Roadmap for big data and data analytics

Description and state of the art	
Definition	*Big Data* is a term for data sets with sizes and complexity beyond the ability of commonly used software tools to capture, curate, manage and process data within a tolerable elapsed time. According to Gartner's definition, Big data is high volume, high velocity, and/or high variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation[84]. That definition, which includes the 3Vs (Volume, Velocity, Variety) has been recently complemented to include also Value of data as well as Veracity, coining in this manner a 5V Big Data definition. The term often refers simply to the use of Big Data Analytics to collect, organize and analyse large sets of data to discover hidden patterns, unknown correlations and other useful information[85]. *Data Analytics* refers to the discovery, interpretation, and communication of meaningful patterns in data, based on the simultaneous application of statistics, computer programming and operations research to quantify performance. It further often favours data visualization to communicate insight. The goal of Data Analytics (big and small) is to get actionable insights resulting in smarter decisions and better business outcomes[86]. Data Analytics can be descriptive (explaining in more detail a phenomenon which is represented with data), predictive (trying to forecast the future behaviour of a system for which past and present data is available) or prescriptive (targeting the prediction of the impact of the behaviour of a system in a future scenario)). Data analytics are closely related with Big Data, as the advert of the latter propelled the rapid development of novel analytics methods, capable of handling bigger data loads and of providing more evidence-based results with less uncertainty due to the bigger data samples available.
Addressed	Societal need: Inclusive well-being and health

societal /business or public sector need	
 Existing solutions /applications /services	Several big-data platforms and infrastructure are already in use in the Healthcare sector: • **Philips HealthSuite** Digital Platform provides a cloud-based infrastructure for connected healthcare[87, 88] • **European Medical Information Framework** (EMIF)[89] • **Open PHACTS Discovery Platform**[90] • **NIH Big Data to Knowledge** (BD2K) initiative[91] • **Asthmapolis** (GPS-enabled tracker that monitors inhaler usage by asthmatics)[92] • **Ginger.io** (mobile application for patients with diabetes for example to assist with behavioral health theories)[92] • **mHealthCoach** (supports patients on chronic care medication, providing education and promoting treatment adherence through an interactive system)[92] • **RiseHealth** (customized accountable-care-organization dashboard)[92] • four hospitals which are part of the **Assistance Publique-Hôpitaux de Paris** have been using data from a variety of sources to come up with daily and hourly predictions of how many patients are expected to be at each hospital[93] • **Blue Cross Blue Shield** have started working with big data experts at **Fuzzy Logix** and have been able to identify 742 risk factors that predict with a high degree of accuracy whether someone is at risk for abusing opioids[93] • **University of Florida** made use of Google Maps and free public health data to prepare heat maps targeted at multiple issues, such as population growth and chronic diseases[93] • **Cancer Moonshot program**: Medical researchers can use large amounts of data on treatment plans and recovery rates of cancer patients in order to find trends and treatments that have the highest rates of success in the real world. For example, researchers can examine tumor samples in biobanks that are linked up with patient treatment records.[93] • **Precision medicine initiative** launched by President Obama[94] • **European Medical Information Framework** (EMIF)[95] • **Open Phacts Discovery Platform[96]** • **Flatiron Health** has developed a service called the OncologyCloud, based on the idea that 96% of potentially available data on patients with cancer is not yet analyzed. It aims to take this data gathered during diagnosis and treatment, and make it available to clinicians to further their study.[97]

	• **Comet K-Project** DEXHELPP – AT[98] • **The Shared Care Platform** – DK[98] • **E-Estonia** – National Identity Scheme – EE[98] • **AEGLE** (An analytics framework for integrated and personalized healthcare services in Europe) – UK, IT, GR, SE, BE, NL, PT, FR[98] • **The Business Intelligence database system** – GR[98] • **PASSI** (Progressi delle Aziende Sanitarie) – IT[98] • **Arno Observatory** – IT[98] • **The Swedish Big Data Analytic Network** – SE[98] • **Clinical Practice Research Datalink** (CPRD) – UK[98] • **Sentinel Stroke National Audit Programme** (SSNAP) – UK[98] • **Hospital Episode Statistics** (HES) – UK (England)[98] • **The YODA Project** (Yale University open data access) – US[98] • **FDA Adverse Event Network Analyser** - US[98] • **CEPHOS-LINK –** FI, AT, RO, NO, SI, IT[98] • **Twitter (Adverse drug reactions and public health)** – International[98] • **Flatiron** – US[98] • **UK Biobank** – UK[98] • **Semantic Data Platform for Healthcare** (SEMCARE) – DE, NL, AT, UK, ES[98] • **Integrated BioBank of Luxembourg** (IBBL) – LU[98] • **Spanish Rare Diseases Registries Research Network** (SpainRDR) – ES[98] The solutions listed below[99] are related to the platforms and initiatives listed under big data above, but have their focus on (big) data analytics in health care: • The **Hadoop Distributed File System** (HDFS): HDFS enables the underlying storage for the Hadoop cluster. It divides the data into smaller parts and distributes it across the various servers/nodes. • **MapReduce**: MapReduce provides the interface for the distribution of sub-tasks and the gathering of outputs. When tasks are executed, MapReduce tracks the processing of each server/node. • **PIG and PIG Latin** (Pig and PigLatin): Pig programming language is configured to assimilate all types of data (structured/unstructured, etc.). • **Hive**: Hive is a runtime Hadoop support architecture that leverages Structure Query Language (SQL) with the Hadoop platform. It permits SQL programmers to develop Hive Query Language (HQL) statements akin to typical SQL statements. • **Jaql**: Jaql is a functional, declarative query language designed to process large data sets. To facilitate parallel processing, Jaql converts "'high-level' queries into 'low-

	level' queries" consisting of MapReduce tasks. • **Zookeeper**: Zookeeper allows a centralized infrastructure with various services, providing synchronization across a cluster of servers. Big data analytics applications utilize these services to coordinate parallel processing across big clusters. • **HBase**: HBase is a column-oriented database management system that sits on top of HDFS. It uses a non-SQL approach. • **Cassandra**: Cassandra is also a distributed database system. It is designated as a top-level project modelled to handle big data distributed across many utility servers. It also provides reliable service with no particular point of failure and it is a NoSQL system. • **Oozie**: Oozie, an open source project, streamlines the workflow and coordination among the tasks. • **Lucene**: The Lucene project is used widely for text analytics/searches and has been incorporated into several open source projects. Its scope includes full text indexing and library search for use within a Java application. • **Avro**: Avro facilitates data serialization services. Versioning and version control are additional useful features. • **Mahout**: Mahout is yet another Apache project whose goal is to generate free applications of distributed and scalable machine learning algorithms that support big data analytics on the Hadoop platform. From industry side big data systems such as **IBM Watson, Cloudvara and Hortonworks** exist, but the industry is still in the pioneering stages.[100]
 Main actors regarding R&D of this technology	**Big data:** • Fraunhofer-Gesellschaft zur Förderung der Angewandten Forschung e.V., • Atos Spain SA, • Universidad Politecnica de Madrid, • Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, • University College London **Data analytics:** • Centre National de la Recherche Scientifique • Fraunhofer-Gesellschaft zur Förderung der Angewandten Forschung e.V. • University College London • University Of Oxford • University Of Manchester

 Current research activities	**Big data:** AutoMat, BACI, BD2Decide, BIG, BigDataEurope, BigStorage, BYTE, CIMPLEX, CoherentPaaS, DAIAD, DATA SIM, datACRON, DEDALE, EDSA, ENLIGHT-TEN, EO4wildlife, EoT, ERA-PLANET, EUDAT2020, EuDEco, EXCELL, FERARI, FREME, GROWTHCOM, iKaaS, INDIGO-DataCloud, INSIGHT, i-PROGNOSIS, L3MATRIX, LeanBigData, LinDA, MixedEmotions, MUSING, OCEANDATAMODELS, PANORAMIX, PETRA, PHEME, PROTEUS, RETHINK big, SAGE, SEE.4C, SENSATION, SoBigData, STREAMLINE, TOREADOR, TrendMiner, VaVeL, VELaSSCo, VICINITY BigDieMo (BMBF, DLR, PTKA)[98] Big data projects (BMBF):[99] News-Stream 3.0, iPRODICT, BigPro, BDSec, FEE, GeoMultiSens, HUMIT, BigGIS, AGATA, ABIDA [100] BMWi projects:[3] iTESA, PRO-OPT, SmartEnergyHub, Fast Genomics, EnOB: BigData, NetzDatenStrom, LeichtFahr EUROSTARS projects: WINDELIN [101], PBD[102], ReProsis[103]; CDTI (Spain):SISAMED[104] The British government has announced a joint project with IBM[105] Digital Agenda Germany[106] Big data competence centres (BMBF)[107] Big-Data research (BMBF)[107] **Data analytics:** There are 89 EU research projects in the area of 'data analytics'. For the public sector, the following projects might be relevant: ASGARD (analysis of raw data), DataBio (bioeconomy), PULSE (participatory urban living), BIMEDA (medial domain), NICHE (healthcare), AEGLE (healthcare), AEGIS (public safety), BYTE (societal externalities), CityPulse (smart cities) BMBF:[3] LINDA, CODA, SELFPASS, Smart Urban Services, STEPS, Wachstumskern Potenzial - iLaP – B, E! 10196 MoVieStA, EINS3D SMICE, FLORIDA, Visual Analytics for Security Applications BMWi:[3] SERVICE-FACTORY, EMuDig 4.0, MIA, PRO-OPT
 Impact assessment	**Public sector modernization:** • Sustainability • Quality of Services Provided **Public sector as an innovation driver:** • Entrepreneurship • Innovation

	• Prosperity and well-being • Quality of health • Public Safety • Transport Infrastructure

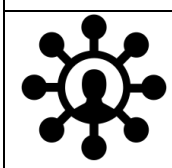
Necessary technological modifications	
 Potential use cases	The following areas in healthcare will most benefit from the application of big data technologies:[101] **Healthy living** (prevention, health promotion) • Lifestyle support • Better understanding of triggers of chronic diseases for effective early detection • Population health • Infectious diseases **Healthcare** • Precision Medicine • Collecting patient reported outcomes and total pathway costs for value based healthcare • Optimizing workflows in Healthcare • Infection prevention, prediction and control • Social-clinical care path • Patient support and involvement • Shared decision support • Home care • Clinical research More specific examples are: Organizations would be able to identify potential health issues and alert a care manager to intervene. For example, if a patient's blood pressure spikes, the system will **send an alert** in real time to a care manager who can then interact with the patient to get his blood pressure back into a healthy range.[102] Another important future use is **predictive analytics**. The use cases for predictive analytics in healthcare have been limited up to the present because we simply haven't had enough data to work with. Big data can help fill that gap.[102] One example of data that can play a role in predictive analytics is **socioeconomic data**. Socioeconomic factors influence patient health in significant ways. Socioeconomic data might show that people in a certain zip code are unlikely to have a car. There is a good chance, therefore, that a patient in that zip code who has just been discharged from the hospital will have difficulty making it to a follow-up appointment at a distant physician's office.[102] Another use for predictive analytics is predicting the **"flight path" of a patient.** Leveraging historical data from other patients with similar conditions, predictive algorithms can be

	created using programming languages such as R and big data machine learning libraries to faithfully predict the trajectory of a patient over time.[102] Healthcare applications would benefit from processing and analysis of multimodal data. The fusion of different health data sources could enable the **study of phenotypes that have proven difficult to characterize** from a genomic point of view only.[101] Analysis of lifestyle data collected from apps on smartphones can be used within (learning) **recommender systems that help monitor patients, raise alarms**, or give advice for the better handling of a disease.[101] The use of knowledge bases constructed from sophisticated ontologies has proven to be an effective way **to express complex medical knowledge** and support the structuring, quality management, and integration of medical data.[101]
 Technological challenges	**Data quality:[101]** As the complexity of operations grows, with new analysis methods being developed quite rapidly, it becomes key to record and understand the origin of data which in turn can significantly influence the conclusion from the analysis. **Cleaning:[103]** Dirty data can quickly derail a big data analytics project, especially when bringing together disparate data sources that may record clinical or operational elements in slightly different formats. Data cleaning ensures that datasets are accurate, correct, consistent, relevant, and not corrupted in any way. **Data quantity:[101]** There is a need to deal with this large volume and velocity of data to derive valuable insights to improve healthcare quality and efficiency. To enhance scientific analysis and relevant applications of Big Data in Health it is recommended to adapt and expand existing Big Data sources (e.g. data repositories in hospitals) in order to include necessary information not yet captured (e.g. biomedical data) and to complement them with newly explored sources.[98] **Multi-modal data:[101]** The combination and analysis of multi-modal data poses several technical challenges related to interoperability, machine learning and mining. Integration of multiple data sources is only possible if there are on the one hand, de jure or de facto standards and data integration tooling, and on

	the other hand, methods and tools for integrating structured, unstructured (textual, sound, image) data. **Data access:[101]** There is a high degree of fragmentation in the health sector: collected data is not shared among institutions, even not within departments. This leads to the existence and spread of different isolated data silos that are not fully exploited. Access to complementary sources of Big Data enables improved analytical insights and facilitates data analysis. To utilize this asset, it is recommended to support secure open use and sharing of government data, non-proprietary private data, and data of different healthcare providers for research in public interest on a national and international level.[98] **Healthcare knowledge:[101]** This knowledge exists in books and research papers, but also in the heads of healthcare professionals. New approaches are needed that bring together big data and knowledge, such that knowledge can be used to make better sense of data, and data can be used to generate more knowledge. **Analytical methods**:[98] To fully exploit this potential, it is recommended to constantly improve and update existing analytical methods and tools. Furthermore, their development and use (e.g. data mining, living laboratories, and rich open data repositories) should be facilitated while keeping in mind the practical use of Big Data.	
Necessary activities (in or for the public sector)		
 Development of a specific training necessary	**Open task**	One roadblocks to the general use of big data in healthcare is the **technical expertise** required to use it.[102] The value for big data in healthcare today is largely limited to research because using big data requires a very specialized skill set. Hospital IT experts familiar with SQL programming languages and traditional relational databases aren't prepared for the steep learning curve and other complexities surrounding big data.[102] In fact, most organizations need data scientists to manipulate and get data out of a big data environment. These **experts are hard to come** by and expensive and only research institutions usually have access to them. Data scientists are in huge demand across industries like banking and internet companies with deep

		pockets.[102] A report of DG Health recommends that the **digital health literacy of healthcare professionals and allied health professionals (e.g., managers) should be increased** through information and education. To achieve this, existing training and education programmes for public health or healthcare should integrate data handling in the curricula to ensure the development of the necessary skills and competencies.[98]
 Advanced or adapted ICT infrastructure needed	**Open task**	**Data storage** is a critical cost, security, and performance issue for the IT department. As the volume of healthcare data grows exponentially, some providers are no longer able to manage the costs and impacts of on premise data centres.[103] Data continue to increase at an exponential rate and the need for **cross-border exchange** of biomedical and healthcare data, **cloud-storage**, and **cloud-computing** is inevitable. Until many issues of data safety and security are solved, however, local solutions will be favoured.[104]
 Change of (public sector internal) processes necessary	**Open task**	Healthcare organizations that want to become data driven have to make several changes to their internal processes: they must commit to **valuing data as a strategic asset**, making data part of their culture, developing an understanding of the complete flow of data and acting upon data-driven insights. These organizations need to **encourage and reward the sharing of data and insights**, have management and executive teams who champion transformation and **build programs to develop data and analytics skills** across their enterprises. Progressing across the analytics continuum toward being a data-driven organization involves a **shift in the type of technologies and systems** involved in working with the data, as well as an evolution in the types of business questions being asked.[105]

 Promotion / information of stakeholders necessary	**Open task**	Traditionally, the healthcare industry has lagged behind other industries in the use of big data. Part of the problem stems from **resistance to change**—providers are accustomed to making treatment decisions independently, using their own clinical judgment, rather than relying on protocols based on big data.[92] The report of the Directorate-General for Health and Food Safety recommends **to raise awareness of the practical use of Big Data in Health and its benefits to make it more tangible and understandable for the public and concerned citizens.** It is therefore necessary to encourage a positive public mind set towards Big Data in Health by strengthening both the dialogue between the stakeholders in the field and the fact-based information towards the European citizens and patients.[98]
 Need to deal with cyber security issues	**Open task**	One roadblocks to the general use of big data in healthcare is a lack of **robust, integrated security** surrounding it.[102] In healthcare, HIPAA (Health Insurance Portability and Accountability Act) compliance is non-negotiable. Nothing is more important than the privacy and security of patient data. Although security is coming along, it has been an afterthought up to this point. But when opening up access to a large, diverse group of users, security cannot be an afterthought.[102] **Cyber thieves routinely target medical records**, and reportedly earn more money from stolen health data than by pilfering credit card details. In February, the largest ever healthcare-related data theft took place, when hackers stole records relating to 80 million patients from Anthem, the second largest US health insurer.[97]
 New or modified legislative framework or	**Open task**	A new **General Data Protection Regulation** (GDPR), replacing the previous Data Protection Directive (1995), was adopted in April 2016 and aims at harmonising legislation across EU Member States. As a "regulation" the GDPR will apply to all Member States without the need of transposition into national legislation. The

regulations necessary		GDPR will be implemented by mid-2018 to allow public and private sector to adapt their organisational measures to the new legal framework.[101] On EU-level, the implementation of national governance mechanisms for Big Data in Health can be supported by giving guidance on the process of data access approval and the technical implementation of data platforms, e.g. by **providing information on models of good practice for good data governance at research level** such as the International Human Epigenome Consortium (IHEC). [98]
 Development of a common standard necessary	**Open task**	By setting **common standards** across the Big Data value chain in Health, pooling, ex-changing and analysing data will become more efficient. It is therefore recommended to adopt, or to develop where non-existent, standards with global scope addressing the issues of **interoperability** (cf. ICT Standardisation Priorities for the Digital Single Market) e.g. in areas related to patient consent in the use of Big Data in Health or nomenclature of genotyping or ethics to name only a few.[98]
 Need for a more economical solution	**Open task**	Due to its unstructured nature and open source roots, big data is much less expensive to own and operate than a traditional relational database.[102] However, other important economic factors are related to data storage and personnel costs.
 Ethical issues	**Open task**	Ethical issues mainly refer to privacy and data protection. A lot of scepticism with regards to "where the data goes to", "by whom it is used" and "for what purpose" is present in most public opinion and, so far, European and international fragmented approaches together with an overly complex legal environment did not help.[101]
 Societal issues		No societal issues identified.

 Health issues		No health issues identified (apart from the expected positive impact on the health of the citizens)
 Public acceptance	Open task	People have specific expectations of confidentiality where their health data is concerned. They believe that when big data is used in the context of health data, these expectations are ignored or not sufficiently taken into account.[106]