrate various low-cost sensors for measuring parameters such as humidity, temperature, total solar radiation (TSR), and photosynthetically active radiation (PAR) [68].
Biodiversity planning.	- Machine learning and statistical algorithms have proved to be useful for the prediction of several numeric target attributes simultaneously, for instance, to help natural resource managers to assess vegetation condition and plan biodiversity conservation [249].

Natural Disasters, Safety and Security

- Advancements in the field of IoT, machine learning, big data, remote sensing, mobile applications can improve the effectiveness of disaster management strategies and facilitate implementation of evacuation processes. The requirements faced by ICT developers are similar to those in the other domains already discussed
  - the need to integrate multimodal data (images, audio, text from social sites such as Twitter and Facebook);
  - the need to syncronize the activities of many stakeholders involved in four aspects of emergency (preparedness, response, mitigation and recovery);
  - the need to install measuring devices for collecting and real-time analysis in order to understand changes (e.g. in water level, ocean waves, ground motions, etc);
  - the need to visualize information;
  - the need to communicate with people (first responders and/or affected people and track their responses and behaviour) or to alert officials to initiate rescue measures.

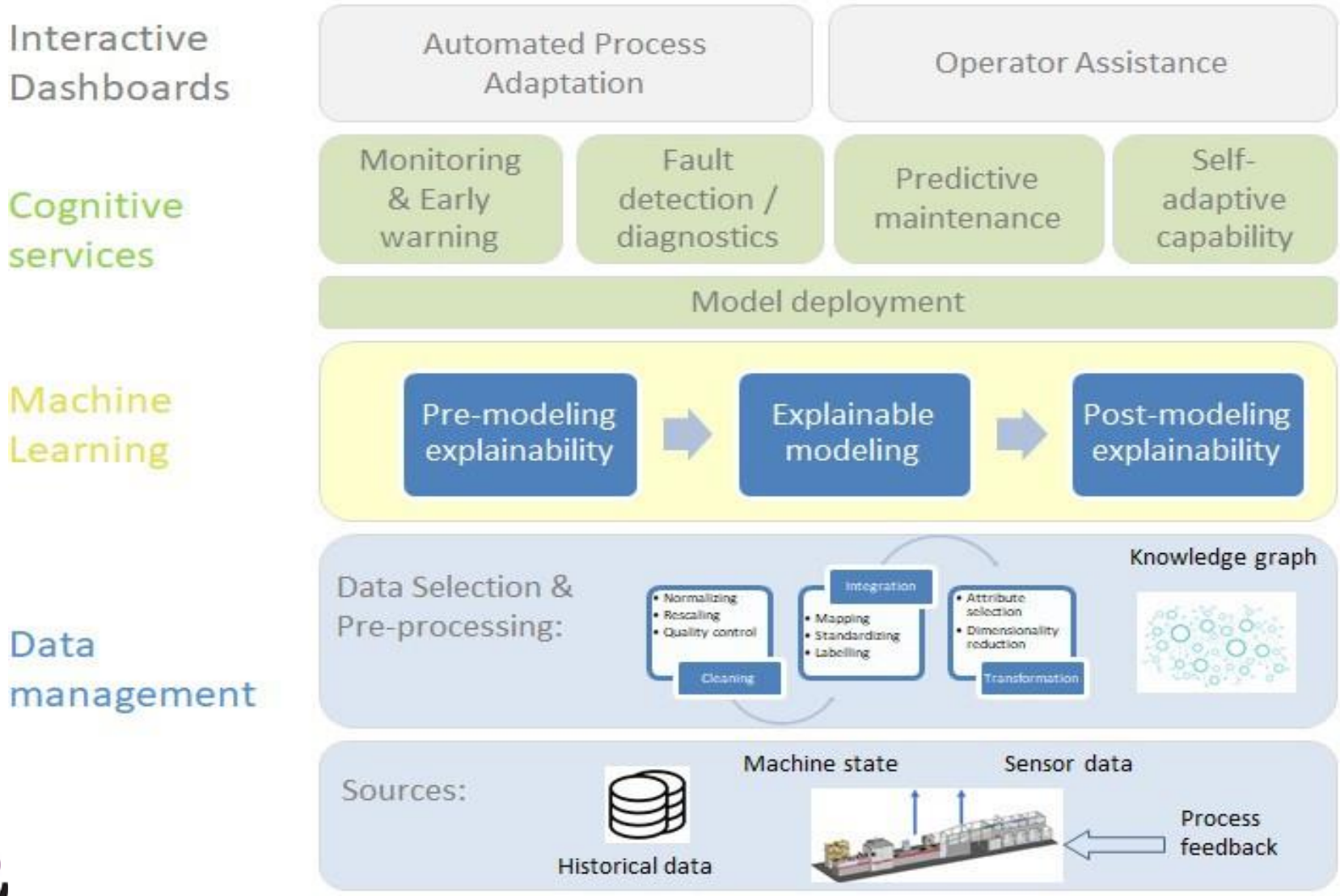
Safety of critical infrastructures

- Big data processing is especially important for protecting critical infrastructures like airports, railway/metro systems, and power grids. Large infrastructures are difficult to monitor due to their complex layout and the variety of entities that they may contain such as rooms and halls of different sizes, restricted areas, shops, etc.
- Besides processing the large amount of heterogeneous data extracted from multiple sources while considering the challenges of volume, velocity and variety, what is also challenging today is
  - real-time visualization and subsequent interaction with computational modules in order to improve understanding and speed-up decision making;
  - development of advanced semantic analytics and Machine Learning techniques for new pattern recognition that will build upon pre-defined emergency scenarios (e.g. based on rules) and generate new early warning procedures or reliable action plans.

Manufacturing

- Industry 4.0 is about automating processes, improving the efficiency of processes, and introducing edge computing in a distributed and intelligent manner.
- More complex requirements are imposed in process operations while the process frequently forfeits robustness, complicating process optimization.
- Smart manufacturing services have to operate over multiple data streams, which are usually generated by distributed sensors in al- most real-time.
- Cognitive applications make use of process data (processed on the edge) and provide high level supervisory control and support the process operators and engineers.

Manufacturing (Example)



Manufacturing (Scenarios)

- **Human-Computer Interaction.** In complex situations, operators and machines need to quickly analyze situations, communicate and cooperate with each other, coordinate emergency response efforts, and find reasonable solutions for emerging problems. In such situations, **collaborative intelligence services are needed that require fewer human-driven decisions as well as easy- to-use interfaces that accelerate information-seeking and human response.** Interpretability and explainability are crucial for achieving fair, accountable and transparent (FAT) machine learning, complying with the needs and standards of the business sector.
- **Dynamic process adaptation.** Many industrial processes are hard to adapt to changes (e.g. related to status and availability of all relevant production resources, or in case of anomaly detection). This affects product quality and can cause damage to equipment and production lines. Hence, a **semantic framework for storing contextual information and an explainable AI approach can be used for fine-tuning of process parameters to optimize environmental resources, fast reconfiguration of machines to adapt to production change, or advance fault diagnosis and recovery.**

Takeaway Tips on Research Trends & Challenges

LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 809965.

Challenges (Summary)

- **Data challenges** related to the characteristics of the data itself (e.g. data volume, variety, velocity, veracity, volatility, quality, discovery,,);
- **Process challenges** related to techniques (how to capture data, how to integrate data, how to transform data, how to select the right model for analysis and how to provide the results);
- **Management challenges** related to organizational aspects such as privacy, security, governance and ethical aspects.

Trends (Summary)

- ..there is a shift of focus **from operational data management** systems, data-warehouses and business intelligent solutions (present for instance in Finance and Insurance domain in 1990s) to **parallel and distributed computing**, as well as scalable architectures for storing and processing data in the cloud (" Analytics in Cloud").
- Emerging paradigms such as the **IoT and blockchain additionally influence cloud computing systems**.
- Wide availability of cheap processing power and vast amounts of data in recent years have enabled **impressive breakthroughs in machine learning, semantic computing, artificial neural networks and multimodal affective analytics**.

End-to end data processing pipelines

- In addition to operational database management systems (present on the market since 1970s), different NoSQL stores appeared that lack adherence to ACID principles (atomicity, consistency, isolation, durability).

- **The Data Lake concept emerged as a new storage architecture where raw data can be stored regardless of source, structure and (usually) size.**

- **TDWI –Transforming Data with Intelligence:** Hadoop technologies will soon become a common complement to (but not a replacement for) established products and practices for business intelligence (BI), data warehousing (DW), data integration (DI), and analytics.

Applications (Summary)

- Healthcare and Pharma
- Transportation and Smart Cities
- Energy Production and Smart Grids
- Energy Consumption and Home Automation
- Banking and Insurance
- Social Networks and e-Commerce
- Environment Monitoring
- Natural Disasters, Safety and Security
- Telecommunications
- Manufacturing

- **ENERGY -** …Big Data Analytics… a significant area in emerging smart grid technology, for instance, where different predictive models and optimization algorithms serve to improve end-to-end performance, end-user energy efficiency and allow increasing amounts of renewable energy sources to be embedded within the distribution networks (e.g. solar photovoltaic (PV), wind power plants).

Applications (Summary)

- **ENERGY -** Research into real-time data analytics by addressing the volume and velocity dimension of big data is a significant area in emerging smart grid technology, for instance, where different predictive models and optimization algorithms serve to improve end-to-end performance, end-user energy efficiency and allow increasing amounts of renewable energy sources to be embedded within the distribution networks (e.g. solar photovoltaic (PV), wind power plants).
- **SECURITY -** Analytics on real-time data streams combined with GIS and weather data improves detection of significant events, enhances situational awareness and helps identify hazardous road conditions (e.g. snow), which may assist drivers and emergency responders in avoiding such conditions and allow for faster emergency vehicle routing and improved response time.

Acknowledgement

- This project has received funding from the European Union's Horizon 2020 research and innovation programme, GA No 809965

- Project Partners
  - Institute Mihajlo Pupin, Serbia (Coordinator)
  - Fraunhofer Institute for Intelligent Analysis and Information Systems, Germany
  - Institute for Computer Science - University of Bonn, Germany
  - Department of Computer Science - University of Oxford, UK

Thank You for Your Attention !