European Commission

JRC SCIENCE AND POLICY REPORTS

Collaborative research-grade software for crowd-sourced data exploration: from context to practice

Part I: Guidelines for scientific evidence provision for policy support based on Big Data and open technologies

Jacopo Grazzini

Francesco Pantisano

2015

European Commission
Joint Research Centre
Institute for Environment and Sustainability

Contact information
Jacopo Grazzini
Address: Joint Research Centre, Via Enrico Fermi 2749, TP 266, 21027 Ispra (VA), Italy
E-mail: jacopo.grazzini@jrc.ec.europa.eu
Tel.: +39 0332 786658 – Fax: +39 0332 786325

JRC Science Hub
https://ec.europa.eu/jrc

JRC 94504
EUR 27094 EN

ISBN 978-92-79-45377-9 (PDF)
ISSN 1831-9424 (online)

doi: 10.2788/329540

Luxembourg: Publications Office of the European Union, 2015

Abstract

The scope and focus of the research reported in this document is to implement a collaborative high-level research-grade application software platform for scientific experimentation and data analysis. By doing so, we aim at exploring, extracting value from, and making sense of massive, interconnected datasets. Namely, the software is designed as an application layer that makes use of suitable statitiscal, exploratory or descriptive techniques, as well as visualisation tools, in order to produce reasonable interpretations of data – e.g. consisting in crowd-sourced data from social media, as well as other domain-orientated data, like sensor-based and geospatial data – that are logical but not definitive in their claims. We believe that by starting small and building quickly through a pilot, and by gaining experience from its deployment, it is possible to foster interdisciplinary and collaborative research that conjoins domain expertise. All together, it will lead to more holistic and extensive approach of entire complex systems. In order to consider all the potential issues and address all possible challenges in the future implementation, we adopt a multi-stage approach that aims to first acquire a clear vision of how to use data analysis and analytics, and thereafter this vision to the strategic needs of our research institution. Part I of the report provides the current Big Data and open technologies context (landscapes) in terms of European policies, states the motivation for our approach, and the foundations for an open, verifiable, reproducible, collaborative, and participatory framework for its deployment, and formulates applicable recommendations for implementation. Indeed, while it relies mainly on secondary literature – on Big Data, open technologies and data-driven decision making, as well as policy documents – this report actually defines a set of practical guidelines for the deployment and implementation of a Big Data software solution in our institution.

"Given the dynamic changes in R[esearch]&I[nnovation], policy making should also take account of emerging thinking and paradigms (such as big data, open innovation, and Science 2.0). Objective information and evidence is an integral part of policy making, including foresight and systematic ex-ante and ex-post evaluations."

"Research and innovation as sources of renewed growth" [28]

"To be able to seize these opportunities and compete globally in the data economy, the EU must [...] develop its enabling technologies, underlying infrastructures and skills, [...] extensively share, use and develop its public data resources and research data infrastructures."

"Towards a thriving data-driven economy" [30]

"The future of data analysis can involve great progress, the overcoming of real difficulties, and the provision of a great service to all fields of science and technology. Will it? That remains to us, to our willingness to take up the rocky road of real problems in preference to the smooth road of unreal assumptions, arbitrary criteria, and abstract results without real attachments."

"The future of data analysis" [259]

"Researchers need to adapt their institutions and practices in response to torrents of new data, and need to complement smart science with smart searching. [...] Above all, data on today's scales require scientific and computational intelligence"

"Community cleverness required" [209]

"Success in organizational change is not achieved simply by making the right decision at a particular time but rather through developing a social process that facilitates organizational learning."

"Advancing scientific knowledge through participatory action research" [267]

The purpose of this report is to present:

- some considerations on the use of Big Data[1] and open technologies for scientific support to policy through European initiatives [28] and within the large context of a 'data-driven economy' [30],
- a set of guidelines – and best practices – for the deployment, within a research organisation, of a Big Data system aiming at scientific evidence provision for decision-making [172, 218], that are derived from the previous considerations, and rely on a reasonable framework and applicable recommendations [26],
- a proposal for an actual interpretation and a practical implementation of these guidelines for the exploration of (crowd-sourced[2], but not only) data [260] within the smaller context of a pilot research project.

Beyond just the guidelines mentioned above, the final goal of the research reported herein is to provide a collaborative high-level research-grade application software for scientific experimentation and data analysis, and to gain experience from its deployment. By 'high-level', we mean to provide a platform that enables to focus on the application itself, without having to manage low-level technical aspects – *e.g.*, depending on the underlying infrastructure. By 'research-grade', we intend to have a high degree of flexibility to integrate and deploy algorithms very quickly onto the platform. The proposed solution is thought as an integrated software that will make use of suitable statistical, exploratory or descriptive techniques [139, 143], as well as visualisation tools [243, 165] to produce reasonable interpretations of heterogeneous datasets – *e.g.*, consisting in crowd-sourced data from social media, as well as other domain-orientated data, like sensor-based and geospatial data – that are logical but not definitive in their claim.

For that purpose, an 'intelligent' design is adopted, in order to offer a (as) clean and flexible (as possible) approach to long-term issues and potential future requirements – *e.g.*, scalability and accessibility are the most common requirements encountered when dealing with Big Data: we *"plan the whole"* in order to *"build the parts* [185], though we are also interested in 'evolutionary' development, by continuously implementing small changes to solve immediate problems. Having in mind the general and ambitious objective of making scientific processes more efficient, transparent and effective, and also because we believe the principle of openness should go beyond providing open access to scientific publications and analyses [10], we use open (source) technologies to develop the software, and we make it open source as well [208, 153]. We are aware of the fact that faults may emerge – as commonly they *"are found in data, models, computations as well as networks, storage media, and computational units; and they may emerge at the interfaces of software and hardware components that by themselves may work just fine"* [156] – but we also believe that *"anything less than the release of source programs is intolerable for results that depend on computation"* [153].

Considering the current context – in terms of research policies and legislation – inside the European Union and the position of our institution, our first consideration is to adopt

[1] Following the approach in [64], we have chosen to capitalise "Big Data" idiom throughout the report *"to make it clear that it is the phenomenon[s] we are discussing"*.

[2] 'Crowd[-]sourcing' is regarded as a process, not a phenomenon: we keep it with a lower case.

a multi-stage approach implementing some of the recommendations of [111] enounced to *"leverage analytics*[3] *in the public sector"*, namely:

[i] *"get a clear vision of how to use analytics"*, statistical analysis, data mining, and data exploration to solve problems, and

[ii] locally – *i.e.*, at the level of our institution – *"map to strategic needs"* in terms of research and technology infrastructures and human resources.

Part I of this report relies on the analysis of secondary literature – on Big Data, Open Data, open technologies and data-driven decision making – as well as policy documents, so as to provide – as part of the 'intelligent' design – the 'big' picture. Next, the scope of Part II is to replicate generic requirements and transpose policy guidelines across our project so as to:

[iii] *"address challenges for implementation"*, *e.g.*, related to Big Data intrinsic and systemic challenges, using open technologies, so as to

[iv] *"start small and build quick wins through a pilot"*, *i.e.*, aiming specifically at data exploration,

so as to follow an 'evolutionary' design as well. As for the practice, Part III will report some small-scal(abl)e show cases.

[3]Though 'analytics' is commonly used in the modern Big Data landscape, we will in the following be more traditional and stick to the more generic 'data analysis' idiom.

Contents

Abstract

The scope and focus of the research reported in this document is to implement a collaborative high-level research-grade application software platform for scientific experimentation and data analysis. By doing so, we aim at exploring, extracting value from, and making sense of massive, interconnected datasets. Namely, the software is designed as an application layer that makes use of suitable statitiscal, exploratory or descriptive techniques, as well as visualisation tools, in order to produce reasonable interpretations of data – *e.g.* consisting in crowd-sourced data from social media, as well as other domain-orientated data, like sensor-based and geospatial data – that are logical but not definitive in their claims.

We believe that by starting small and building quickly through a pilot, and by gaining experience from its deployment, it is possible to foster interdisciplinary and collaborative research that conjoins domain expertise. All together, it will lead to more holistic and extensive approach of entire complex systems.

In order to consider all the potential issues and address all possible challenges in the future implementation, we adopt a multi-stage approach that aims to first acquire a clear vision of how to use data analysis and analytics, and thereafter this vision to the strategic needs of our research institution. Part I of the report provides the current Big Data and open technologies context (landscapes) in terms of European policies, states the motivation for our approach, and the foundations for an open, verifiable, reproducible, collaborative, and participatory framework for its deployment, and formulates applicable recommendations for implementation. Indeed, while it relies mainly on secondary literature – on Big Data, open technologies and data-driven decision making, as well as policy documents – this report actually defines a set of practical guidelines for the deployment and implementation of a Big Data application software solution in our institution.

Outlines

This document is organised as follows. Next Section describes the Big Data phenomenon and the background of the so-called data revolution, and provides some figures. Section 2 shortly reviews the current EU Open Data policies that serve as a foundation for Big Data initiatives and support the transition towards Big Data, but also examines how the Big Data landscape endows new strategies to help design and deliver policies. We further explore the potential benefits and risks of using Big Data for effective data-informed – rather than driven – decision-making in Section 3. We introduce the so-called data value chain in Section 4 and present the fundamental systemic issues along the chain that a Big Data system needs to deal with, namely: complexity, heterogeneity, accessibility, scalability, timeliness, efficiency, and robustness. We also present in Section 5 relevant cross-cutting legal, methodological, technological and organisational challenges that accompany the data revolution and need to be specifically addressed by a BD system, that is: access to data, ethics & privacy, technologies & infrastructure, methods & techniques, as well as skills & experience sharing. In this aspect, we support the need for research and innovation to leverage the potential of Big Data and show how our organisation already proposed to move into this direction in terms of scientific activities. In Section 6, we introduce exploratory data analysis and its principles in large, and further relate it to problem-centric data-informed approach for scientific evidence provision. Motivated by the emerging consensus in the research world that science must be open, we introduce the requirements for an open, verifiable, reproducible, collaborative, and participatory framework for deploying a software application layer. We also formulate applicable recommendations for its implementation. This way, we aim at providing a set of guidelines and possible best practices to favour in-house collborative computational research activities. Finally, the conclusion 7 introduces a practical interpretation of these guidelines in the form of an open high-level research-grade application software that will be further presented in the following parts of the report.

1 Introduction: facing the data revolution

In recent years, diverse groups – researchers, businessmen, political scientists, and other domain experts – have identified a phenomenon based on the flow of new digital data that has been coined as *"data revolution"* – or *"data deluge"* – in popular news media [100, 194]. The so-called *"Big Data"* (BD) phenomenon was underway[1]. The following quote[2] provides a rather intuitive picture of this phenomenon in its current state:

> *"Every single minute, the world generates 1.7 million billion bytes[3] of data, equal to 360,000 DVDs. This works out at over 6 megabytes of data for each person every day. This information comes from many different sources like people, machines or sensors. This could be climate information, satellite imagery, digital pictures and videos, transaction records or GPS signals."*

1.1 Steady evolution

With the increasing usage of the World Wide Web, the tremendous amount of data and its steady rate of growth have been building for some time [214, 53, 268]. Research has shown that data has[4] grown globally from 5 exabytes[5] in 2003 to 2,700 exabytes in 2012, and is expected to grow in Western Europe from 538 exabytes to 5.0 zettabytes between 2012 and 2020, close to 40% a year [199, 117]. Though the sustainability of such a growth[6] is questionable, this trend[7] is expected to continue indefinitely [241, 199]:

- the advent of the Internet has generated exponential growth in the global user community,
- the interaction of these users with Internet applications has resulted in unprecedented levels of data and transaction volumes,

while digital information derived from all types of human activities is always easier to access [150, 31]:

[1] See for instance Wikipedia definition of BD at http://en.wikipedia.org/wiki/Big_data, though this definition may also vary depending on the target audience [265] or the technological considerations [151].

[2] Press release of October 2014 on *"2.5 billion € partnership to master Big Data"* launched by the European Commission and data industry: http://europa.eu/rapid/press-release_IP-14-1129_en.htm.

[3] 1 million billion (10^{15}) bytes = 1 petabyte, while 1 billion (10^{9}) bytes = 1 gigabyte; check Wikipedia table of orders of magnitude: http://en.wikipedia.org/wiki/Orders_of_magnitude_%28data%29.

[4] Singular or plural "data"? Check discussions: *"Data are or data is?"* (http://www.theguardian.com/news/datablog/2010/jul/16/data-plural-singular) and *"Is 'data' singular or plural?"* (http://www.quickanddirtytips.com/education/grammar/is-data-singular-or-plural).

[5] 1 exabyte = 1000 petabytes (10^{18} bytes).

[6] *"A full 90% of all the data in the world has been generated over the last two years"*: http://www.sintef.no/home/Press-Room/Research-News/Big-Data--for-better-or-worse/.

[7] More figures related to BD at http://wikibon.org/blog/big-data-statistics.

- an ever-expanding access to computing power and bandwidth benefit to users,
- a better connectivity enables larger amount of data per usage, thus creating larger and larger streams of continuously evolving data.

In this context, the term "Big Data" is a "catch-all" phrase[8] that is used to designate not only data, but also associated technologies and information processing methods and has become a major topic, a *"megatrend"* [152], in particular in the Information and Communication Technologies (ICT) field [263, 265]. BD[9] is in fact used to describe data collections which merge data from multiple sources – especially unstructured data – into a single one beyond the ability of standard techniques and commonly used hardware and software platforms to handle – capture, store, manage, process and analyse – the data within a tolerable time span [199]. While much emphasis is placed on the notion of data being big[10] – owing to the obvious quantity and volume of data – it is not just the increase in volume, but rather the types of data more frequently being collected and their ability to be leveraged that constitute the true "revolution".

1.2 A true revolution

The volume of data, its velocity and variety, and the need for veracity[11], have exploded [183] because of new social behaviours and new societal transformations [233]:

- the core social networks (*e.g.*, Facebook, Twitter, ...), by their very nature, have generated massive new ways for people to communicate and interact,
- many specialised social media (*e.g.*, LinkedIn, ...), micro-blogs, have also arisen, providing continuous streams of user shared activities,
- crowd-sourced[12] initiatives produce an ever-growing amount of data stemming from experiences and behaviours of active individuals, groups and communities,

or new business models migrating to the web [239]:

- the overall expansion of the worldwide economy has spurred massive data growth for traditional commerce,

[8] *"BD is, in many ways, a poor term"* [64].

[9] *"A very short history of Big Data"*: http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/.

[10] *"It is fair to say that the I[C]T world has been facing big data challenges for over four decades – it's just that the definition of 'big' has been changing"* [61].

[11] Check these infographics: http://www.ibmbigdatahub.com/infographic/four-vs-big-data and http://whatsthebigdata.com/2013/07/25/big-data-3-vs-volume-variety-velocity-infographic/, to grasp intuitively the main concepts related to the "three +1" V dimensions in BD analysis. Other V's worth considering are: validity, veracity, value and visibility (http://roblivingstone.com/2013/06/big-data-or-black-hole/), unless we want to avoid the *"wanna V" confusion* (http://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077?).

[12] See Wikipedia definition: http://en.wikipedia.org/wiki/Crowdsourcing. Note that rather than just considering *'crowd-sourced data'* – which gives the idea of data voluntarily contributed – we aim at exploring more generally *'community-contributed data'* – may they be actively or passively collected [88].

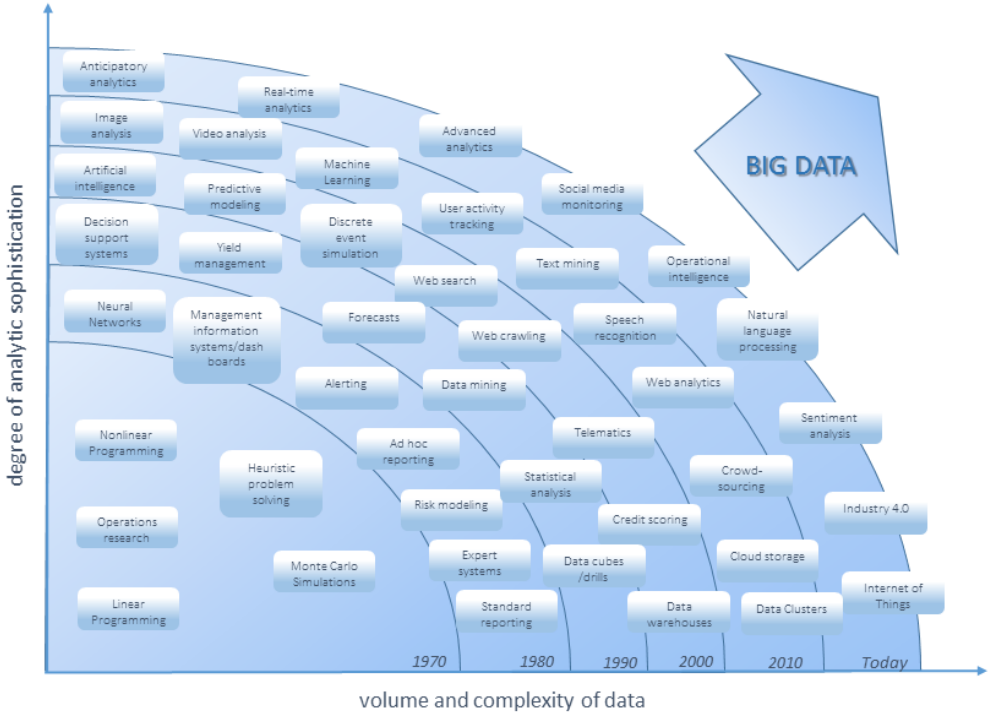


Figure 1: The "data (r)evolution" explained by data complexity, both intrinsic – the data per se, e.g. the V-dimensions – and extrinsic – in terms of analytical tools required to process the data. Source: [103].

- the shift to online advertising (*e.g.*, supported by the likes of Google, Yahoo, ...) is a key driver in the data boom we are seeing today,
- an entirely new breed of social network applications has been spawned, leveraging the inter-connection of social network users,
- web- and advertising-analytics applications crawl and analyse virtually every aspect of the users' interactions,
- a steeply growing volume of personal and behavioural data is collected in exchange of free digital services or liability programs,

as well as the vast spread – caused by the so-called *Internet of Things*[13] (IoT) – of software systems and sensors connecting people and devices [134, 219]:

- ubiquitous computing and electronic communication technologies generate data from both digital and analog sources,
- mobile/smart phones are by far the most commonly used communication vehicle for much of the world's population, and new wearable devices generate new types of data,

[13] *"The Internet of Things will radically change your Big Data strategy"*: http://www.forbes.com/sites/mikekavis/2014/06/26/the-internet-of-things-will-radically-change-your-big-data-strategy/.

- quick, voluminous data is also being produced as a by-product of use of other readily available, high-growth electronic devices, tracking devices, ...

Also of major relevance are datasets produced by specialised infrastructures and large-scale research projects [66]:

- Earth observation through remote sensing technologies (satellite monitoring) produces datasets with various granularity, size, diversity,
- modern equipments in medicine and genomics now generate health-related records in digital format,
- scientific experiments and simulations (telescopes, particle accelerators) are also associated to a bulk of data.

Figure 1 represents the complexity and the variety of topics and issues met when dealing with BD. As it is often claimed [199], BD is *"the next frontier for innovation, competition and productivity"*.

1.3 New opportunities for policy-making

The potential benefits of the use of these massive quantities of information – about people, things, and their interactions – are commonly acknowledged [30, 19] together with *"the hope to harness the knowledge they hide to solve the key problems of society, business and science"* [16]. In its recent report to the United Nations (UN) Secretary-General [133], the Independent Expert Advisory Group (IEAG) on data revolution for sustainable development suggested a comprehensive programme of actions in four areas: principles and standards, technology, research and innovation, capacity and resources, and leadership and governance. Indeed, BD draws not only on existing and new sources of ever-growing data, it builds also on new methodologies and emerging ICT resources, and advances thanks to innovative initiatives [133]:

> *"The data revolution is a revolution of possibilities – of new technologies, data production and dissemination systems and new resources opening up to produce more and better data, as well as expanding what can be done with data. [...] It consists of the entire data ecosystem – national statistical agencies, mapping authorities, government administrative data collection systems, academia, independent think tanks, researchers, civil society, private sector, media and individuals. [...] A healthy data ecosystem is typically characterized by strong complementarities and robust engagement and free debate among these actors."*

In this context, several organisations in the private sector have not only created sustained competitive advantage from the extensive use of data [185, 8, 274], but also leveraged data-driven decision-making – basing decisions and actions on data, rather than purely on intuition – to enable and drive their strategies and performance in increasingly volatile and turbulent environments [226, 40, 45]. Since policy advice is also becoming increasingly supported by data [13, 218, 226, 12], public organisations are embracing the data revolution as well, and the debate for a next-generation policy-making has already been initiated [272, 118, 207, 21].

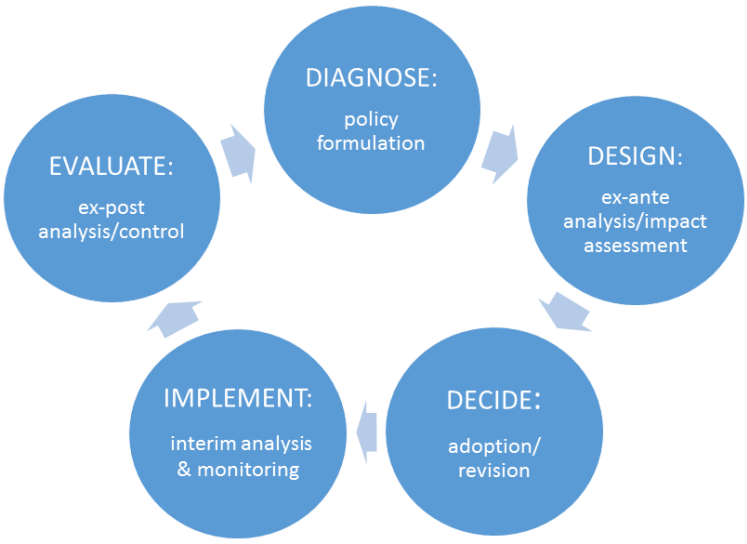


Figure 2: Commonly acknowledged phases of the policy cycle.

Following the corollary *"more data and better processing improve policy-making"* [205], the institutions of the European Union (EU) – at the forefront of which is the European Commission (EC) which designs the policies[14] of the EU – expressed their interest to use BD [30] to support, or even drive, the accomplishment of key policies – *e.g.*, for developing these policies and simulating their impacts on socio-economic and environmental systems [1, 12, 206]. It is believed that, through better data and knowledge management to inform decisions, the combination of BD together with emerging ICT will improve current governance processes by enabling data-driven policy-making and potentially reduce the costs and risks of policy decisions [205]:

> *"The concept of distributed policy-making and intelligence, with open governance and integration of policy intelligence to harness collective intelligence, realised into a 'distributed platform' based on ICT-enabled policy modelling (appropriately supported by participative and user-friendly simulation and visualisation tools), may prove to be instrumental to further implement policies and achieve socio-economic impacts. This would generate a 'cascade' of public and private decision-making on society's course of change and affecting the interactions that precede formal policy-making processes."*

While simply reporting the political decision is no longer acceptable[15], the data and the detailed information about sources, the underlying assumptions (models and methods) and also the tools (software) used to support the decision are today expected to be accessible [272]. At the time where the citizens' demands for more transparency in the EU institutions are growing [27], it also underpins the movement towards not only more open, transparent and integrated [55, 276], but also more participative [205, 110] policy-making systems. BD presents – together with Open Data (OD), see next Section – substantial promises for E-government

[14] http://ec.europa.eu/policies/index_en.htm.

[15] Ultimately, the purpose is to build public trust in the data, the models and the tools [258].

services[16], openness and transparency[17], and the interaction between governments, citizens, and the business sector [54, 207]. In addition, while citizen science becomes more formalised and more widely accepted among researchers, policy-makers, and communities [212, 24], crowd-sourced data [182, 88] can, through new ICT developments, also broaden participation in ways that were not previously possible [205]:

> *"The corollary of the enabling conditions and their main impact is that there will be a growing amount of data and of computing power that, through advancement in modelling techniques, may produce a quantum leap in our capacity to support policy making with real time, robust, evidence bases insights. This could happen not simply as a result of the work of traditional 'experts' (scholars, policy analysts, and policy consultants) but as a collaborative and participative effort of potentially all citizens."*

1.4 A challenge for scientific evidence-based decision-making

Because the data revolution affects all phases of the policy cycle [205, 206], including some of the underlying research activities that underpin policy-making with scientific evidence[18], the Joint Research Centre[19] (JRC), through its mandate to provide the EC with independent, evidence-based scientific advice and support, is also directly concerned with BD [255]:

> *"The nature of science is changing. It is more global. It is increasingly multi-disciplinary. And the raw stuff of science – the data – grows ever-more abundant. This data-intensity poses new opportunities and challenges across the scientific world."*

Data-driven science[20] reconfigures in many instances the way research is conducted [172]. In particular, domain experts and researchers should be able to formulate and simulate hypotheses – *'what-if'* policy scenarios – with BD so as to access a much deeper understanding of behaviour and interactions between global processes and systems [16]:

> *"Data science may empower [...] with the means to gain a better understanding of complex socio-economic systems, methods for introspection of complex global processes, tools for assessing the implications of decisions beforehand, and hence to improve our capacity to sustainably manage our society on the basis of well-founded knowledge [...]."*

[16] As a general trend, *"the accelerating digitisation of public services, driven by the need to modernise, cut costs and provide innovative services, opens up further opportunities to optimise data storage, transfer, processing and analysis"* [30].

[17] Note that Facebook also claims its mission is *"to make the world more open and transparent"* when collecting and using users' data: https://www.facebook.com/principles.php. This is clearly a *"transparency paradox"* of BD, as identified in [229].

[18] According to Wikipedia (http://en.wikipedia.org/wiki/Scientific_evidence), *"scientific evidence is evidence which serves to either support or counter a scientific theory or hypothesis; [s]uch evidence is expected to be empirical evidence and in accordance with scientific method."*.

[19] As the European Commission's in-house research service [15], the Joint Research Centre serves as the interface between science and policy-making, *i.e.* it provides technical and scientific evidence-based answers to the questions posed and advises decision-makers what course of action to take: https://ec.europa.eu/jrc/.

[20] Also referred to as *Data-Intensive Scientific Discovery* (DISD) in [220].

Therefore, it provides new capabilities to create scientific evidence aimed at supporting policy-making – *e.g.*, from formulation to evaluation, through design, implementation, and monitoring [205] – by guiding decisions (see policy cycle in Figure 2).

2 Policy support to Big Data and Big Data support to policy

There is undoubtedly a growing enthusiasm about this – *e.g.* web-based, crowdsourced – data deluge and the possibilities of using BD, especially for making better and quicker decisions in terms of policy [136, 12].

2.1 European policies aim at promoting Big Data

Many organisations have come to realise that the BD revolution and its effects are here to stay [8, 13, 58]. While it is evident that BD means business opportunities [122, 8, 22] – and the main actors in this field belong to the private sector[21] – it also raised the interest of policy makers [7, 30]. This is suggested by a policy focus for setting out some operational instructions – and sound investments[22] as well – to support and accelerate the transition towards BD while addressing the rise of related issues [19, 112]. Indeed, it is often implicitly assumed – not yet fully assessed[23] – that the use of BD will benefit both individuals and the society, as well as the economy [202]:

> *"The benefits to society will be myriad as Big Data becomes part of the solution to pressing global problems like addressing climate change, eradicating disease and fostering good governance and economic development."*

In this regard, the EU institutions and the EC confirmed a comprehensive approach to BD[24], in line with the targets set out in the EU-flagship initiative *Digital Agenda* [3] of the *'Europe*

[21] Check the top-100 BD companies: http://www.bigdatalandscape.com/bigdata100, and a list of influential BD vendors in [220]. Also find out *"who's big in Big Data"* using BD technology: http://treparel.com/news/whos-big-in-big-data/.

[22] In Europe, relevant actions at national and international levels are promoted and funded by the EU *e.g.* through the Seventh Framework Programme for research: http://ec.europa.eu/digital-agenda/en/what-big-data-can-do-you; see also footnote 2. In the United States, the Obama administration announced large investments in research and development through a BD initiative: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf. The UN Global Pulse (http://www.unglobalpulse.org/) is also working on building its networks around the world, partnering with governments, the private sector, agencies and academia on doing research.

[23] Carr argues in [74] that a *"data-driven society [...] would encourage us to optimize the status quo rather than challenge it [as it] will tend to perpetuate existing social structures and dynamics"*. Some skepticism has also been expressed by the IEAG on data revolution [133]: *"while [it] is already happening, it would be incorrect to assume its effects will be inevitably positive [...] Left alone, [the data revolution] is likely to reinforce existing inequities and patterns of marginalization"*. Read also *"The next technology revolution will drive abundance and income disparity"*: http://www.forbes.com/sites/valleyvoices/2014/11/06/the-next-technology-revolution-will-drive-abundance-and-income-disparity/.

[24] The European institutions, themselves, may be described as enormous data generation engines: http://europa.eu/publications/official-documents/index_en.htm.

2020 Strategy' [4] with a number of – legislative and non-legislative[25] – measures directed at the use of data sources in Europe[26]. Besides defining the key enabling role of ICTs, the Digital Agenda emphasizes the importance of maximising the benefits of the *Public Sector Information* (PSI), and specifically the need for opening-up public data resources as *"zettabytes of useful public and private data will be [made] widely and openly available"* [22]. In this aspect, the EU institutions have pushed for actions at regulatory and policy levels: the *'Open Data Strategy'* [5] – which is an amendment of the PSI directive[27] – fosters more transparency and openness by encouraging availability, access, use and reuse of public sector data, while they have already performed some actions at technical level[28]. The EC is also funding e-Infrastructures that implicitly address many aspects of interdisciplinary data management and indirectly promote related open practices[29]. When addressing BD[30], the EU institutions are however aiming well beyond OD, as it is believed the combination of OD and BD can create real-time solutions to challenges in many different policy areas, promote greater openness, and guide a new era of decision- and policy-making. As evidenced by a speech by former Commission's Vice-President Neelie Kroes [178], the leaders of the EU have recognised in BD one of the key characteristics of a data-driven economy and further recognised its potential for sparking technological innovation, creating new jobs and building up the knowledge-based economy[31]:

> *"Now we stand facing a new industrial revolution: a digital one. With cloud computing its new engine, big data its new fuel. [...] The impact and difference to people's lives are huge: in so many fields."*

In this aspect, the communication from the EC *"Towards a thriving data-driven economy"* [30] constitutes an important policy pillar in developing improved framework conditions for maximising the benefits of BD, in particular supporting the creation of a single market for BD[32]. This approach seems confirmed in the recently published communication on the EC's 2015 Work Programme [20] while it is recognised that *"the Digital Single Market*

[25] Further readings: https://ec.europa.eu/digital-agenda/en/legislative-measures and https://ec.europa.eu/digital-agenda/en/non-legislative-measures-facilitate-reuse.

[26] Website of the Digital Agenda for Europe: http://ec.europa.eu/digital-agenda/digital-agenda-europe.

[27] Directive 2013/37/EU of June 2013 (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:EN:PDF) on the re-use of public sector information, amending directive 2003/98/EC of November 2003 (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:EN:PDF).

[28] The EU Open Data Portal (https://open-data.europa.eu/en/data/) is a single point of access to data produced by the institutions and other bodies of the EU which promotes interoperability and use/reuse: making data accessible in downloadable, machine-readable formats to larger groups of people can enable innovative solutions.

[29] *e.g.*, European strategy on research infrastructures *ESFRI*: http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri.

[30] *"Big Data at your service"*: http://ec.europa.eu/digital-agenda/en/news/big-data-your-service.

[31] Conclusions EUCO 169/13 of October 2013 European Council (http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/139197.pdf).

[32] See press release on *"Commission urg[ing] governments to embrace potential of Big Data"*: http://europa.eu/rapid/press-release_IP-14-769_en.htm.

holds one of the main keys to a new dynamic across the European economy as a whole, fostering jobs, growth, innovation and social progress". Building an industrial community around BD in Europe is in particular the priority of the *Big Data Public Private Forum* initiative[33] which aims at defining a clear strategy that tackles the necessary efforts in terms of research and innovation, and supporting actions for the successful implementation of the data-driven economy while also providing a major boost for technology adoption. In parallel, the EU institutions are also addressing BD at a scientific level. As another facet of the *'Europe 2020 Strategy'*, the research programmes supported by the EC – *e.g.*, within *'Horizon 2020'*, or under the umbrella of *Future and Emerging Technologies* actions[34] – share together with the socio-economic policies proposed by the EC the same vision on data-driven innovation and growth[35]. Current research activities on ICT technologies carried-out in the EU have, in particular, integrated the BD landscape – often using a different terminology – in their framework [11, 158]. Overall, the availability of BD calls for a reconsideration of the mechanisms orchestrating the exploitation of data for designing, implementing, and evaluating public policies [16].

2.2 Big Data can help support European policies

Besides the data-driven economy[36] – and a pure market opportunity [199, 8, 169, 121] – the promise of data-driven decision-making[37] is also being broadly recognised [122, 226, 12, 207] considering that [133]:

> *"data are the lifeblood of decision-making, and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time; designing, monitoring and evaluating effective policies becomes almost impossible."*

Building on the pyramid of knowledge (see Figure 3), processed data become information that, when analysed, become knowledge, which can lead in turn to insights and informed decision-making [275]. This is also the point made in a section entitled *"Replacing/supporting human decision-making with automated algorithms"* of the McKinsey Global Institute report on BD [199]:

> *"Sophisticated analytics can substantially improve decision-making, minimize risks, and unearth valuable insights that would otherwise remain hidden. [...] Decision-making may never be the same."*

The huge volume of – potentially meaningful – digital information derived from all types of human activities is indeed being increasingly exploited to help solve complex systems which

[33] http://www.big-project.eu/.

[34] http://ec.europa.eu/programmes/horizon2020/en/h2020-section/future-and-emerging-technologies.

[35] Not to mention that some aspects of the OD policy have been integrated in *'Horizon 2020'*, like the need for open publications.

[36] Further readings: http://ec.europa.eu/digital-agenda/en/making-big-data-work-europe and http://europa.eu/rapid/press-release_MEMO-14-455_en.htm.

[37] In particular, BD analytics refers to the process of examining and interrogating BD assets to derive insights of value for decision-making [272].

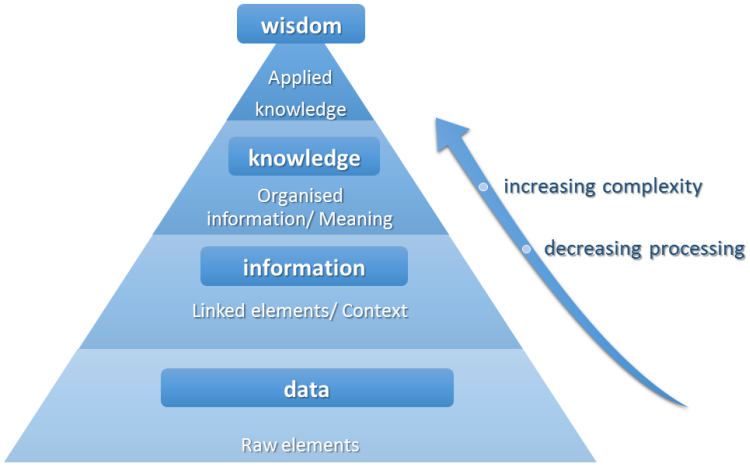


Figure 3: Pyramid of knowledge: From (raw) data to (organised) knowledge (and wisdom). Inspired by [275].

require to choose a course of action despite an incomplete understanding of the overall context [123, 115]. While decisions are taking place in an environment of uncertainty and rapid change[38], transforming BD resources into actionable knowledge is a major concern [251]. Above all, BD is about extracting valuable information from data to use in intelligent ways such as to improve decision-making [185, 22] – *e.g.*, not only for businesses, but also in science and society, and for policy support as well [207]:

> *"The emergence of open data portals and the explosion in data availability in the government context is opening great opportunities for collaborative governance and innovative way of using modelling techniques for improving policy-making."*

Decisions that were previously based on constructed models of reality can now be made based on the – structured and unstructured – data itself [21]. A conceptual framework for producing scientific evidence and gaining knowledge about complex – *e.g.* societal, economic, environmental – processes from the use of BD can be decomposed into the following phases:

[i] *describe*: use data to characterise a current or prior state of the system, for example monitoring and identifying outliers;

[ii] *investigate*: explore data to discover the boundaries and characteristics of a system, frame a problem or find supporting/discrediting evidence (*e.g.*, anomalies);

[iii] *verify*: perform rapid impact assessment;

[iv] *explain*: use data and analytic methods to determine causes and effects, build models and construct narratives;

[v] *predict*: apply analytic models to determine possible/probable future states of the system;

[38] *"Decisions are made in conditions of ontological uncertainty – that is, in situations where the future development of entities and their future relations are profoundly unknowable ahead of time"* [156].

[vi] *prescribe*: use data in models to define policy, procedure, and rules for taking action, and possibly automate them,

so as to "make sense"[39] of data contents – possibly well after it has been collected – and, ultimately, make inference and draw conclusions. The underlying idea is to define interdependent relationships to create a context that provides information about data (phase [i]) to further support (but not "replace"[40]) their analysis and generate new knowledge. BD value most surely comes from the patterns that can be derived by making connections between pieces of data – about an individual, about individuals in relation to others, about groups of people – or simply about the structure of information itself [64]. However, while we recognise the potential of BD for detecting anomalies and rare events [109, 89] – *e.g.*, detecting excessive variations from trends in context, see next Section – we believe, considering the current state of the art, that deriving direct assertions (decisions) from BD is still premature[41]. Instead, it should trigger a process of investigation and verification (phases [ii] and [iii]) so as to ensure both sensitivity and specificity of the derived information (phase [iv]): not missing a signal[42] and not picking up a false signal [190]. Hence, preliminary phases, where hypotheses are defined and tested (phases [i] to [iv]), merit special attention so as to avoid narrowing vision, missing important information, and distort evaluations [242]:

> *"The numbers have no way of speaking for themselves. We speak for them. [...] Data-driven predictions can succeed – and they can fail. It is when we deny our role in the process that the odds of failure rise. Before we demand more of our data, we need to demand more of ourselves. [...] If the quantity of information is increasing by 2.5 quintillion bytes per day, the amount of useful information almost certainly isn't. Most of it is just noise, and the noise is increasing faster than the signal. [...] Our predictions may be more prone to failure in the era of Big Data. As there is an exponential increase in the amount of available information, there is likewise an exponential increase in the number of hypotheses to investigate".*

Say it otherwise, the role of models in decision-making should not be overwhelmed by the data and the algorithms[43] to analyse them [41]. Hence we prefer the term *"data-informed"* to "data-driven"[44] as it gives the idea data should be used to inform decisions but decisions

[39] Note that, in the literature, "sensemaking" relates to an emerging class of technology designed to help organisations make better sense of their diverse observational space as to make better decisions, faster. More generally, it refers to a set of theoretical assumptions that lead explicitly to an overall approach to framing questions, gathering data, and conducting analyses for arriving at substantive theory [225].

[40] *"'Big Data hubris' is the often implicit assumption that Big Data are a substitute for, rather than a supplement to, traditional data collection and analysis"* [186].

[41] *"Extrapolating beyond the data is risky"* [251] and *"absence of evidence is not the same as evidence of absence"* [62], while *"claims to objectivity and accuracy are misleading"* [64].

[42] As already mentioned in Section 1 and represented in Figure 3, 'data' and 'information' are the basic building blocks in the field of information science – information can be seen as a refinement of mere data [275]. We also use the term 'signal' to represent indifferently the data content [197, 86].

[43] *"The problem with our data obsession"*: http://www.technologyreview.com/review/511176/the-problem-with-our-data-obsession/.

[44] *"Know the difference between data-informed and [versus] data-driven"*: http://andrewchen.co/know-the-difference-between-data-informed-and-versus-data-driven/.

should not be based solely on – mostly quantitative – data[45]. Because decision-making is a very complex issue[46], the decision-maker should, in the first place, aim at accurate description and general comprehension[47] ('nowcasts') instead of exact prediction[48] ('forecasts'), so as to derive the right question/hypothesis. It is assumed that detailed predictions of complex systems will initially be difficult to make (phase [v]), therefore the decision-maker will at first be given the means to monitor – *i.e.*, assess after implementation – the impact of (already existing) policies, looking for coarse-grained descriptions [1]. In that sense, different monitoring methodologies of BD by govern(ment)ance have been described so as to provide decision-makers with continuous feedback [171]. However, descriptions should eventually lead to (quantitative) prescriptions[49] for essential features (phase [vi]). Indeed, she should at last be provided with the means to predict – *i.e.*, evaluate prior to implementation – the impact of (foreseen) policies through integrated simulations of possible 'what-if' scenarios that will build on large datasets of past experiences and current observations [258, 16]. In addition, the information – say it otherwise, evidence – derived from data and supplied to the decision-maker should come with an assessment of its *relevance, validity, and reliability* [156]:

> *"Whenever policies are backed by simulations, they rely on the results of a computation. Yet the fact of the matter is that those computations inevitably contain faults. [...] Big Data implies big faults".*

In short, it is not enough to provide just the results [35]. Rather, one must provide to the decision-maker supplementary evidence that explains how each result was derived, and based upon precisely what inputs [35]:

> *"Ultimately, a decision-maker, provided with the results of analysis, has to interpret these results. This interpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. [...] For all of these reasons, no responsible user will cede authority to the computer system. Rather she will try to understand, and verify, the results produced by the computer."*

Thus, to really understand the opportunities offered by this seemingly ungovernable deluge of data, it is more than ever required to truly analyse the *strengths, weaknesses, opportunities and threats* (SWOT) related to the use of BD sources [64, 218]. While this is not in the scope of this document to further discuss its *relevance, validity, and reliability* – neither do we discuss the regulation – we will still try to understand how BD has been and can be used for efficient data-informed decision-making. We refer the reader to [6, 13, 22, 58] for further analysis and discussions on SWOT and real benefits of exploiting BD.

[45] *"Data informed, not data driven"*: http://www.youtube.com/watch?v=bKZiXAFeBeY.

[46] *"As we are considering developments that may change how policy making and decision making with implications for the collective good will take place we must remind ourselves that we are dealing with socio-economic issues far more important than network requirements and technological applications and pertaining to the broadly defined field of human and social sciences"* [205].

[47] Similarly to the concept of 'narratives' introduced in the orientation paper [156], which *"can help structure the exploration and communication of complex and uncertain issues, such as the scope and limits of modeling, or the unintended consequences of political actions. Narratives can help us follow a process from its beginning to the possible outcomes without losing 'the big picture'."*

[48] Though, potentially, *"Big Data opens up the realm of reliable predictive analytics"* [272].

[49] See semantic and methodological distinction between *'descriptive'*, *'predictive'* and *'prescriptive' analytics* in [131].

3 Effective data-informed decision-making

The first key question that arises in the presence of massive, interconnected datasets is what to do with all this data, also because *"turning an ocean of messy data into knowledge and wisdom is an extremely challenging task"* [16].

3.1 Decisions can be driven by data

In the public sector, the potential of BD has been recognised by statistical offices and institutions[50] – often experiencing a decreasing willingness of citizens to respond to surveys – for the production of official statistical figures [14, 18, 228]. Indeed, BD – available from open sources, but also possibly making use of data from private institutions or companies – competes with traditional methods of information gathering, and can be used in different ways, by replacing partially or completely traditional statistical sources[51], or by providing completely new statistical figures that may complement the available statistical information [25]. In particular, IoT applications supply with a tremendous capacity to enable crowd-sourcing of consumer and user data[52] in ways that can not only increase civic engagement [24, 219], but can also contribute to decision-making [205]:

> *"When objects can both sense the environment and communicate, they become tools for understanding complexity and responding to it swiftly. What is revolutionary in all this is that these physical information systems are now beginning to be deployed, and some of them even work largely without human intervention. [...] Future applications are opening up huge opportunities for private and public sector organisations alike. Despite the fact that many of the technologies which underpin the future internet infrastructure are not new [...], the conditions for their application may result in innovative and disruptive usages. This innovation could support many public policies, such as logistics, security, transport, environment and energy, education and health, and others."*

Statistics derived from such dynamic, fast moving data may have a competitive advantage for policy design in this regard [152]. The value and potential of BD have been already reported [226, 22, 21] for addressing the following societal challenges[53]:

- health-care: demographic change and well-being,
- energy: secure, clean and efficient energy,
- mobility & transport: smart, green and integrated transport,

[50] Following the Scheveningen Memorandum (http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf), several national statistical offices, the Directorate General Eurostat (DG ESTAT), and other bodies, have started to draft a road map for integrating the BD landscape into official statistics [162, 91].

[51] See comment in Footnote 40.

[52] http://ec.europa.eu/digital-agenda/en/internet-things.

[53] These challenges are also of major relevance for research and development within *"Horizon 2020"* framework: http://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges. They are also key fields of research for the JRC [15].

- environment & agriculture: climate action, resource efficiency and raw materials.

For example, widely aggregated data from the Internet can be a more accurate predictor than what is produced by a "community of experts" in a given sector. In certain areas – *e.g.* price statistics[54] – official statistics can be reliably issued on the basis of Internet data [91, 45]. This means that BD may push different types of official statistics toward more accurate predictions or more efficient methodology [181]. Supporting statistics through BD – on the basis of projections (now/forecasts) from huge volumes of mined or collected data[55] – offers big opportunities for enhanced insight and decision-making [190]:

- better early warning, to enable faster response,
- close to real-time (possibly cheap) observations to allow better awareness, to know what is currently happening on the ground,
- close to real-time feedback, to see what is not happening versus what was intended, to allow further adaptation and/or correction.

Understandably, there is a demand for official statistics not only to be more disaggregate[56], but also to be more frequent and timely. The benefit of BD in official statistics could also be an acceleration of the statistics production. One of the reasons of better timeliness is that the data collection phase of the generic statistical processes can be substituted in BD sources by automatic and instantaneous data availability [14]. This feature may influence in particular short-term indicators [181]. This idea is for instance illustrated by some studies which found that light emissions picked up by satellites could track GDP[57] growth [195, 148]. There are many other anecdotes about the success of using BD – *e.g.*, Internet-based data and phone records – for practical societal and economic problems[58] – *e.g.*, measuring and monitoring changes in society [224]. In particular it has been used in persuasive cases to:

- measure economic impact, like inflation – *e.g.*, through analysis of social networks and financial transactions: the "UN Global Pulse" correlates tweets in Twitter accounts regarding the high prices of rice with its actual price[59]; it reported as well some correlation between mobile phone usage and urban traffic [190];

[54] MIT *Billion Prices Project* (http://bpp.mit.edu/) and DG ESTAT *Consumer Price Index* (http://www1.unece.org/stat/platform/display/bigdata/Eurostat+-+Consumer+Price+Index+from+internet+price+data) conduct economic research using collected internet price data.

[55] *"Bigger is usually better for sample size"* [251] but *"bigger data are not always better data"* [64].

[56] The analysis of BD should probably go beyond common 'average' principles (joins, group-by, aggregation): *"[w]hile it may be useful to reason about the averages, social phenomena are really made up of millions of small transactions between individuals. There are patterns in those individual transactions that are not just averages, they're the things that are responsible for the flash crash and the Arab spring. You need to get down into these new patterns, these micro-patterns, because they don't just average out to the classical way of understanding society. We're entering a new era of social physics, where it's the details of all the particles – the you and me – that actually determine the outcome."*, Pentland cited in http://edge.org/conversation/reinventing-society-in-the-wake-of-big-data. Obviously, while this "personalisation" of data allows rapid access to more relevant information [135], it presents difficult ethical questions [250] such as the *"identity paradox"* identified in [229]; see next Section.

[57] *Gross Domestic Product.*

[58] *"GDP, welfare and the rise of data-driven activities"*: http://www.bruegel.org/nc/blog/detail/article/1044-blogs-review-gdp-welfare-and-the-rise-of-data-driven-activities/.

[59] http://www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress.

- forecast social impact and trends – *e.g.*, through analysis of consumer behaviour in social networks: Twitter mentions are used to predict both financial and popular success of movies in [227, 247];
- identify potential socio-political threats – *e.g.*, by monitoring critical events and precursor signs in the media: the JRC developed, in collaboration with the World Food Program, the "Humanitarian Early Warning System"[60], a system based on the "Europe Media Monitor"[61] that allows humanitarian organisations to detect possible conflict situations from information gathered on news portals world-wide [97];
- detect early digital signs of emerging trends, like disease outbreaks – *e.g.*, through fusion of social networks and geographical data: "Google Flu Trends"[62] tracks health-related online data – namely, search terms relating to illness – and map that data with the United States Center for Disease Control [126], while mobility data from mobile-phone records can help researchers recommend where to focus health-care efforts [266];
- build real-time indicators – *e.g.*, through integrated analysis of web traffic, log information and other events in order to find dependencies between economical entities: the studies in [75, 83] handle "Google Trends"[63] data to predict present (*"nowcast"*) economic metrics, *e.g.* track consumer purchases; Google Trends data are also used to predict daily price moves in the Dow Jones in [223];
- monitor environmental risks, like disasters – *e.g.*, through machine- or human-sensor networks ubiquitously collecting data: a study of JRC relates the (spatial and temporal) frequency of terms relating to fire used in social media like Twitter and Flickr with the actual occurrence of such fire, and map that data to "European Forest Fire Information System"[64] [87]; another example is the construction of an earthquake reporting system using again Twitter in [231];
- follow transportation – *e.g.*, through analysis and visualisation of live and detailed traffic network data: "IBM Smarter Traffic"[65] is a model for optimising an urban transportation system using movement data collected from millions of cell-phone users [52].

It is then expected that indicators based on BD sources will enable to track the evolution of some key short-term variables with some degree of accuracy [43, 123, 114]. By means of advanced statistical and computational techniques, decision-makers may disclose (frequent) patterns and (hidden) anomalies within BD that are not ordinarily evident on the basis of traditional information channels, such as census [35]:

> *"[E]ven noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. Further, interconnected Big Data forms large heterogeneous information networks, with which information*

[60] http://www.hewsweb.org.
[61] http://emm.newsbrief.eu/.
[62] http://www.google.org/flutrends.
[63] http://www.google.com/trends/.
[64] http://forest.jrc.ec.europa.eu/effis/.
[65] http://www.ibm.com/smarterplanet/us/en/traffic_congestion/ideas/.

redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent clusters, and to uncover hidden relationships and models."

3.2 Data should rather inform decisions

It seems very logical, and appealing, to claim that the same data and tools could be deployed into analytical decision-making systems, so that one could make full use of the abundance of data on social, economic, financial, and environmental systems available today[66]. However, while the previous applications have given weight to the promise, there are many counter examples as well, *e.g.* originating from the occurrence of spurious correlations[67]:

- in latest years, Google's (data-based, theory-free, real-time) model has been overestimating the spread of flu-like illnesses by almost a factor of two [69, 142];
- a study of Twitter and Foursquare data before, during and in aftermath of Hurricane Sandy in the New York metropolitan area supported the idea that Manhattan was the most hit place by the Hurricane [130],

or from the formulation of wrong hypotheses and the use of unadequate models/tools:

- as its title suggests, the study of [119] *"wanted to predict elections with Twitter and all [it] got was [a] lousy paper"* while trying to forecast the outcomes of elections using Twitter data; the same author also demonstrates that Twitter message allow no prediction of election results [120];
- a claim [72] by Princeton University paper stating that Facebook was set to lose 80% of its users by 2018 is heavily discredited by an answer from the company itself which demonstrates that, using the same flawed, non-robust model, the University would have no students by 2021[68].

Besides, statistics derived from BD sources do not provide any guarantee of high quality[69], neither professional independence, nor regular availability[70]. In addition to these isssues,

[66] *"How, exactly, is Big Data going to change the World?"*: http://www.scientificamerican.com/article/pentland-how-exactly-is-big-data-going-to-change-the-world-video.

[67] *"Correlation does not imply causation"* [251]. Find also relevant discussions on causation ('causality') *vs.* correlation interpretations of BD in the following weblogs: *"Correlation, causation, and Big Data"* (http://info.brightcomputing.com/Blog/bid/200074/Correlation-Causation-and-Big-Data) and *"Big Data news roundup: correlation vs. causation"* (http://www.forbes.com/sites/gilpress/2013/04/19/big-data-news-roundup-correlation-vs-causation/). Funny examples of spurious correlations at http://www.tylervigen.com/ illustrate *"our predilection for causal thinking"* and *"the ease with which people see patterns where none exists"* [159]. Recommended readings on causal inference and deductive/inductive reasoning are [197, 82, 143].

[68] *"Debunking Princeton"*: https://www.prod.facebook.com/notes/mike-develin/debunking-princeton/10151947421191849.

[69] Note that quality regards not only the accuracy of data, but also its 'trustworthiness' – the data contents is uncertain, it is not guaranteed to be reliable, though it may be 'fit-for-purpose' – and it is not anymore an attribute of the data only, it is also an attribute of the generating source.

[70] *"Not all data are equivalent"* [64] while *"no measurement is exact"* [251].

there is the question of erroneous data, data volatility or data representativity[71]: large Internet-based data sets, for instance, can be unreliable, prone to outages and losses, and these errors and gaps are magnified when multiple data sets are merged together [64, 186]. More generally [173]:

> *"Such approaches are not without critique, with detractors arguing that data analytics [...] struggle with the social (people are not rational and do not behave in predictable ways; human systems are incredibly complex, having contradictory and paradoxical relations); and with context (data are largely shorn of the social, political and economic and historical context); create bigger haystacks (consisting of many more spurious correlations making it difficult to identify needles); have trouble addressing big problems (especially social and economic ones); favour memes over masterpieces (identifies trends but not necessarily significant features that may become a trend); and obscure values (of the data producers and those that analyze them and their objectives)."*

All these problems – besides the major concern of ethics[72] and privacy[73] – need to be taken into account and addressed when designing BD systems and applications [173, 228].

4 The Data Value Chain: systemic components and issues

BD comes with a range of new requirements, raising difficult issues and posing new challenges [105]. The so-called *Data Value Chain* (DVC) aims at putting in place the "systemic" prerequisites for effective generation, acquisition, storage and processing of data assets [22].

4.1 Components of the Big (and small) Data Value Chain

Designing and deploying a BD system is not a trivial task. BD alone is not very interesting or useful. Investing in BD capabilities can be successful only if data is managed properly through its life-cycle, by setting the right context[74] to perform well-designed operations – *e.g.*, exploration and analysis – and integrating the insights gained from this data into the right processes, all in a timely fashion [176]. It is when data can be used and become actionable that it can *"create value"* [185]. In this regard, the concept of DVC – derived from systems-engineering [151] – is used to refer to the full (end-to-end) life-cycle of digital data [13, 49]. Its strategic importance for BD has been largely recognised [22]:

> *"Europe must aim high and mobilise stakeholders in society, industry, academia and research to enable a European Big Data economy, supporting and boosting agile business*

[71] *"Sample bias does not necessarily go to zero, even with Big Data"* [62] and *"bias is rife [and] data can be dredged or cherry picked"* [251].

[72] *"Just because it is accessible doesn't make it ethical"* [64].

[73] *"How Big Data is failing us"*: http://gawker.com/how-big-data-is-failing-us-1638605758?utm_source=recirculation&utm_medium=recirculation&utm_campaign=thursdayPM.

[74] *"Disruptions: data without context tells a misleading story"*: http://bits.blogs.nytimes.com/2013/02/24/disruptions-google-flu-trends-shows-problems-of-big-data-without-context/.

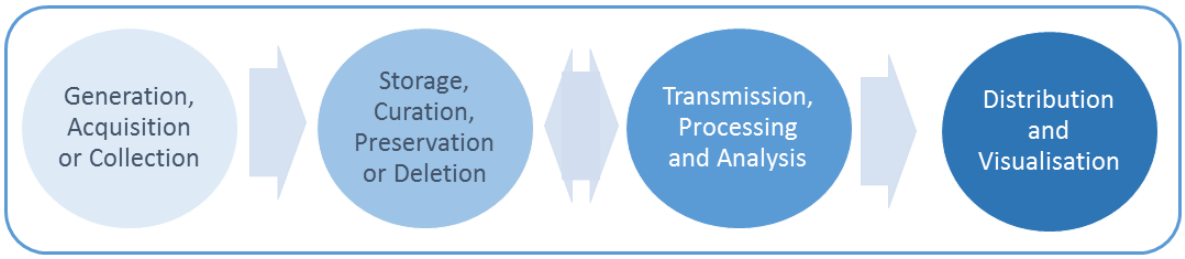


Figure 4: (Big) Data Value Chain representation, ranging from data generation and colletion through data processing and analysis, then to curation, use, and distribution. Inspired by [22].

> *actors, delivering products, services and technology, and providing highly skilled data engineers, scientists and practitioners [so as to offer appropriate solutions] along the entire Big Data Value chain."*

As shown in Figure 4, the DVC involves multiple distinct phases of the data management and processing pipeline[75]:

[a] data generation and acquisition [or collection];

[b] data storage, [access,] curati(ng)on, [preservation or deletion];

[c] data [transmission,] processing and analysis;

[d] data [distribution and] visualisation for [products and] services creation and provisioning.

Say it otherwise, a BD-dedicated system is not only responsible for archiving and accessing data, it needs also to support sophisticated forms of computation over data. Its overall purpose should be the creation of 'value' – *"beyond monetisation"* [11] – by collecting and combining data from different sources, processing data and providing access to it [185].

4.2 Systemic issues are CHASTER

Each single phase above introduces specific requirements, hence significant issues that are crucial to address [105]. First of all, the design of the data generation (and acquisition, and collection: [a]) phase will typically not all be laid out in advance, as we may need to figure out questions/hypotheses based on the data[76] and because the data stucture may be unknown in advance. Still, issues extend beyond this phase and are present all along the entire (Big) DVC. Owing to both intrinsic – the data *per se* – and systemic – the tools used

[75] In brackets, we mention the different phases which are, in our opinion, missing in the DVC representation of [22]. The order of implementation of these phases can also be discussed: for instance, storage and curation occur after data processing and analysis in case of stream processing, but will also be operated before in case of batch processing [151]; more generally, not only raw data, but also processed data are stored. See following discussions in this section, as well as Part II of this report.

[76] Though, again, it can be argued that *"with the new, [big] data-is-abundant model, we collect first and ask questions later"*, contrary to *"old, data-is-scarce model, [where we] had to decide what to collect first, and then collect it"* [90] – see Section 2.

Table 1: CHASTER challenges along the DVC accounted for in reference documents on key requirements for innovation in relation to BD. Namely: report on strategic research & innovation agenda [22] by the European BDVA, report on the future of cloud computing BD [158] by DG CONNECT, report on the future of global research data infrastructures [255] by the GRDI2020 Consortium, reports on BD [6] and on the future of software engineering [29] by the NESSI, white paper on BD [35] by an association of American researchers.

complexity	[22] *Data cleaning, integration, curation tools and services are required for data users to be able to differentiate noise from valuable data and to be able to integrate them and make them ready for analysis processes.* [35] *Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information. [...] The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured.* [158] *Not all data is structured and the relationship between information and its digital representative is neither straight-forward nor unique. Interpretation and analysis of the data is accordingly difficult and error prone.* [255] *There are problems in data modelling. For instance, metadata [...] can be as valuable as the data themselves. Data provenance, or lineage, is another challenge: where did the information come from? How was it updated? Another problem: data context. [...] How do we retain that context as we work more and more through remote computer systems? The uncertainty of any data is also important to understand and retain – error, incompleteness, inconsistency, ambiguity. [...] And then there is data quality: how 'good', useful, or appropriate are the data for the task in hand?*
heterogeneity	[22] *Data silos have to be unlocked by creating interoperability standards and efficient technologies for the storage and exchange of semantic data and tools to allow efficient user-driven or automated transformation. [...] There is an urgent need to build an interoperability layer upon all different systems taking advantage of transformation and semantic integration techniques* [35] *Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository. [...] With adequate metadata, there is some hope, but even so, challenges will remain due to differences in experimental details and in data record structure. [...] Greater structure is likely to be required by many (traditional) data analysis systems. However, the less structured design is likely to be more effective for many purposes.* [158] *Heterogeneity is increasing. Specialisation needs to be exploited, rather than homogenized, but thus require portability and interoperability beyond current mechanisms. As the heterogeneity grows, performance decreases and maintenance and porting issues increases exponentially.* [255] *Sounds straightforward, but many problems hamper interoperability. Heterogeneity arises when the way a users asks for information differs from one system to another, when models differ for the way information objects are represented, or when the semantics of the systems differ.*
(availability &) accessibility	[35] *Even for simpler analyses that depend on only one dataset, there remains an important question of suitable database design. Usually, there will be many alternative ways in which to store the same information. [...] Newer storage technologies do not have the same large spread in performance between the sequential and random I/O performance, which requires a rethinking of how we design storage subsystems for data processing systems.* [158] *Communication is the core bottleneck for availability, elasticity, sharing, consistency etc. [...] Bandwidth increases steadily. Latency hardly decreases, but reaches the physical limitations across larger and larger areas and routes. [...] Communication constraints reduce performance and accessibility on all levels. Not only data needs to be accessible, but also code. [...] At the same time, the data consumption rate grows faster than the bandwidth of storage.* [255] *Science needs well managed data – easy to get, find, store and analyse. But as data intensity rises, so do the challenges of management. [...] Another problem is data integration – combining data from different sources into a unified view.[...] Linking data also becomes more important.*

[continued on next page]

[continued from previous page]

scalability (& adaptability)	[22] *Being able to apply complex analytics techniques at scale and for data in motion is crucial in order to extract knowledge out of the data and develop decision support applications. [...] Improvement of the scalability and processing speed for [...] algorithms in order to tackle linearization and computational optimization issues.* [29] *How can we support the engineering of Big Data applications through targeted methods and platforms? [...] How to address the fundamental issues of scalability, performance, and availability that become necessary when dealing with Big Data systems and applications that have to cope with unprecedented size, speed, diversity and noise of data? [...] How can we build novel algorithms that store and cluster data objects [...] and subsequently facilitate searching and retrieving information, as well as presenting it in a useful manner? How can we leverage efficient storage techniques to greatly decrease the processing time of data?* [158] *Application[s] ha[ve] to adapt to changing usage context, as well as end points etc. [...] all offer some degree of adaptability, but are difficult to guide and typically slow. Many frameworks support automatic adaptation to different form factors, but software adaptability is still widely manual. [...] Workflow models and time analysis exists for single instances, but not with partially shared data or resources. [...] Statistical models for incomplete analysis over distributed data have not reached that maturity yet.* [255] *Better software and other data tools for science [...] will need better algorithms to analyse extremely large data sets 'approximately' rather than exactly, to analyse problems with many processors at once, and to 'steer' long-running computations so priority can be given to the most important data. They will also need better ways of visualising data, so it is to validate models, interpret information, play 'what if' scenarios, form hypotheses and look at data from multiple perspectives.*
timeliness	[6] *[A]nother relevant performance/scalability issue worth of significant efforts relates to the need that Big Data analysis is performed within time constraints, as required in several application domains. [...] Handling large amounts of streaming data, ranging from structured to unstructured, numerical to micro-blogs streams of data, is challenging in a Big Data context because the data, besides its volume, is very heterogeneous and highly dynamic. It also calls for scalability and high throughput.* [35] *There are many situations in which the result of the analysis is required immediately [e.g. interactive services demand that the data processing engine return back an answer within a fixed response time cap].* [22] *New Big Data-specific parallelization techniques and (at least partially) automated distribution of tasks over clusters are crucial elements for effective stream processing. [...] [There is a] need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.* [158] *The network and interconnects are growing way slower than data production and consumption [rate]. Modern approaches need to account for the delay of accessing data on all levels, whereas traditional means do not account for communication costs.*
efficiency (& performance)	[22] *The performance of the algorithms [...] must be scaled by orders of magnitude while reducing energy consumption with the best effort integration between hardware and software.* [35] *In the past, this challenge [of managing large and rapidly increasing volumes of data] was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.* [158] *Processor is not increasing anymore. Instead, modern processors incorporate more and more computing units, necessitating parallelism and memory management [...] addressing the communication complexity and needs beyond data-flow based processing. [...] Modern processors are not getting faster, but only bigger and more domain specific, meaning that applications now have to be developed for parallelism and, what is more, for communication / data access awareness. [...] With the number and co-dependency of all devices exposed to even just a single user, failures can have disastrous impact and increase in probability. Similarly, the energy consumption, though reduced per individual component, is increased by the number of devices required for a single application.*

[continued on next page]

[continued from previous page]

robustness	[6] *[T]here is the need for novel effective solutions dealing with the issue of data volume per se, in order to enable the feasible, cost-effective, and scalable storage and processing of enormous quantities of data. Promising areas that call for further investigation and industrially applicable results include effective non-uniform replication, selective multi-level caching, advanced techniques for distributed indexing, and distributed parallel processing over data subsets with consistent merging of partial results.* [29] *Problems like data replication, data consistency, temporary failures, communications latencies and concurrent processing need to be explicitly addressed in the system design. Such issues are amplified in a Big Data context, where systems need to dynamically grow to utilize data geographically distributed.* [35] *This [new] level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger clusters (that are required to deal with the rapid growth in data volumes). [...] Implications of [the] changing storage subsystem potentially touch every aspect of data processing, including query processing algorithms, query scheduling, datab ase design, concurrency control methods and recovery methods.* [158] *Storage systems and database systems [...] still grow slower than the total data production and consumption on the web and cannot compensate the network problems, making relocation and replication over different locations problematic.*

to handle it – properties of BD, the major key bottleneck points are commonly identified in phases [b] and [c], namely [8, 274, 220]:

- the capacity to manage, store, catalogue and organise complex data: data has to be managed in context, which may be heterogeneous, noisy, uncertain, and schemaless,
- the capacity to adequatly transmit, process and analyse data: data may need to be handled at large scale and in timely manner – *e.g.* real-time with low latency while ensuring as much as possible integrity and quality.

In this respect, a BD system – aimed at organising BD for 'value' extraction – will have to cope with unprecedented size, speed, diversity and noise of data [35]:

"Heterogeneity, scale, timeliness, complexity, and privacy[77] *problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. [...] Data analysis, organization, retrieval, and modeling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analysed."*

Alltogether, we identify – through a thorough analysis of various reference documents prepared by: the European Big Data Value Association (BDVA) [22], the Directorate General of Communications Networks, Content & Technology (DG CONNECT) [158], the GRDI2020 Consortium [255], the Networked European Software and Services Initiative (NESSI) platform [6, 29], an association of leading researchers from the United States [35], see Table 1

[77] Privacy is rather regarded as a crosscuting challenge, see next Section.

– the most fundamental issues that need to be addressed by any BD system and which we refer to as CHASTER, namely[78]:

- intrinsic to the data: *(i) Complexity* and *(ii) Heterogeneity* (say it otherwise, *interoperability* issues), and
- related to the system itself: *(iii) availability & Accessibility*, *(iv) Scalability & adaptability*, *(v) Timeliness*, *(vi) Efficiency & performance*, and *(vii) Robustness*.

Those should be focused priorities for BD solution systems and applications [193]:

> *"At the end of the day, an efficient and successful Big Data analytics platform is about achieving the right balance between several competing factors: speed of development, ease of analytics, flexibility, scalability, robustness, etc..."*

5 Crosscutting challenges

Having presented the main stages in the DVC and considered related CHASTER issues – see previous Section – crosscutting challenges clearly emerge. In this respect, the Organisation for Economic Co-operation and Development (OECD) has identified, in various reports on the potential of BD [13] and data-driven innovation [21], key points to address in order to overcome obstacles and to exploit BD opportunities. It states that efficient data-(informed)driven innovation requires policy actions – and practice – in five different domains: *access to data*, *ethics & privacy*, *technologies & infrastructure*, *measurements (methods & techniques)*, as well as *skills & experience sharing*, in order to deliver value. These domains have also been recognised as strategic challenges[79] by different (European) instistutions and consortiums throughout various policy-related workplan documents, *e.g.*, the European BDVA [22], the European Statistical Systems [77], and the High Level Group for Modernisation of Statistical Production and Services of the United Nations Economic Commission For Europe [25].

5.1 Access to data should be granted

Considering the generation phase [a] of the DVC (see previous Section), there is a broad range of data types and sources: structured and unstructured data, human- or machine-generated data, static or dynamic data, multi-lingual and sensorial data, ... In BD, value is also added by processing, validating, augmenting data – *i.e.*, through phases [c] to [d] of the DVC. Obviously, those considerations are central to the DVC activities while making datasets and assets available is paramount [22]. The various PSI initiatives and OD policy strategies (see Section 2) aim at doing so by actively seeking to engage with the data revolution [276]. Note that it is usually recognised – as a set of best practices [108] – that data, beyond being just available, should also be: discoverable – modeled and hosted in a way that they can

[78] Note that we did not include *elasticity* [99] in this list, not only because it would then make the acronym we chose meaningless, but also because we regard it more as a specifically business-oriented issue.

[79] It is also commonly taken note of both societal [21] and financial [13] challenge in the (Big) DVC.

be discovered through search, accessible – so they can be interrogated, when discovered, intelligible – so they can be read and understood by both human and machine, assessable – hence the reliability of their sources can be evaluated, and processable – they are actionable and can be (re)used. Other concerns are the integrity and quality of available data. We will however not discuss these issues in the rest of the report, instead further refer the interested reader to the various concerned policy documents [4, 5, 10, 30].

5.2 Ethics and privacy need protection

The increased importance of data intensifies the debate on data ownership and usage, ethics, data protection and privacy, security, liability, Intellectual Property Rights and the impact of insolvencies on data rights [22]. Therefore, legislation plays a crucial role in determining the framework conditions for the DVC – *e.g.* for data collection, processing and distribution. For instance, the 'deletion' we mention in phase [b] is of particular relevance for ethical and privacy issues on personal data, while it is considered less important in the context of BD – where the default is to keep data for long periods if not indefinitely. As underlined in [13], the DVC does usually not mention this process, although it deserves a prominent role from a (EU) policy perspective[80]: the Data Protection directive[81] and the so-called *"right to be forgotten"*[82] are indeed directly addressing the issue. Similarly, a council of policy advisors in the United States judges 'deletion' – together with data anonymisation and distinction between data and metadata – to be fundamental for privacy protection [17].

Besides the extent of personally identifiable information available in social media, transactional data, mobile telecommunications, *etc...* [135], new, increasingly powerful fusion [59], mining [204] and visualisation [243] tools and algorithms also provide many ways – *e.g.*, through phases [c] and [d] – to link data, suggesting possible scenarios and tangible examples of invasion of privacy [17]:

> *"The challenges to privacy arise because technologies collect so much data (e.g., from sensors in everything from phones to parking lots) and analyse them so efficiently (e.g., through data mining and other kinds of analytics) that it is possible to learn far more than most people had anticipated or can anticipate given continuing progress."*

This can potentially allow "Big Brother" to watch[83] ('nowcast') over us, or anticipate[84]

[80] *"Privacy and data protection in the EU"*: https://www.privacyinternational.org/reports/european-union/i-privacy-and-data-protection-in-the-eu.

[81] Directive 1995/46/EC of 24 October 1995 (http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&from=EN) on the protection of individuals with regard to the processing of personal data and on the free movement of such data, and all supbsequent superseding directives.

[82] Article 17 of the proposal for General Data Protection Regulation (http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf): *"Right to be forgotten and to erasure"*.

[83] As questioned in [63], *"will data analytics help make people's access to information more efficient and effective? Or will it be used to track protesters in the streets of major cities?"* It is generally recognised and accepted that *"[official] statistics [...] are a key pillar of democratic societies, providing a quantitative assessment of governments' policies and allowing for comparisons between countries, regions or effects of alternative actions"* [18], hence contradictions may arise.

[84] As argued by the authors of [202], the fact that BD seeks correlations rather than causation – see also Footnote 67 – could *"cut off common sense"* and lead to a *"dictatorship of data"*: data correlations are *"singularly unfit to decide who to punish and who to hold accountable"*.

('forecast') our actions [33, 264, 170]. General recommendations to ensure a 'reasonable' use of BD and 'preserve privacy' values [19, 254] have, so far, not fully address BD's risks and opportunities[85]. We further refer the interested reader to references [76, 254, 188, 17, 6, 181] as well as the special issue[86] of the Stanford Law Review Online on BD and privacy (among others: [170, 229, 222]).

5.3 Specific technologies and infrastructure have to be adopted

The inherent characteristics of BD – including the previous mentioned V-dimensions, see Section 1 – have deep implications [244] on a BD system, both in terms of hardware and software [176]:

> *"Through the development of new classes of software, algorithms, and hardware, data-intensive applications[87] provide timely and meaningful analytical results in response to exponentially growing data complexity and associated analysis requirements."*

First of all, the huge 'volume' of datasets requires a BD system to manage complex storage systems and overcome physical limitations [183]. Because of these physical limitations[88], it also encompasses new technologies that support, among others, the access and processing of this new data to create highly scalable applications without a human in the loop [6]. In addition, the 'variety' – *i.e.*, the heterogeneity of data types, representation, and semantic interpretation – and the 'velocity' – *i.e.*, the high rate at which data arrive and the short time in which it must be acted upon – further impose important requirements on the BD system [215]. BD encompasses a set of technologies and infrastructure that embody assumptions and design constraints different from traditional/standard systems [199]. Above all, depending on the data characteristics – most data generated is in fact originally streaming data – and the application purpose – whether the problem being solved is real-time ingestion of data for later use, real-time access to data, or real-time in-stream processing – it can use two alternative processing paradigms [160]: batch and streaming processing[89], which will cause conceptual, architectural and methodological distinctions in the associated platforms [151]. Following the nomenclature of [151], a BD system can be decomposed into a structure consisting in three hierarchical layers (see also Figure 5), from bottom to top:

[i] the *infrastructure layer* consists of a pool of ICT resources allocated to support the Big DVC, *e.g.* the hardware components for acquisition, storage, analysis, distribution, *etc...*

[85] In the United States, the Obama administration conducted a broad review on BD and privacy (http://www.whitehouse.gov/issues/technology/big-data-review) while still backing up, at the same time, bulk data collection by the National Security Agency (http://www.computerworld.com/article/2489349/data-privacy/nsa-phone-metadata-collection-program-renewed-for-90-days.html): this supports the idea in [229] of a *"power paradox"* with BD.

[86] http://www.stanfordlawreview.org/online/privacy-and-big-data.

[87] Note that 'data-intensive' is used not only to designate (BD) applications which require large volumes of data, but also applications that may devote most of their processing time to I/O and data manipulation.

[88] As one of its definitions suggests [199], BD is beyond the capability of current hardware and software platforms.

[89] *"In-stream Big Data processing"*: http://highlyscalable.wordpress.com/2013/08/20/in-stream-big-data-processing.

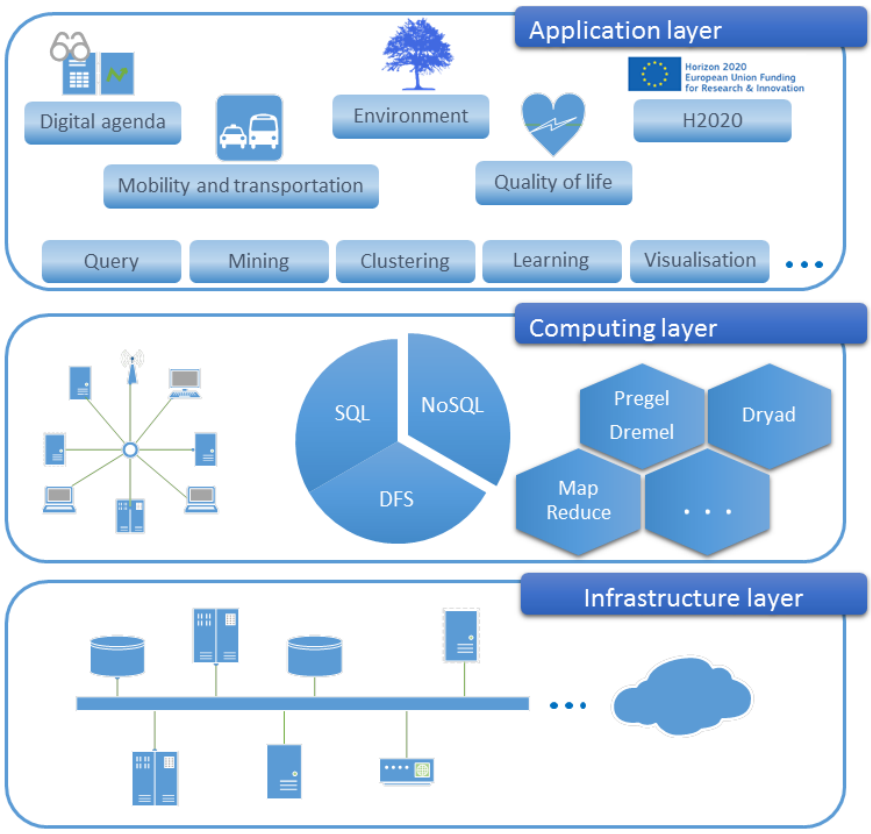


Figure 5: Conceptual layered architecture of BD system. It can be decomposed, from top to bottom, into an application layer (software/algorithms), a computing layer (middleware/framework), and an infrastructure layer (hardware/infrastructure). Our focus is the top layer: we aim at implementing a high-level research-grade application layer for scientific experimentation and data analysis. The hardware and middleware need dedicated research focus. Inspired by [151].

[ii] the *computing layer* encapsulates various data tools into a middleware layer that runs over the infrastructure layer: typical tools include data management and the programming model,

[iii] the *application layer* exploits the interface provided by the programming models to implement various data analysis and visualisation tools and combine them into dedicated applications.

Modern ICT infrastructures (layer [i]) aim at a good balance between enforcing data storage and access control (phase [b] of the DVC), and facilitating data processing (phase [c]). Hence, they should be able to scale up and out[90] and be dynamically configured so as to further accommodate diverse applications [176, 42]. Their deployment also depends on privacy (see above) and security [157] protection capabilities, as well as energy consumption [46, 174], throughput network and bandwidth efficiency [257], and fault tolerance [125]. To address these requirements, different technological strategies are often considered [216], *e.g.*: improved hardware performance and capacity [253] – using High Performance Computing (HPC), using more CPUs, faster GPUs[91]; reduced volume of accessed data [56] – using

[90] According to Wikipedia (http://en.wikipedia.org/wiki/Scalability), scale up means *"to add resources to a single node in a system, typically involving the addition of CPUs or memory to a single computer"*, while scale out means *"to add more nodes to a system, such as adding a new computer to a distributed software application"*. Check also *"Scale up vs. scale out"*: http://blog.clustrix.com/2013/01/24/scale-up-vs-scale-out/.

[91] *Central* and *Graphics Processing Units resp.*

appropriate cleansing and compression methods – and in-memory caching [273]; Distributed File Systems (DFS) [34] – storing data on many processing nodes, and taking advantage of higher-throughput network to apply processing on these nodes. Another prominently adopted technology for storing and providing fast access to large datasets is storage virtualisation, enabled by the emerging Cloud Computing (CC) paradigm [70, 158] and the new generation of Data-Intensive Scalable Computing (DISC) platforms [128, 67]. Exhaustive descriptions of the various possible solutions for the infrastructure layer are provided in [220, 151].

Data (integration and) management (layer [ii]) refers to mechanisms and tools that provide persistent and reliable data storage – with the necessary data pre-processing operations – and highly efficient management [65]. It has a tremendous and important history of achievements [127], and, since recently, numerous tools and algorithms to deal with Massively Parallel Processing (MPP) and DFS [61, 73]. Owing to certain essential features – *e.g.* being schema-free, non-relational, distributed, and scalable, supporting easy replication – the NoSQL database[92] has been largely adopted in BD systems, most of which are organised by the data model, namely: key-value[93], column-based[94], document-based[95], array-based[96], and graph[97] stores, as to make distinctive trade-offs to optimise specific performance [47]. Because relational (SQL-based) databases and NoSQL databases have their own advantages and disadvantages, many other projects have been implemented to integrate the advantages of both – like NewSQL[98] databases[99] – and support different types of data stores. Alongside the database structure, the programming model is critical to implementing the application logics and facilitating the data processing applications [151]. While various batch processing models have been introduced to address domain-specific application problems[100], the general purpose model based on the MapReduce (MR) algorithm is the most widely used – though improvements have also been suggested [68, 101] – and mostly supported by the various existing databases [93, 180]. It allows writing distributed applications that rapidly process large amounts of data in parallel on large clusters of computer nodes (or commodity servers) [102, 94]. Combined with a distributed storage system[101] cloning the Google's

[92] NoSQL commonly translates into 'not only SQL', where SQL is the *Structured Query Language*. Check SQL and NoSQL respective Wikipedia definitions: http://en.wikipedia.org/wiki/SQL and http://en.wikipedia.org/wiki/NoSQL. A list of NoSQL databases is provided here: http://nosql-database.org/.

[93] *e.g.*, Redis: http://redis.io/, Voldemort: http://www.project-voldemort.com/voldemort/, MemcacheDB: http://memcachedb.org/.

[94] *e.g.*, Google's Bigtable [79] and its derivatives, such as Accumulo: https://accumulo.apache.org/, Cassandra: http://cassandra.apache.org/.

[95] *e.g.*, MongoDB: http://www.mongodb.org/, CouchDB: http://couchdb.apache.org/.

[96] *e.g.*, rasdaman: http://rasdaman.org/, SciDB: http://www.scidb.org/.

[97] *e.g.*, Neo4j: http://neo4j.com/, InfoGrid: http://infogrid.org/.

[98] *"NewSQL – The new way to handle BD"*: www.opensourceforu.com/2012/01/newsql-handle-big-data/. Check also the Wikipedia definition: http://en.wikipedia.org/wiki/NewSQL.

[99] *e.g.*, HStore: http://hstore.cs.brown.edu/, NuoDB: http://www.nuodb.com/, Google's Spanner [85].

[100] *e.g.*, Microsoft Dryad [155]: https://github.com/MicrosoftResearch/Dryad, Google's Pregel [198].

[101] Hadoop DFS (HDFS), possibly extended by HBase: http://hbase.apache.org/. Note that dedicated solutions have been also implemented for efficient distribution, transmission, and synchronisation – *e.g.*, Avro: http://avro.apache.org/, Kafka: http://kafka.apache.org/, ZooKeeper: http://zookeeper.apache.org/.

DFS [124], and increasingly user-friendly SQL-like[102] data manipulation and query mechanisms [98], it provides almost unlimited scalability through the to-be-mentioned Hadoop[103] framework, which has become – because of its capability to handle massive data storage and processing[104] – the standard to cope with BD problems [221]. Beyond the storage, MR and query stack[105], several general-purpose data processing platforms have been developed for building applications with improved performance and increased flexibility, either on top of this framework[106] or independently[107]. While it is possible to extend the MR paradigm for stream processing[108], specific streaming models[109] have been designed for managing and processing continuous flows of data in real-time [249]. This raises the question of how an optimal architecture of a BD system should deal with historic data and real-time data at the same time [61].

It is not clear how all these technologies and approaches could converge to a solid solution (layer [iii]) that covers all different use cases – query processing, batch processing, and in-stream processing. There are already examples[110] of integrated solutions where the different techniques are put together to better support the user of the system. The Lambda architecture[111] proposed in [200] solves the problem of arbitrary processing any data in real-time by decomposing it into three components: a batch component, a serving component and a speed component, and thus defines the principles for building robust and scalable data systems. In fact, existing frameworks[112] often combine scalable batch processing for complex analysis, real-time query processing for online analysis, and in-stream processing for continuous querying [179]. This way, they offer a combination of different technical capabilities: accessibility, scalability – both in terms of storage and processing, low latency – both in terms of processing and retrieval, as well as the ability to run arbitrarily complex analysis. A technology independent reference architecture for BD systems – which is based on analysis of published implementation architectures of BD use cases – has also been proposed in [215]. Still, with so many kinds of databases and programming models, no one can be best for all workloads and scenarios [151]. As a comprehensive comparison is out of the scope of this

[102] *e.g.*, Yahoo's Pig [213]: http://pig.apache.org/, Facebook's Hive [256]: https://hive.apache.org/, Google's Tenzing [80].

[103] http://hadoop.apache.org/.

[104] Hadoop also benefits from a 'next generation' MR model, Hadoop YARN: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html, further extended by Tez: http://tez.apache.org/.

[105] *"The SMAQ stack for Big Data"*: http://radar.oreilly.com/2010/09/the-smaq-stack-for-big-data.html.

[106] *e.g.*, Pig, Chukwa: http://chukwa.apache.org/ Cascading: http://www.cascading.org/.

[107] *e.g.*, Asterix [38]: http://asterixdb.ics.uci.edu, Flink (ex-Stratosphere [37]): http://flink.incubator.apache.org/.

[108] *e.g.*, Twitter's Summingbird: https://github.com/twitter/summingbird/.

[109] *e.g.*, Twitter's Storm: https://storm.apache.org/, S4 [211]: http://incubator.apache.org/s4/, Google's MillWheel [36].

[110] *e.g.*, Google's BigQuery together with the Cloud Dataflow (http://googlecloudplatform.blogspot.it/2014/06/sneak-peek-google-cloud-dataflow-a-cloud-native-data-processing-service.html), Spark (through components Spark SQL/streaming): https://spark.apache.org.

[111] http://lambda-architecture.net. Check also *"BD Lambda architecture"*: http://www.databasetube.com/database/big-data-lambda-architecture/.

[112] *e.g.*, Google's Dremel [203] and PowerDrill [138], Drill: http://incubator.apache.org/drill/, Cloudera Impala: http://impala.io/.

report, we refer the interested reader to references [255, 6, 132, 29, 151, 244, 220, 215] and the literature therein.

5.4 Methods and techniques have to be advanced

Because the current technology allows to efficiently store and query large amounts of data, the focus is now on techniques that make use of the complete dataset, instead of sampling, and possibly operate at finer grain and faster speed [234]. Besides, compared to traditional approaches, the need is for analysis and processing rather than simply accessing large volumes of data – *i.e.*, phases [c] to [d], not only [b] – as it is believed truly *"mining and modelling BD can improve our understanding of the world"* [23]. Hence, the use of BD sources requires application of new types of methods and techniques. The advent of DISC platforms in particular has tremendous implications [193]. The aforementioned computing – such as HPC, MPP, CC[113] – and processing – batch or streaming – paradigms are also required to provide new approaches, though they are not trivial to adopt [184, 238, 248, 71]. In particular, to have specific parallel and distributed versions of some methods, a lot of research is needed with practical and theoretical analysis [57, 81]. Key aspects such as integration, interoperability and linking of data, information and content, low latency, scalability and real-time in processing data – *e.g.*, beyond batch and *ad-hoc* analytic tools – all have to be advanced through new approaches in areas like data mining [235, 270], machine learning [51, 246], pattern recognition and classification [145, 104], natural language processing [191, 192], as well as (geo)spatial [140, 144, 245] and temporal [141, 78, 261] processing, to name a few. In this regard, various scalable/distributed machine learning and data mining frameworks have been already proposed[114]. In addition, because it is very difficult to find user-friendly visualisations of large amounts of data [230], new techniques and tools[115] are also needed for so-called *'visual analytics'* [164, 107], so as to efficiently communicate the outcomes of the analysis and tell the narratives behind BD [146, 237], or further support the 'analytics' [166, 167].

More generally, the technological changes to BD system described above also indicate a trend in the methodological approach [215], towards an integration of technologies, infrastructure, techniques, and methods [6]:

> *"A holistic approach is needed for developing techniques, tools, and infrastructure which spans across the areas of inductive reasoning (machine learning), deductive reasoning (inference), high performance computing (parallelization) and statistical analysis, adapted to allow continuous querying over streams (i.e., online processing)."*

[113] Essentially, HPC, MPP, CC have been identified as *key software paradigms* to address in the field of ICT for the future [132].

[114] *e.g.*, MLbase [177]: http://www.mlbase.org/, MADlib [147]: http://madlib.net, Pegasus [161]: http://www.cs.cmu.edu/~pegasus, Mahout: http://mahout.apache.org/, Yahoo's SAMOA [92]: http://samoa-project.net/, Microsoft's Azure ML: http://azure.microsoft.com/en-us/services/machine-learning, IBM's SystemML [60], Google's Sybil: https://plus.google.com/+ResearchatGoogle/posts/7CqQbKfYKQf.

[115] Notice that *"data visualization is an assistance method for data analysis"* [151], *i.e.*, another way of exploring data and extracting information: in fact, we do not make much distinction between 'analytics' and 'visual analytics [165, 164], though the latter enables further accessible reasoning interactions [167, 237].

Following, various general-purpose data processing platforms[116] as well as domain-specific engine applications[117] offer full and integrated solutions. At the same time, the measurements are dependent on the data sources. Combining scalable processing with flexible data structures means that any digital content can be treated as data: new technologies coupled with new data enable new practices [269]. In particular, appropriate methods and techniques for processing and analysing data provided by the IoT – *e.g.*, social data, and other domain orientated data, like sensor-based and geospatial data – are needed. This opens up opportunities for interdisciplinary research [271], so that computational science is expanding into scientific research fields where expertise in computation is essential to advance [187, 163]. In the recent years, BD has indeed demonstrated relevant to almost all scientific areas [201]. In fact, an important aspect to consider is the combination of models and tools of different kinds/from different fields that can complement each other to better exploit the complexity of the information carried by BD [16]. We further refer the interested reader to references [189, 51, 91, 270, 105, 165] and useful literature reviews therein.

5.5 Skills and experience sharing are required

In a communication on its new Work Programme [20], the EC has identified the need for *"boosting digital skills and learning"*[118] as one of the main objectives to reach for an actual *Digital Single Market*. Within the BD landscape, it is particularly required [22, 77] to ensure the availability of highly and rightly skilled people – from so-called 'data scientists' to ICT staff – who have an excellent grasp of the methodologies, technologies and best practices – in phases [a] to [d]. In fact, the access, management, processing and analysis of BD require specific new skills, or combinations of complementary skills [25]:

> *"Generally speaking, analysts [...] are not programmers so they cannot assess the validity of a particular program routines, the programmers are not methodologists and mathematicians and so they do not have the requisite background knowledge needed to code a new [statistical] routine or evaluate an existing one."*

In particular, it has been recognised that there is the need for data scientists and engineers[119] who have expertise in statistics, machine learning, data mining and data management [84]. Another important element for delivering value along the DVC is to share experience on projects, applications, pilots and BD sources [77]. Not only developing skills, but as well

[116] *e.g.*, Spark (with subprojects GraphX/MLlib), Flink, Mesos: http://mesos.apache.org/.

[117] *e.g.*, GeoTrellis: http://geotrellis.io/ for high-performance geographic data processing, EarthServer: http://www.earthserver.eu/ for Earth data analytics, HIPI [252]: http://hipi.cs.virginia.edu/ for parallel image processing, Graphlab [196]: http://graphlab.org for complex networks analysis, IBM's Blast: http://researcher.watson.ibm.com/researcher/view_group.php?id=4947 for analytics on spatio-temporal data.

[118] http://ec.europa.eu/priorities/digital-single-market/index_en.htm.

[119] According to [77], the skills necessary to manage the complete workflow should in fact include: data analysts – able to use statistical software and visual analytics tools, data scientists – able to manipulate complex data sets, data engineers – designers of the ICT architecture, data integrators – running the data collection and integration processes, and systems managers – setting up and managing the physical infrastructure.

sharing knowledge and know-how and federating best practices[120] are important tasks within applications and proposed solutions for BD systems [22].

5.6 Research and innovation need to leverage the potential of Big Data

Research requires BD systems (see Section 1.4), but at the same time, research itself can be used to help improve these systems [201]. No need to mention that BD represents a major research challenge *per se* and a new disruptive approach to science [137]. Big (or small) data-driven science reconfigures in many instances the way research is conducted – by blending aspects of scientific induction, deduction[121], and abduction [172]:

> *"Data-driven science [...] differs from the traditional, experimental deductive design in that it seeks to generate hypotheses and insights 'born from the data' rather than 'born from the theory'. [...] In other words, it seeks to incorporate a mode of induction into the research design, though explanation through induction is not the intended end-point. It forms a new mode of hypothesis before a deductive approach is employed. [...] The strategy adopted within data-driven science is to use guided knowledge discovery techniques to identify potential questions (hypotheses) worthy of further examination and testing."*

In this aspect, it is even often considered as a *"paradigm shift"*[122] in science [168, 149, 202]. Instead, we adhere to the standpoints of [64, 172, 173] and recognise in data-driven science – or just *data science* – another pillar of research, standing equally alongside theory and experiment [173]:

> *"[T]he epistemological strategy adopted within data-driven science is to use guided knowledge discovery techniques to identify valuable insights that traditional 'knowledge-driven science' might fail to spot and then to investigate these further."*

Still, there is very little clarity about the readiness of our institution to take advantage of the BD opportunities and benefit from this new component of science [26], may it be a true

[120] *"Building data science teams"*: http://radar.oreilly.com/2011/09/building-data-science-teams.html.

[121] *"The two classical types of inference [are] known as induction, that is, progressing from particular cases (training data) to general (estimated dependency or model) and deduction, that is, progressing from general (model) to particular (output values)"* [82].

[122] Popular magazine *Wired* has portended one of the greatest "paradigm shift" in the history of science, claiming that *"the data deluge makes the scientific method obsolete"* [39]. This perspective has been largely put forward in the research community [32, 176], implying this deluge would dispose of experiment and theory in science [255]:

> *"From all this, an entirely new method for science is emerging, which some have called a fourth paradigm – moving beyond the older scientific methods of observing, theorizing or simulating, and into a new process of looking for meaningful correlations across vast data sets. Correlation supersedes causation as the source of new knowledge; and science can advance without painstaking cause-and-effect models, grand theories or any mechanistic explanation at all."*

Drawbacks and limitations of such considerations have been however pointed out in the literature [173], in particular spurious examples of correlation relationships (see also Footnote 67).

Table 2: Some of the JRC Task Force's recommendations on BD on the need to address challenges across the BD Value Chain for future scientific requirements of our institution [26]. Note that: (i) Open access to data is somehow taken for granted as it is expected that, through OD and Open Research initiatives, data from administrative sources will contribute to the BD phenomenon; (ii) the particular area of ethics and privacy is not explicitly mentioned, instead it is considered that different applications of BD, with different requirements, will have different treatment.

ethics & privacy	*["Towards a thriving data-driven economy"] is focused too much on the opportunities in BD and should be more balanced, also addressing certain risks (privacy, discrimination, etc.). [...] Whereas Big Data from technical JRC experiments require little privacy or IPR protection, external satellite data might bring along certain [IPR] requirements and public health data are very sensitive in terms of data protection and privacy.*
methods & techniques	*The fact that the Technology Academy of Finland awarded the prestigious 2014 Millennium Prize to Professor Stuart Parkin for facilitating the occurrence of the BD revolution shows that Big Data have indeed become a serious scientific area in the recent years. Based on its experience in handling large or complex or dynamic datasets, JRC has solid data science competence which currently is deployed in particular projects rather than generically to the issue of BD. [...] Due to the undisputed importance of BD, and in its responsibility of being the scientific service of the European Commission, the JRC should conduct a structured analysis of the risks and opportunities of entering into BD science.[...] Pay particular attention to matching a clear definition of scientific needs with a well-documented Data Centre service portfolio.*
technologies & infrastructure	*In many areas of work, the scientists are of the opinion that use could be made of new computing paradigms (such as HPC, parallelisation, cloud computing, SaaS etc.) but were neither sure in which way to exploit these best nor when a suitable point in time might come to consider these new paradigms. [...] Choosing an underlying ICT strategy is difficult, as current market actors such as NASA or CERN and Google are following different approaches. Whereas the former institutions operate huge and central High Performance Computing centres with mainframe capability, Google is fully based on a light-weight, scalable and distributed architecture with hundreds of data centres across the globe, equipped with a huge number of rather standard computing cores. Interestingly, both approaches are complemented by an overflow capacity which is made available over cloud computing concepts in cases of need. [...] Maximum flexibility at both sides (scientists and data centre) [is required] to bring the loose ends together efficiently.*
skills & experience	*Transfer knowledge and experience about ICT concepts relevant to Big Data a structured way to scientists who might need it in the future. Train suitable JRC scientists (being application domain experts) on Big Data concepts and technologies, benefitting from the Big Data knowledge already existing in other places of JRC. [...]Most international players are currently taking a closer look at Big Data with different results. Mutually monitor the progress of respective Big Data work and lift synergies between the JRC competences and the [DG] ESTAT approaches towards the use of Big Data. Additionally, [it is] recommended to look into the feasibility of twinning technical pilots of [DG] ESTAT and JRC.*

shift or not[123]. While the capacity of producing and storing data[124] is increasing daily at a very fast pace [150], our ability to understand and interpret such an overwhelming amount of complex data has indeed not grown at the same rate. It appears clearly that future research efforts should integrate BD solutions in our daily work [63]:

> *"Big Data not only refers to very large data sets and the tools and procedures used to manipulate and analyze them, but also to a computational turn in thought and research".*

We already indicated the need for advanced data handling and processing, as well as the infrastructure to support it, and the development of new skills. In particular, we need to leverage our research expertise to transform the ability of our organisation to handle BD (see Section 4) so as to advance in scientific areas where the use of computational science is critical.

When addressing and providing recommendations on BD systems, the IEAG highlighted the role of research (and innovation) activities – alltogether identified as a major pillar of the data revolution, see Section 1 – and encouraged innovation and experimentation [133]:

> *"Engage research centres and innovators in the development of publicly available data analytics tools to better evaluate long-term trends affecting sustainable development, identify the most effective policies for achieving it, to make better decisions at all level and support improved organizational planning, operations and evaluations."*

In this regard, our institution adopted a suitable approach. Indeed, most of the challenges presented in Section 5 have been addressed by the internal Task Force on BD assessing the future requirements of the JRC [26] as we detail in Table 2. Still, to realise the benefits of BD, there are a number of requirements for moving beyond common – *'business as usual'* – approach: naturally, a genuine ability to understand data and its usability for a given problem, and a solid scientific foundation to select an adequate design, but efficient systems to implement it are required as well [111]. Additionally, it is also necessary to create an integrated framework to provide researchers with the ability to both perform the analysis and repeat it with different hypotheses, parameters, or data, hence translating questions that are asked into a series of computational methods [44]:

[123] As noticed in [172]:

> *"Rather than empiricism and the end of theory, it is argued by some that 'data-driven science' will become the new paradigm of scientific method in an age of Big Data because the epistemology favoured is suited to extracting additional, valuable insights that traditional 'knowledge-driven science' would fail to generate."*

We further draw the attention on Ioannidis' assertion that *"the greater the financial and other interests and prejudices in a scientific field"* and *"the hotter [this] field"*, *"the less likely the research findings are to be true"* [154], hence some sense of proportion should be kept [41]:

> *"It becomes clear that the data deluge is the current wave of the future or, at least, is so regarded by many. The problem is that when 'waves of the future' show up they often wash away a number of worthy things and leave a number of questionable items littering the beach."*

We refer the reader to the interesting discussions in [63, 41, 172] about the epistemological change in data-driven science and whether BD phenomenon should be considered as a paradigm shift or not.

[124] Also because *"Big Data is what happened when the cost of keeping information became less than the cost of making the decision to throw it away"*, Dyson cited in http://www.science20.com/science_20/the_big_data_problem_will_also_be_a_problem_for_science_20-140575.

"If you can program a computer, you have direct access to the deepest and most fundamental ideas in statistics."

Owing to the nature of the underlying systemic issues and crosscutting challenges in the DVC, intermittent solutions – methods and tools – are necessary to drive the research and ensure that the efforts are aligned to the actual objectives, though there is no single clear path towards them [185]. Solutions need to be proposed – and released – with this specific consideration in mind so as to be feasible and still drive the foundational advances in BD [50].

6 Guidelines for scientific data-informed evidence provision

In order to enable efficient and effective decision-making – both bottom-up (inductive, data-based, driven-without-knowledge: from data to hypothesis) and top-down (deductive, prior knowledge-based, model-driven: from hypothesis to data) – adequate scientific approaches need be adopted to properly handle and analyse BD, processes that require tools and capabilities for exploring and understanding it [160].

6.1 Problem-centric question-driven exploration of Big Data

In the field of information science [275], data represents the raw material for information, and information is the raw material for knowledge – as a hierarchically organised structure (see again Figure 3). Regardless of the kind of ITC tools we use [220], the fundamental problem nowadays is that we cannot consume and make sense of all this data fast enough [172]:

"The challenge[s] of analysing Big Data is coping with abundance, exhaustivity and variety, timeliness and dynamism, messiness and uncertainty, high relationality, and the fact that much of what is generated has no specific question in mind or is a by-product of another activity."

Transforming – low-level, complex – data and resources into potentially useful – high-level, structured – knowledge is the purpose of "Knowledge Discovery and Data Mining"[125] (KDD), classically defined as *"the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns [or structures, or models, or trends, or relationships] in data"* through automatic analysis (mining) and discovery [106]. KDD is indeed designed to analyse data – usually large amounts of data, *i.e.* massive datasets – in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings

[125] We will not discuss herein the conceptual difference(s) – whether or not their differ by scope, purpose and focus – between knowledge discovery, data mining, DISD and *"analytics"* [274]. *"Although the buzzwords describing the field have changed – from 'Knowledge Discovery' to 'Data Mining' to 'Predictive Analytics', and now to 'Data Science', the essence has remained the same – discovery of what is true and useful in the mountains of data"*, Piatetsky-Shapiro at http://www.kdnuggets.com/2012/02/kdnuggets-15th-anniversary.html. We accept that these computational processes all imply the extensive use of data and somehow involve methods at the intersection of artificial intelligence, statistics, machine learning and pattern recognition, simulation and optimisation, and data integration, management and linking [106, 139, 274].

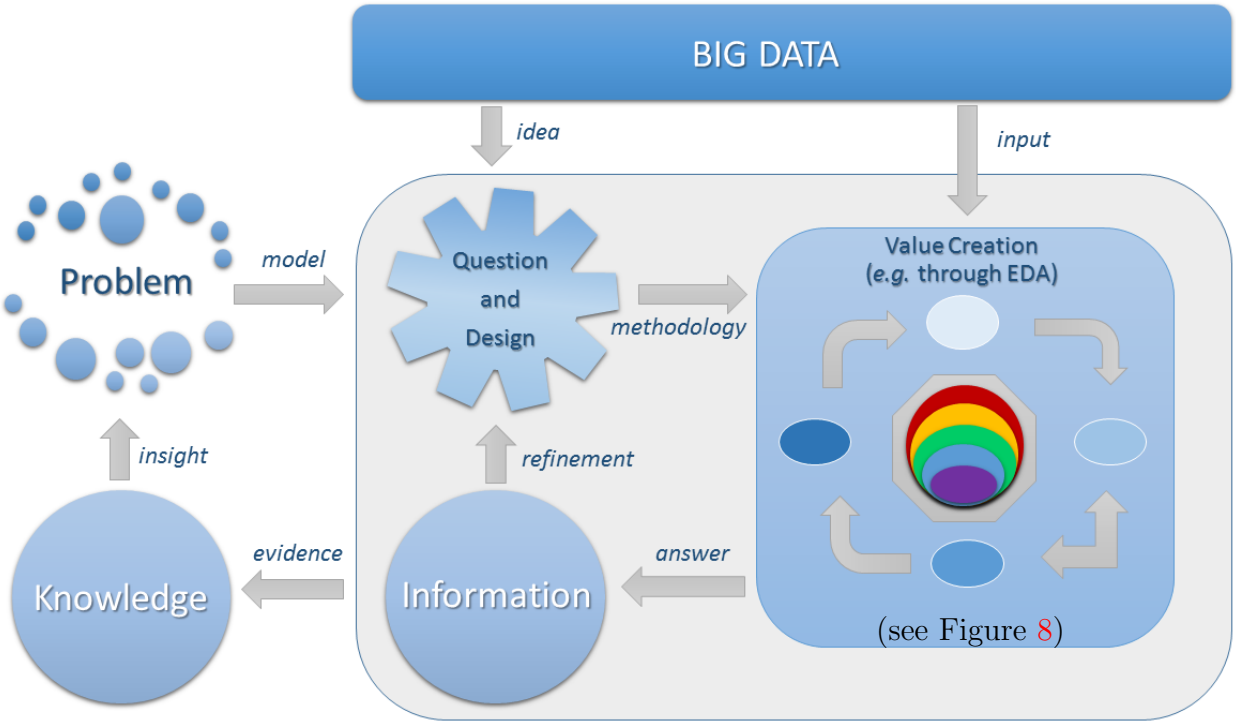


Figure 6: Problem-centric – and question-driven – data-informed workflow for scientific evidence provision: through EDA, BD supports the process more than it drives it. Compared to the "data value [single] cycle" of [21], we introduce multiple cycles in the workflow, as the information output by the analysis of BD ('answer') w.r.t to a specific problem ('question') can be used no only to produce 'evidence' that feeds knowledge, but also to reformulate ('refine') the question – thus the methodology – to recollect or reprocess data. Besides, BD, by itself, is not only a raw 'input', it is also potentially a source of new 'ideas' for research, as it is already the case for business (through 'value' creation). Inspiration is the original data analysis schema of Tukey [260].

by applying the detected patterns to new datasets [139]. In this analytical treatment of data, the so-called *"Exploratory Data Analysis"* (EDA) introduced by Tukey [259] plays an important role as it helps – say, the researcher – reduce the amount of information in order to focus on the pertinent aspect of relevant data before true analysis can be achieved. It essentially consists in a preliminary processing of data aiming at identifying interesting and previously unknown patterns without preconceived assumptions or models[126] for the purpose of generating hypotheses [240]. Actually, Tukey asserts that data analysis must proceed by approximate answers since the knowledge of the issue under investigation is likewise usually approximate [260]:

> *"Ideas come from previous exploration more often than from lightning strokes. [...] Broad general inquiries are also important. Finding the question is often more important than finding the answer."*

[126] In contrast to *"Confirmatory Data Analysis"* which is concerned with statistical hypothesis testing: it begins with the research hypothesis and determines if the data support it [260].

Hence, starting with some high-level hypotheses in mind, EDA helps at iteratively reformulating or revising them by learning from the data. This is consistent with the generic above-mentioned data-informed approach to science where [172]:

> *"the data are not subject to every ontological framing possible, or every form of data-mining technique in the hope that they reveal some hidden truth. Rather, theoretically informed decisions are made as to how best to tackle a data set such that it will reveal information which will be of potential interest and is worthy of further research."*

In short, the benefits of EDA are that (see also Figure 6):

- the data may suggest new research hypotheses – *e.g.*, exploiting humans ability to identify patterns not captured by automatic tools,
- it may provide ideas for further (re)collection and exploration of data,
- it can guide the researcher in selecting the appropriate successive procedures – *e.g.*, helping to select the right statistical tools for KDD.

Following a series of preliminary EDA procedures, one should indeed be able to identify definite patterns or detect any strange observations in the data, gain insight into the variability contained within the data, decide how to proceed with further analysis, *etc...* When searching for interesting patterns – generally not knowing *a priori* exactly what to look for – in BD, and if one wants to react quickly, EDA is becoming a new and necessary processing paradigm [116]:

> *"As Big Data and statistics engage with one another, it is critical to remember that the two fields are united by one common goal: to draw reliable conclusions from available data."*

With new developments in open technologies and recent breakthroughs in data collection and management techniques, data modelling, simulation and optimisation sciences, machine learning and pattern recognition methods, the opportunities of EDA – and KDD – for improving efficiency and effectiveness appear in particular limitless [270]. Still, BD also needs *"sound judgment"* [272] as an abundance of data and computing power does not automatically guarantee good decision-making [173]:

> *"The challenge with big data is to cope with abundance and exhaustivity (including sizable amounts of data with low utility and value), timeliness and dynamism, messiness and uncertainty, high relationality, semi-structured or unstructured content, and the fact that much of them are generated with no specific question in mind or are a by-product of another activity."*

In particular, it is expected that the large amounts of crowd-sourced data available – *e.g.*, from social media – will provide information – possibly real-time – to integrated policy models on human, societal, economic and environmental processes and systems [171, 236, 218]. In this context, we suggest to first focus on (big and not so big) data exploration using suitable exploratory and descriptive techniques – *e.g.* summary statistics, clustering, anomaly detection, *etc...* – and visualisation tools made available through open technologies [240]

with the scope of seeking interpretations of data that make reasonable and logical sense[127], but that are not definitive in their claim [64]:

> *"When researchers approach a data set, they need to understand – and publicly account for – not only the limits of the data set, but also the limits of which questions they can ask of a data set and what interpretations are appropriate."*

Therefore, it is necessary to provide an integrated BD application solution that spans the life cycle of data analysis – *i.e.*, the DVC: data collection, storage, processing, and visualisation. For that purpose, we rely on a strong justified framework and practical applicable recommendations.

6.2 Open framework for deployment of computational resources

In 2012, the (former) EC adopted a scientific information package consisting of a communication *"Towards better access to scientific information: Boosting the benefits of public investments in research"* [10] that set up a framework for improving access to – as well as preservation of – (publicly funded) scientific information. In that sense, the grand vision of *Digital Science*[128] (DS) – now also referred to as *Open Science* [9] – implies for research processes to be more open, more collaborative, and closer to society [96]:

> *"Various aspects of Digital Science improve its trustworthiness and availability of science for policy making. Furthermore, and more importantly, empowering citizens to access, understand and participate in scientific processes enhances their acceptance of policies which are based on scientific evidence and which enable citizen participation in their development and monitoring."*

The ultimate goal of DS is in fact *Open Knowledge*, *i.e.*, to open up science and its results, for everyone to see, participate, use, and benefit from, using digital technologies[129]. For the goals of communicating results in computational science and data analysis, interoperability between publications, analyses, data, models and tools (software) is required in a research institution [175] and needs to be supported [153]:

> *"Research funding bodies should commission research and development on tools that enable code to be integrated with other elements of scientific research such as data, graphical displays and the text of an article. [They] should provide metadata repositories that describe both programs and data produced by researchers."*

[127] *"Sure, Big Data is great. But so is intuition"*: http://www.nytimes.com/2012/12/30/technology/big-data-is-great-but-dont-forget-intuition.html?ref=technology&_r=0.

[128] Digital Science is about the way research is carried out, disseminated, deployed and transformed by digital tools, networks and media: http://ec.europa.eu/digital-agenda/digital-science.

[129] *"Now digital technology and tools offer the chance for a new transformation: improving research and innovation and making them more relevant for citizens and society"*, as *"Science 2.0 is revolutionising the way we do science – from analysing and sharing data and publications to cooperating across the globe. It is also allowing citizens to join in the search for new knowledge [so that] the whole scientific process is becoming more transparent and efficient [...]."*, citing respectively former Commission Vice-President N. Kroes and European Research, Innovation and Science Commissioner M. Geoghegan-Quinn in press release on *"Have your say on the future of science: public consultation on Science 2.0"* (http://europa.eu/rapid/press-release_IP-14-761_en.htm). At this stage, it is not clear yet if this vision is still shared by the *'Connected Digital Single Market Strategy'* of the new EC [20]

Because JRC is both a research organisation and a service of the European Commission, it is critical for our computational resources to be developed with the same rigor, open access and review as the scientific evidence it intends to support [262] – hence, *"leading by example"* [10]. In particular, when considering the many policies and actions aiming at – implicitly or explicitly – promoting DS: *Open Access*[130] (OA) – which facilitates access to scientific results and data, *Global Systems Science*[131] (GSS) – which promotes scientific evidence provision to support policy-making, *Interoperability Solutions for Europe*[132] (ISA) programme – which support the sharing of existing tools, common specifications, standards and solutions and encourages collaborative development of solutions between public organisations, *Collective Awareness Platforms for Sustainability and Social Innovation*[133] (CAPS) – which aim at creating awareness of problems and possible solutions requesting collective efforts, enabling new forms of social innovation, *Citizen Science*[134] (CS) – which supports science outreach activities for society's benefit, we should be able to transpose (implement) their recommendations – made at European levels – to our own context. We believe, in particular, that the principle of openness that applies to JRC should go beyond providing open access to academic publications, analyses and data [217].

It is generally recognised that publications and analyses alone might not be the right way to make the results of a research accessible to practitioners in computational science [50]. In particular, 'reproducibility' (see Figure 7) further requires that researchers make data available and accessible to others, as well as the original code used in the research so that the data can be analysed in a similar manner [217]. In this context, we introduce the following requirements for future computational resources and research software to be developed or deployed in-house:

- *open*: this aspect is indeed crucial for reusability and reproducibility (see also below), it also ensures transparency: [217]:

  > *"[A]nyone doing any computing in their research should publish their code. It does not have to be clean or beautiful, it just needs to be available. Even without the corresponding data, code can be very informative and can be used to check for problems as well as quickly translate ideas. [...] Non-open source software can only be changed by their owners, who may not perceive reproducibility as a high priority."*

- *reproducible*[135]: anyone should be able to reproduce[136] the experiments and computational workflow that lead to given scientific outcomes [113]. Researchers across a

[130] https://ec.europa.eu/digital-agenda/en/open-access-scientific-information; see also Section 2.

[131] http://global-systems-science.eu/gss/content/gss-portal.

[132] http://ec.europa.eu/isa/.

[133] http://ec.europa.eu/digital-agenda/en/collective-awareness-platforms.

[134] http://ec.europa.eu/digital-agenda/en/news/green-paper-citizen-science-europe-towards-society-empowered-citizens-and-enhanced-research-0.

[135] *"Making scientific computations reproducible"*: http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible:cip.

[136] Following the definition of [153], we mean by 'reproducibility' the *"reproduction of a scientific paper's [project's, experience's] central finding, rather than exact replication of each specific numerical result down to several decimal places"*. Indeed, successfull reproduction may vary widely depending on the data and the types of experiments performed [95].

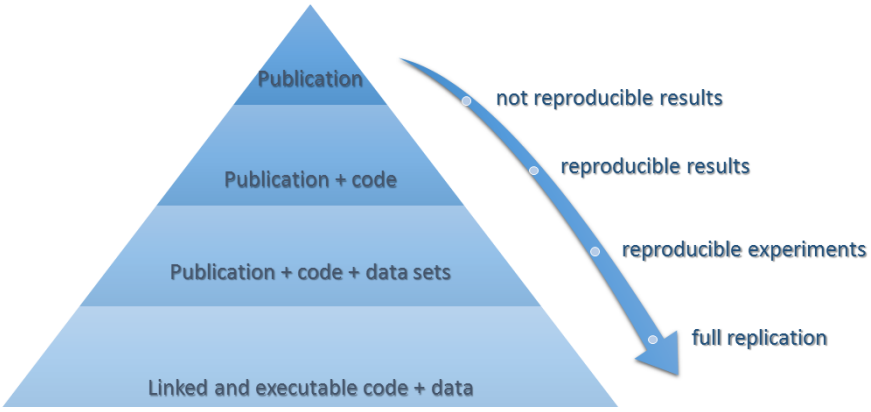


Figure 7: Spectrum of reproducibility. Inspired by [217].

range of computational science disciplines have been calling for reproducibility, or reproducible research, as an attainable minimum standard for assessing the value of scientific claims [262]. As for the software [217]:

> *"A critical barrier to reproducibility in many cases is that the computer code is [not] available. Interactive software systems often used for exploratory data analysis typically do not keep track of users' actions in any concrete form."*

In fact, it can be argued that [153]:

> *"With some exceptions, anything less than the release of source programs is intolerable for results that depend on computation. The vagaries of hardware, software and natural language will always ensure that exact reproducibility remains uncertain, but withholding code increases the chances that efforts to reproduce results will fail."*

- *verifiable*: given certain assumptions – *i.e.*, hypotheses and models – it should be possible not only to reproduce the experiments [217]:

  > *"The fact that an analysis is reproducible does not guarantee the quality, correctness, or validity of the published results. [...] In cases in which questionable results are obtained, reproducibility is critical to tracking down the 'bugs' of computational science."*

  but also to fully test and validate the methodology – and implementation – used to produce the scientific outcomes of a research [210]:

  > *"No research paper can ever be considered to be the final word, and the replication and corroboration of research results is key to the scientific process."*

- *collaborative*: computational resources should facilitate the participation of all and the integration of any additional scientific contribution [48]:

  > *"Freely provided working code – whatever its quality – improves programming and enables others to engage with your research."*

Above all, it could help adopting more efficient ways of working and support effective interdisciplinary research [255]:

> *"[T]he research world tends to purpose-build, application-specific software for each discipline, of limited long-term value without a consistent computer science perspective, rather than devise common requirements that permit re-use of software across many disciplines. It is instead necessary to develop generic technology, to support the whole research cycle. These will permit to follow new paths, try new techniques, build new models and test them in new ways. The result: innovative multidisciplinary and interdisciplinary work."*

- *participatory*: computational resources should also allow contribution from external communities – *e.g.*, scientific, but possibly any other pre-defined/meritocratic criterion – as well [156]:

  > *"Advances in ICT provide the key to an exciting prospect, that of actively engaging the public in the process of creation of the understanding."*

  It is claimed that new forms of participatory knowledge production can contribute to new models of interaction between citizens, authorities and scientists to engage in policy-making processes [24]:

  > *"Participants provide experimental data and facilities for researchers, raise new questions and co-create a new scientific culture. While adding value, volunteers acquire new learning and skills, and deeper understanding of the scientific work in an appealing way. As a result of this open, networked and trans-disciplinary scenario, science-society-policy interactions are improved leading to a more democratic research based on evidence-informed decision making."*

Table 3 shows how this approach maps the strategic needs of the DS agenda. We believe the deployment of computational resources under the above-mentioned requirements would enable to track the totality of decision-making processes and their progress as well. In addition, it is also important to ensure – as a set of best practices – that the technical expertise is not drained off, whatever the reason(s) for it[137]. Beyond enabling the reproduction and the verification of the outcomes of the research[138], the practical benefits are in allowing code reuse. Even if there is no guarantee of quality, it can still potentially contribute to new experiments and help develop/deploy more advanced analysis methods in-house [48]. In doing so, we should have in mind to specifically comply with rules 1: *"for every result, keep track of how it was produced"*, and 10: *"provide public access to scripts, runs, and results"* of the Ten Rules for reproducible computational research [232], while it is also consistent with rule 6:

[137] *E.g.*, because the resources can not be located, because the principal investigator/operator has moved, because the underlying infrastructure has changed, *etc...* See also the *"data entropy"* diagram that describes the decrease of knowledge over time: http://innet-project.eu/sites/default/files/rda-pewi-innet_0.ppt.

[138] *"What the Reinhart & Rogoff debacle really shows: verifying empirical results needs to be routine"*: http://blog.stodden.net/2013/04/19/what-the-reinhart-rogoff-debacle-really-shows-verifying-empirical-results-needs-to-be-routine/.

Table 3: Proposal for core requirements for the development of computational resources and research software at JRC. We conceptually map these requirements to global European initiatives: OA, GSS, CAPS, ISA, *and* CS *which aim at supporting Digital Science to "make scientific processes more efficient, transparent and effective by new tools for scientific collaboration, experiments and analysis and by making scientific knowledge more easily accessible".*

		open	verifiable	reproducible	collaborative	participatory
OA	*provide researchers, businesses and citizens with improved and free of charge online access to EU-funded research results, including scientific publications and research data.*	✓	✓	✓		
GSS	*provide scientific evidence to support policy-making, public action and civic society to collectively engage in societal action.*		✓	✓		✓
CAPS	*harness collective intelligence for taking better informed and sustainability-aware decisions.*	✓			✓	✓
CS	*foster the interaction between the Citizen Science stakeholders and the EU policy officers.*	✓			✓	✓
ISA	*foster interoperability between public administrations by helping to establish common approaches that will make collaboration a lot easier; [s]haring and reusing tools such as common platforms and common components, along with the sharing of services like common infrastructures.*	✓		✓	✓	

"publish your code (even the small bits)" of the Ten Rules for the care and feeding of scientific data [129]. However, it is legitimate[139] to question if it is really necessary to build our own BD system/platform while commercial-off-the-shelf (COTS) software are available on the market. Still, we believe there is a need for 'research grade' and independent computational resources, hence open source – though also susceptible to downsides – is the right approach. In this aspect, open technologies are a guarantee of openness and adaptability, and presents very tangible benefits over traditional COTS – 'closed-source', 'black-box' – technologies. There is moreover a need for building internal skills, that can also be supported by open

[139] *"8 considerations when selecting big data technology"*: http://www.computerworld.com/article/2475840/big-data/8-considerations-when-selecting-big-data-technology.html.

technologies. With an open source (software) solution, any skilled person can modify the code to suit his/her needs, learn from its use and further contribute to its improvement. In an open source environment, there is a process and a set of positive incentives for useful modifications to be shared back to the benefit of the entire community. In addition, the use of open source can facilitate data and metadata exchange, thus using contributions from external users [24]. In this aspect, it is worth mentioning again the UN Global Pulse initiative that pursues a strategy that consists in developing innovative methods and techniques for Big Data by assembling free and open source technology toolkits. Finally, it is necessary to highlight the fact that the BD phenomenon is intrinsically related to open source initiatives, as large companies like Google, Facebook, Yahoo!, Twitter, LinkedIn benefit and contribute working on open source projects[140].

6.3 Practical recommendations for implementation of Big Data application software

Several challenging issues accompany the practical implementation – which requires specific focus – of a (big) data analysis solution, *e.g.* in modelling and programming, in handling, managing and distributing data, in processing and running the execution of analyses, to mention a few. To facilitate experimentation in our research activities, we envision the implementation of a collaborative high-level research-grade application software:

- by 'application software', we refer to a BD system that encompasses the application layer [iii] in the sense of [151] (see also Figure 5),
- by 'research-grade', we require a high degree of flexibility to integrate and deploy new algorithms or functionalities very quickly onto the software,
- by 'high-level', we express the need to have a tool that enables to focus on the application itself, without having to manage low-level technical aspects – *e.g.*, depending on the underlying infrastructure, similar to [215] approach.

However, a BD system should not only meet the needs of research applications, but also fulfils some of the common requirements for production environment. In particular, the Information Society Technologies Advisory Group (ISTAG) identified important challenges for software engineering research and innovation, and formulated appropriate related recommendations [132]. More generally, these views have been emphasised by many different groups in the EU – such as the NESSI platform [29] – so as to provide input to the *Horizon 2020* research work programme [2]. In following some of the recommendations shown in Figure 4, we suggest to early adopt some best practice principles for implementation. Say it otherwise, components of the proposed solution will cope with the CHASTER issues, while also addressing the crosscutting challenges related to the DVC (see Figure 8) so as to *"strengthen software engineering capacity and the number of skilled software engineers"* [29].

As stated in Section 4, the main issues for a BD system solution are related to interoperability, availability, scalability, efficiency, and to some extent, real-time processing

[140] See various examples mentioned in Footnotes 93 to 117

Table 4: Recommendations for software implementation, found in the report on the future of software engineering [29] by the NESSI (top) and the report on software technologies by the ISTAG [132] (bottom) resp.

Rec-2.2	*understand the consequences of different implementation alternatives (e.g., quality, robustness, performance, maintenance, evolvability, ...).*
Rec-2.3	*encourage the emergence of open source software repositories associated with development or qualification tools to gather and foster the result of cooperative R&D or local initiatives.*
Rec-3.1	*manage and hide the heterogeneity of existing and future parallel architectures from the programmers, [...] develop programs without basic knowledge of their runtime targets.*
Rec-3.2	*address technical challenges like elasticity, scalability and data management systems and aim at easing the uptake of cloud services.*
Rec-6.1	*issues of flexibility, security, integrity, portability and migration must be natural ingredients of any software development ecosystem.*
Rec-7	*develop new scientific foundations, system design methodologies, development processes and tools to create the technical solutions tackling the challenges posed by system complexity (e.g. system behavior, dynamic growth, availability, fault tolerance, safety and security).*

(a) Recommendations identified in [132].

Innov-1	*leveraging software creation as key enabler for innovation creation.*
Skills-1	*fostering software engineering skill building during research and innovations activities.*
BestPr-1	*perform industry-near research exploiting real-world software engineering cases.*
BestPr-2	*deliver well-documented, working software tools and pilots to make project outcomes more accessible.*
BestPr-3	*larger-scale, integrated projects – in addition to small ones – important for software engineering research.*
BestPr-4	*if pursuing an open source strategy, it needs to be done early on and by leveraging existing ecosystems.*

(b) Recommendations identified in [29].

(timeliness). We further believe it is essential for a BD software solution to define clear constraints on the DVC – *e.g.*, by relaxing some of the DVC requirements – to enable effective implementation and deployment. We explicitly formulate some of these constraints:

- *interoperability*: the challenge is manifold particularly when considering the mix of structured/unstructured data, qualitative and quantitative data. Beyond data collection, data integration – prior to data fusion – of heterogeneous and dynamic data needs to be supported;
- *accessibility*: possibly continuous operations need to be achieved with no human intervention through distributed data management and storage – because issues related to data transmission from source to storage or analysis components need to be considered, *e.g.*, through dynamic scheduling and memory management;
- *scalability* and *timeliness*: parallel implementation enables to perform HPC on mul-

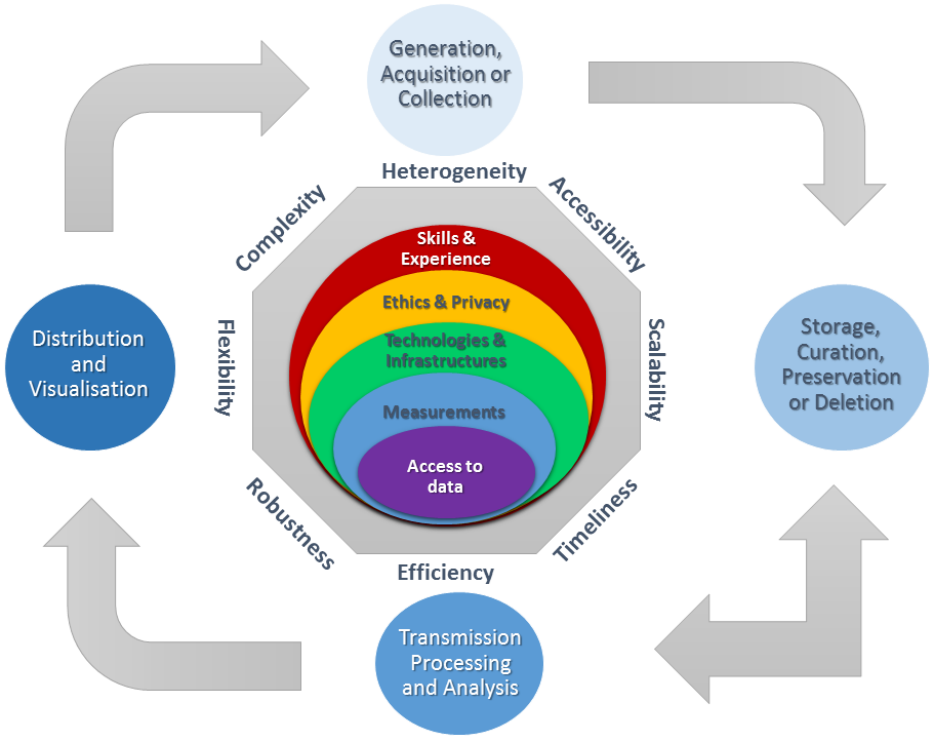


Figure 8: Representation of a BD system aiming at BD exploration. Implementation issues that need to be addressed by an application software: the DVC raises both CHASTER *– in addition to a requirement on flexibility – issues along the chain – here, a cycle – and crosscuting challenges. See also Sections 4 and 5, and Figure 6 as well.*

ticore processors or run MPP analyses on distributed, heterogeneous and parallelised platforms, *e.g.* DFS or clusters. It is required not only for effective data accessibility, but, naturally, also for scaling – when handling various heterogeneous resources to increase throughput, and real-time stream processing – where specific parallelisation techniques and automated distribution of tasks are also crucial elements. Other issues regard latency requirements;

- *efficiency* and *robustness*: the ultimate goal consists in enabling a distributed computation with tasks running effectively on heterogeneous data. A combination of the computing paradigms – batch or streaming – is needed as neither alone is sufficient to satisfy all kinds of data and all types of application requests.

In our opinion, one key feature of the application software should be, in particular, its capability to handle complex BD, by combining stream processing – *e.g.*, to deal with the velocity of the data – with distributed computing – *e.g.*, to deal with its volume – and open source – *e.g.*, for the variety. Beyond the CHASTER requirements – parts of the 'intelligent design' mentioned in the Preface of this report – we further introduce the following key requirement for an application software, as it supports an 'evolutionary design':

- *flexibility and evolvability*: development (implementation) happens without basic knowl-

edge of the runtime targets[141]. In particular, a software solution should tailor applications to the underlying infrastructure – layer [i] in the sense of [151] – at deployment and run time. Another challenge is to allow portability of both functionality and performance on the whole range of computing layers [ii]. It should indeed be flexible *w.r.t* database systems and programming models – such as the MR paradigm. It needs to implicitly support any parallel programming model, *e.g.* through parallel algorithmic skeletons that can be composed. Additional requirements are the ability to integrate or combine open source and other third-party comp[onents (libraries) so that new functionalities emerge out.

From our perspective, raising the level of abstraction is also essential for enabling flexibility. Moreover, a BD system will most likely change during its lifetime, also requiring the application software to be flexible and evolvable, so as to support – hardware and middleware – current technological developments as well as emerging technologies.

7 Conclusion: walk the talk

As we laid the foundations for an open, verifiable, reproducible, collaborative, and participatory deployment and implementation of a BD application solution, we already started the development of a software platform (so-called KINK*i*[142]) for – big and not so big – data exploration that we intend to make open – *i.e.*, through an open license – and available – *i.e.*, through shared code repository. This way:

- we support the current need to direct research efforts towards developing highly scalable and autonomous data management and processing tools,
- we want to promote internal – at the level of our institution – collaboration through common/shared implementation and experimentation.

In doing so, we also demonstrate that the guidelines we defined – *i.e.*, the above framework as well as the practical recommendations based on DVC-related issues – are applicable in our context, that is we *'walk the talk'*. Gaining experience from software deployment – this one, or any other solution – is also essential to address the DVC crosscutting challenges on skills creation and experience sharing for *"boosting digital skills and learning"* [20]. Under that aspect, the KINK*i* solution is a high-level research-grade application software that addresses most of the CHASTER requirements, though relaxing some of the associated constraints in the first stages of its implementation. For instance, it relaxes the high availability requirement: we will not deal with system failures – a business-oriented issue – in the first place. The software should be flexible enough to enable later implementation of this feature as plug-in's. Another aspect regards privacy: we do not intend to take it initially into account. as we believe privacy-enhancing features can also be incorporated later on onto the software. However, we shall still have overlook on legislation, regulation, and associated issues that may impact the implementation or use of the considered technologies. After all, we contend

[141] See again infrastructure/computing layers examples mentioned in Footnotes 93 to 117

[142] KINK*i* is a recursive acronym which stands for 'KINK*i* Is Not Kowalski*i*'.

that scientific evidence provision to support policy making cannot be operated using data(-informed)-driven approach only. To provide insight into global and complex – *e.g.*, human, societal, economic and environmental – processes and systems, a more comprehensive – hence collaborative – approach is required: it should take multiple experts from different domains to really understand what is going on.

References

[1] Guide to cost-benefit analysis of investment projects. Technical report, Directorate General Regional Policy, 2008. Structural Funds, Cohesion Fund and Instrument for Pre-Accession. Available from: http://ec.europa.eu/regional_policy/sources/docgener/guides/cost/guide2008_en.pdf. [pp. 17 and 24]

[2] Playing to win in the new software market: Software 2.0 – Winning for Europe. Technical report, Industry Expert Group on a European Software Strategy – European Commission, 2009. Available from: ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/ssai/European_Software_Strategy.pdf. [p. 53]

[3] A Digital Agenda for Europe. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, May 2010. COM(2010)245. Available from: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52010DC0245R%2801%29&from=EN. [p. 19]

[4] Europe 2020 – A strategy for smart, sustainable and inclusive growth. Communication from the Commission, March 2010. Available from: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52010DC2020&from=EN. [pp. 20 and 35]

[5] Open data – An engine for innovation, growth and transparent governance. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, December 2011. COM(2011) 882. Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF. [pp. 20 and 35]

[6] Big Data: A new world of opportunities. Technical report, Networked European Software and Services Initiative – European Commission, 2012. Available from: http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf. [pp. 24, 31, 32, 33, 36, and 40]

[7] Big Data for development: Opportunities & challenges. Technical report, United Nations Global Pulse, 2012. http://www.unglobalpulse.org/projects/BigDataforDevelopment. Available from: http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf. [p. 19]

[8] Data equity – Unlocking the value of big data. Technical report, Center for Economics and Business Research – SAS UK, 2012. Available from: http://www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf. [pp. 16, 19, 21, and 33]

[9] Open Science for the 21st century – A declaration of ALL European academies. Technical report, ALLEA – The European Federation of National Academies of Sciences and Humanities, 2012. Available from: http://www.allea.org/Content/ALLEA/General%20Assemblies/General%20Assembly%202012/OpenScience%20Rome%20Declaration%20final_web.pdf. [p. 48]

[10] Towards better access to scientific information: Boosting the benefits of public investments in research. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, July 2012. COM(2012) 401. Available from: http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf. [pp. iii, 35, 48, and 49]

[11] Work programme 2013 : Theme 3 – Information and Communications Technologies. European Commission, 2012. C(2012)4536. Available from: http://cordis.europa.eu/fp7/ict/docs/ict-wp2013-10-7-2013.pdf#page=65. [pp. 21 and 30]

[12] Data-Driven Innovation – A guide for policymakers: Understanding and enabling the economic and social value of data. Technical report, Public Policy Division – Software and Information Industry Association, 2013. Available from: http://www.siia.net/index.php?option=com_docman&task=doc_download&gid=4279&Itemid=318. [pp. 16, 17, 19, and 21]

[13] Exploring data-driven innovation as a new source of growth. Mapping the policy issues raised by Big Data. Technical Report 222, Committee for Information, Computer and Communications Policy – Organisation for Economic Co-operation and Development, 2013. DSTI/ICCP(2012)9. doi:10.1787/5k47zw3fcp43-en. Available from: http://www.oecd-ilibrary.org/science-and-technology/exploring-data-driven-innovation-as-a-new-source-of-growth_5k47zw3fcp43-en. [pp. 16, 19, 24, 29, 34, and 35]

[14] What does Big Data mean for official statistics? Technical report, High Level Group for Modernisation of Statistical Production and Services – United Nations Economic Commission For Europe, 2013. Available from: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614. [pp. 25 and 26]

[15] A-Z Guide to the Joint Research Centre. Technical Report EUR26806, European Commission – Joint Research Centre, 2014. doi:10.2788/11668. Available from: http://www.cc.cec/dgintranet/jrc/intranet/news/2014/documents/jrc_general_2014-11-12_pubsy_web_02.pdf. [pp. 18 and 25]

[16] Big Data Analytics: Towards a European research agenda. Technical report, Expert Group on Big Data Analytics – ERCIM, 2014. Available from: https://rd-alliance.org/sites/default/files/attachment/BigDataAnalytics_white_paper_v9.pdf. [pp. 16, 18, 21, 24, 25, and 41]

[17] Big Data and privacy: A technological perspective. Technical report, President's Council of Advisors on Science and Technology – Executive Office of the President, 2014. Available from: http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf. [pp. 35 and 36]

[18] Big data – An opportunity or a threat to official statistics? Technical Report ECE/CES/2014/32, Eurostat – United Nations – Economic and Social Council, 2014. Available from: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2014/32-Eurostat-Big_Data.pdf. [pp. 25 and 35]

[19] Big Data: Seizing opportunities, preserving values. Technical report, President's Council of Advisors on Science and Technology – Executive Office of the President, 2014. Available from: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf. [pp. 16, 19, and 36]

[20] Commission Work Programme 2015 – A new start. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, December 2014. COM(2014)910. Available from: http://ec.europa.eu/atwork/pdf/cwp_2015_en.pdf. [pp. 20, 41, 48, and 56]

[21] Data-driven innovation for growth and well-being: Interim synthesis report. Technical report, Organisation for Economic Co-operation and Development, 2014. Available from: http://www.oecd.org/sti/inno/data-driven-innovation-interim-synthesis.pdf. [pp. 16, 22, 25, 34, and 46]

[22] European Big Data value strategic research & innovation agenda. Technical report, Big Data Value Association – European Commission, 2014. Available from: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=7151. [pp. 19, 20, 22, 24, 25, 29, 30, 31, 32, 33, 34, 35, 41, and 42]

[23] The future of Europe is Science. Technical report, Science and Technology Advisory Council – European Commission, 2014. doi:10.2796/28973. Available from: http://ec.europa.eu/commission_2010-2014/president/advisory-council/documents/the_future_of_europe_is_science_october_2014.pdf. [p. 40]

[24] Green Paper on Citizen Science. Technical report, Socientize Consortium, 2014. Available from: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=4121. [pp. 18, 25, 51, and 53]

[25] How big is Big Data? Exploring the role of Big Data in Official Statistics. Technical report, High Level Group for Modernisation of Statistical Production and Services – United Nations Economic Commission For Europe, 2014. Available from: http://www1.unece.org/stat/platform/download/attachments/99484307/Virtual%20Sprint%20Big%20Data%20paper.docx. [pp. 25, 34, and 41]

[26] JRC directors' task force "Big Data". Technical report, Joint Research Centre of the European Commission, July 2014. Internal document. [pp. iii, 42, 43, and 44]

[27] The promise of the EU. Technical report, Eurobarometer Qualitative Study – Directorate General for Communication, September 2014. Available from: http://ec.europa.eu/public_opinion/archives/quali/ql_6437_en.pdf. [p. 17]

[28] Research and innovation as sources of renewed growth. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, June 2014. COM(2014) 339. Available from: http://ec.europa.eu/research/innovation-union/pdf/state-of-the-union/2013/research-and-innovation-as-sources-of-renewed-growth-com-2014-339-final.pdf. [pp. i and iii]

[29] Software engineering: Key enabler for innovation. Technical report, Networked European Software and Services Initiative – European Commission, 2014. Available from: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=6771. [pp. 31, 32, 33, 40, 53, and 54]

[30] Towards a thriving data-driven economy. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, July 2014. COM(2014) 442. Available from: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014DC0442&from=EN. [pp. i, iii, 16, 17, 18, 19, 20, 35, and xci]

[31] Global mobile data traffic forecast update, 2013-2018. CISCO Visual Networking Index, February. 2014. Available from: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf. [p. 13]

[32] M.R. Abbott. A new path for science? In Hey et al. [149]. Available from: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part3_abbott.pdf. [p. 42]

[33] A. Acquisti and R. Gross. Predicting social security numbers from public data. *Proceedings of the National Academy of Sciences*, 106(27):10975–10980, 2009. doi:10.1073/pnas.09048. Available from: http://www.pnas.org/content/106/27/10975.full.pdf. [p. 36]

[34] A. Adamov. Distributed file system as a basis of data-intensive computing. In *Proc. International Conference on Application of Information and Communication Technologies*, pages 1–3, 2012. doi:10.1109/ICAICT.2012.6398484. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6398484. [p. 38]

[35] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom. Challenges and opportunities with Big Data – A community white paper developed by leading researchers across the United States. Technical report, Computing Research Association, 2012. Available from: http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf. [pp. 24, 27, 31, 32, and 33]

[36] T. Akidau, A. Balikov, K. Bekiroğlu, S. Chernyak, J. Haberman, R. Lax, S. McVeety, D. Mills, P. Nordstrom, and S. Whittle. Millwheel: Fault-tolerant stream processing at internet scale. *Proceedings of the VLDB Endowment*, 6(11):1033–1044, 2013. doi:10.14778/2536222.2536229. Available from: http://research.google.com/pubs/archive/41378.pdf. [p. 39]

[37] A. Alexandrov, R. Bergmann, S. Ewen, J.-C. Freytag, F. Hueske, A. Heise, O. Kao, M. Leich, U. Leser, V. Markl, F. Naumann, M. Peters, A. Rheinländer, M.J. Sax, S. Schelter, M. Höger, K. Tzoumas, and D. Warneke. The Stratosphere platform for Big Data analytics. *The VLDB Journal*, 23(6):939–964, 2014. doi:10.1007/s00778-014-0357-y. Available from: http://stratosphere.eu/assets/papers/2014-VLDBJ_Stratosphere_Overview.pdf. [p. 39]

[38] S. Alsubaiee, Y. Altowim, H. Altwaijry, A. Behm, V.R. Borkar, Y. Bu, M. Carey, I. Cetindil, M. Cheelangi, K. Faraaz, E. Gabrielova, R. Grover, Z. Heilbron, Y.-S. Kim, C. Li, G. Li,

J. Mahn Ok, N. Onose, P. Pirzadeh, V. Tsotras, R. Vernica, J. Wen, and T. Westmann. AsterixDB: A scalable, open source BDMS. *Proceedings of the VLDB Endowment*, 7(14):1905–1916, 2014. Available from: https://asterixdb.ics.uci.edu/pub/p1289-alsubaiee.pdf. [p. 39]

[39] C. Anderson. The end of theory: The data deluge makes the scientific method obsolete. Wired Magazine, 2008. Available from: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory. [p. 42]

[40] P.L. Andrade, J. Hemerly, G. Recalde, and P. Ryan. From Big Data to big social and economic opportunities: Which policies will lead to leveraging data-driven innovation's potential? In *The Global Information Technology Report 2014* [58], chapter 1.8, pages 81–86. Available from: http://www3.weforum.org/docs/GITR/2014/GITR_Chapter1.8_2014.pdf. [p. 16]

[41] G.E. Andrews. Drowning in the data deluge. *Notice of American Mathematical Society*, 59(7):933–941, 2012. doi:10.1090/noti871. Available from: http://www.ams.org/notices/201207/rtx120700933p.pdf. [pp. 23 and 44]

[42] R. Appuswamy, C. Gkantsidis, D. Narayanan, O. Hodson, and A. Rowstron. Scale-up vs scale-out for Hadoop: Time to rethink? In *Proc. Annual Symposium on Cloud Computing*, 2013. doi:10.1145/2523616.2523629. Available from: http://research.microsoft.com/pubs/204499/a20-appuswamy.pdf. [p. 37]

[43] N. Askitas and K.F. Zimmermann. Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(3):107–120, 2009. doi:10.3790/aeq.55.2.107. Available from: http://www.iza.org/highlights/manage_highlights/docs/97_GoogleEconometrics_AEQ_2009.pdf. [p. 27]

[44] V. Backaitis. Faking Big Data #strataconf. CMS Wire, October 2014. Available from: http://www.cmswire.com/cms/big-data/faking-big-data-strataconf-026866.php. [p. 44]

[45] W. Baker, F. Kiewell, and G. Winkler. Using Big Data to make better pricing decisions. McKinsey & Company – Insights & Publications, June 2014. Available from: http://www.mckinsey.com/insights/marketing_sales/using_big_data_to_make_better_pricing_decisions. [pp. 16 and 26]

[46] J. Baliga, R.W.A. Ayre, K. Hinton, and R.S. Tucker. Green cloud computing: Balancing energy in processing, storage, and transport. *Proceedings of the IEEE*, 99(1):149–167, 2011. doi:10.1109/JPROC.2010.2060451. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5559320. [p. 37]

[47] E. Barbierato, M. Gribaudo, and M. Iacono. Performance evaluation of NoSQL Big-Data applications using multi-formalism models. *Future Generation Computer Systems*, 37:345–353, 2014. doi:10.1016/j.future.2013.12.036. Available from: http://www.sciencedirect.com/science/article/pii/S0167739X14000028. [p. 38]

[48] N. Barnes. Publish your computer code: it is good enough. *Nature*, 467:753, 2010. doi:10.1038/467753a. Available from: http://www.nature.com/news/2010/101013/pdf/467753a.pdf. [pp. 50 and 51]

[49] S. Beardsley, L Enríquez, F. Grijpink, S. Sandoval, S. Spittaels, and M. Strandell-Jansson. Building trust: The role of regulation in unlocking the value of Big Data. In *The Global Information Technology Report 2014* [58], chapter 1.7, pages 73–80. Available from: http://www3.weforum.org/docs/GITR/2014/GITR_Chapter1.7_2014.pdf. [p. 29]

[50] S. Beecham, P. O'Leary, S. Baker, I. Richardson, and J. Noll. Making software engineering research relevant. *IEEE Computer*, 47(4):80–83, 2014. doi:10.1109/MC.2014.92. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6798581. [pp. 45 and 49]

[51] R. Bekkerman, M. Bilenko, and J. Langford. *Scaling up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press, 2011. Related presentation available from http://hunch.net/~large_scale_survey/SUML.pdf. [pp. 40 and 41]

[52] M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, and M.-L. Sbodio. AllAboard: A system for exploring urban mobility and optimizing public transport using cellphone data. In H. Blockeel, K. Kersting, S. Nijssen, and F. Železnỳ, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 663–666. Springer, 2013. doi:10.1007/978-3-642-40994-3_50. Available from: http://www.ecmlpkdd2013.org/wp-content/uploads/2013/07/651.pdf. [p. 27]

[53] T. Berners-Lee. The next web. TED Talk, February 2009. Available from: http://www.ted.com/talks/tim_berners_lee_on_the_next_web. [p. 13]

[54] J.C. Bertot, U. Gorham, P.T. Jaeger, L.C. Sarin, and H. Choi. Big Data, open government and E-government: Issues, policies and recommendations. *Information Polity*, 19(1-2):5–16, 2014. doi:10.3233/IP-140328. Available from: http://iospress.metapress.com/content/u7085882v4606644/. [p. 18]

[55] J.C. Bertot, P.T. Jaeger, and J.M. Grimes. Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, 27(3):264–271, 2010. doi:10.1016/j.giq.2010.03.001. Available from: http://www.sciencedirect.com/science/article/pii/S0740624X10000201. [p. 17]

[56] T. Bicer, J. Yin, D. Chiu, G. Agrawal, and K. Schuchardt. Integrating online compression to accelerate large-scale data analytics applications. In *Proc. International Symposium on Parallel & Distributed Processing*, pages 1205–1216, 2013. Available from: http://web.cse.ohio-state.edu/~agrawal/allpapers/ipdps2013_tb.pdf. [p. 37]

[57] A. Bifet, G. Holmes, B. Pfahringer, P. Kranen, H. Kremer, T. Jansen, and T. Seidl. MOA: Massive Online Analysis, a framework for stream classification and clustering. *Journal of Machine Learning Research*, 11:1601–1604, 2010. Source available from http://moa.cms.waikato.ac.nz/. Available from: http://moa.cs.waikato.ac.nz/wp-content/uploads/2010/09/MOA-Framework.pdf. [p. 40]

[58] B. Bilbao-Osorio, S. Dutta, and B. Lanvin. The Global Information Technology Report 2014 – Rewards and risks of Big Data. Technical report, World Economic Forum, 2014. Available from: http://www3.weforum.org/docs/WEF_GlobalInformationTechnology_Report_2014.pdf. [pp. 19, 24, 63, 64, 69, 79, and 80]

[59] J. Bleiholder and F. Naumann. Data fusion. *ACM Computing Surveys*, 41(1):1:1–1:41, 2009. doi:10.1145/1456650.1456651. Available from: http://www.bioinf.jku.at/teaching/ss2012/se-inf/data_fusion_a1-bleiholder.pdf. [p. 35]

[60] M. Boehm, S. Tatikonda, B. Reinwald, P. Sen, Y. Tian, D. Burdick, and S. Vaithyanathan. Hybrid parallelization strategies for large-scale machine learning in SystemML. *Proceedings of the VLDB Endowment*, 7(7):553–564, 2014. Available from: http://www.vldb.org/pvldb/vol7/p553-boehm.pdf. [p. 40]

[61] V.R. Borkar, M.J. Carey, , and C. Li. Big Data platforms: What's next? *ACM Crossroads*, 19(1):44–49, 2012. doi:10.1145/2331042.2331057. Available from: https://asterixdb.ics.uci.edu/pub/XRDS-Big-Data-ASTERIX.pdf. [pp. 14, 38, and 39]

[62] K. Borne. Statistical truisms in the age of Big Data. Statistics View, June 2013. Available from: http://www.statisticsviews.com/details/feature/4911381/Statistical-Truisms-in-the-Age-of-Big-Data.html. [pp. 23 and 29]

[63] d. boyd and K. Crawford. Six provocations for Big Data. In *Proc. Symposium on the Dynamics of the Internet and Society*, 2011. doi:10.2139/ssrn.1926431. Available from: http://softwarestudies.com/cultural_analytics/Six_Provocations_for_Big_Data.pdf. [pp. 35 and 44]

[64] d. boyd and K. Crawford. Critical questions for Big Data – Provocations for a cultural, technological, and scholarly phenomenon. *Communication and Society*, 15(5):662–679, 2012. doi:10.1080/1369118X.2012.678878. Available from: http://www.tandfonline.com/doi/pdf/10.1080/1369118X.2012.678878. [pp. iii, 14, 23, 24, 26, 28, 29, 42, and 48]

[65] E. Brewer. CAP twelve years later: How the "rules" have changed. *IEEE Computer*, 45(2):23–29, 2012. doi:10.1109/MC.2012.37. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6133253. [p. 38]

[66] G. Brumfiel. Down the petabyte highway. *Nature*, 469:282–283, 2011. doi:10.1038/469282a. Available from: http://www.nature.com/news/2011/110119/pdf/469282a.pdf. [p. 16]

[67] R.E. Bryant. Data-intensive scalable computing for scientific applications. *Computing In Science & Engineering*, 13(6):25–33, 2011. doi:10.1109/MCSE.2011.73. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5953577. [p. 38]

[68] Y. Bu, B. Howe, M. Balazinska, and M.D. Ernst. HaLoop: Efficient iterative data processing on large clusters. *Proceedings of the VLDB Endowment*, 3(1-2):285–296, 2010. doi:10.14778/1920841.1920881. Available from: http://www.ics.uci.edu/~yingyib/papers/HaLoop_camera_ready.pdf. [p. 38]

[69] D. Butler. When Google got flu wrong. *Nature*, 494(7436):155–156, 2013. doi:10.1038/494155a. Available from: http://www.nature.com/polopoly_fs/1.12413!/menu/main/topColumns/topLeftColumn/pdf/494155a.pdf. [p. 28]

[70] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6):599–616, 2009. doi:10.1016/j.future.2008.12.001. Available from: http://www.sciencedirect.com/science/article/pii/S0167739X08001957. [p. 38]

[71] S. Campa, M. Danelutto, M. Goli, H. González-VÍez, A.M. Popescu, and M. Torquati. Parallel patterns for heterogeneous CPU/GPU architectures: Structured parallelism from cluster to cloud. *Future Generation Computer Systems*, 37:354–366, 2014. doi:10.1016/j.future.2013.12.038. Available from: http://www.sciencedirect.com/science/article/pii/S0167739X14000041. [p. 40]

[72] J. Cannarella and J.A. Spechler. Epidemiological modeling of online social network dynamics, 2014. arXiv:1401.4208. Available from: http://arxiv.org/pdf/1401.4208v1.pdf. [p. 28]

[73] M.J. Carey. BDMS performance evaluation: Practices, pitfalls, and possibilities. In R. Nambiar and M. Poess, editors, *Selected Topics in Performance Evaluation and Benchmarking*, volume 7755 of *Lecture Notes in Computer Science*. Springer, 2013. doi:10.1007/978-3-642-36727-4_8. Available from: https://asterixdb.ics.uci.edu/pub/tpctc12.pdf. [p. 38]

[74] N. Carr. The limits of social engineering. ACM Opinion – MIT Technology review, April 2014. Review of Pentland's book "Social Physics: How Good Ideas Spread". Available from: http://cacm.acm.org/opinion/articles/174334-the-limits-of-social-engineering. [p. 19]

[75] Y. Carrière-Swallow and F. Labbé. Nowcasting with Google Trends in an emerging market. Technical Report 588, Central Bank of Chile, 2010. Available from: http://www.bcentral.cl/estudios/documentos-trabajo/pdf/dtbc588.pdf. [p. 27]

[76] F.H. Cate. Government data mining: The need for a legal framework. *Harvard Civil Rights - Civil Liberties Law Review*, 43:435–489, 2008. Available from: http://www.law.harvard.edu/students/orgs/crcl/vol43_2/435-490_Cate.pdf. [p. 36]

[77] J.L. Cervera, D. Fazio, M. Scannapieco, R. Brennenraedts, and T. van der Vorst. Big Data in official statistics. Technical report, Eurostat, 2014. "ESS Big Data Event Rome 2014" Workshop. Available from: http://www.cros-portal.eu/sites/default/files//Big%20Data%20Event%202014%20-%20Technical%20Final%20Report%20-finalV01_0.pdf. [pp. 34 and 41]

[78] B. Chandramouli, J. Goldstein, and S. Duan. Temporal analytics on Big Data for web advertising. In *Proc. IEEE International Conference on Data Engineering*, pages 90–101, 2011. doi:10.1109/ICDE.2012.55. Available from: http://research.microsoft.com/pubs/150002/TiMR-TR.pdf. [p. 40]

[79] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R.E. Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems*, 26(2), 2008. doi:10.1145/1365815.1365816. Available from: http://research.google.com/archive/bigtable-osdi06.pdf. [p. 38]

[80] B. Chattopadhyay, L. Lin, W. Liu, S. Mittal, P. Aragonda, V. Lychagina, Y. Kwon, and M. Wong. Tenzing: A SQL implementation on the MapReduce framework. *Proceedings of the VLDB Endowment*, 4(12):1318–1327, 2011. Available from: http://research.google.com/pubs/archive/37200.pdf. [p. 39]

[81] M.H. Chen, R. Craiu, F. Liang, and C. Liu. Statistical and computational theory and methodology for Big Data analysis. Technical report, Banff International Research Station

for Mathematical Innovation and Discovery, 2014. Available from: https://www.birs.ca/workshops/2014/14w5086/report14w5086.pdf. [p. 40]

[82] V. Cherkassky and F. Mulier. *Learning From Data: Concepts, Theory, and Methods*. John Wiley & Sons, Inc., 2007. 2nd edition. Available from: http://www.planta.cn/forum/files_planta/learn_from_data_concepts_theory_and_methods_831_430.pdf. [pp. 28 and 42]

[83] H. Choi and H. Varian. Predicting the present with Google trends. *Economic Record*, 88:2–9, 2012. doi:10.1111/j.1475-4932.2012.00809.x. Available from: http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf. [p. 27]

[84] J. Cohen, B. Dolan, M. Dunlap, J. Hellerstein, and C. Welton. MAD skills: New analysis practices for Big Data. *Proceedings of the VLDB Endowment*, 2(2):1481–1492, 2009. doi:10.14778/1687553.1687576. Available from: http://db.cs.berkeley.edu/jmh/papers/madskills-032009.pdf. [p. 41]

[85] J.C. Corbett, J. Dean, M. Epstein, A. Fikes, C. Frost, J.J. Furman, S. Ghemawat, A. Gubarev, C. Heiser, P. Hochschild, W. Hsieh, S. Kanthak, E. Kogan, H. Li, A. Lloyd, S. Melnik, D. Mwaura, D. Nagle, S. Quinlan, R. Rao, L. Rolig, Y. Saito, M. Szymaniak, C. Taylor, R. Wang, and D. Woodford. Spanner: Google's globally distributed database. *ACM Transactions on Computer Systems*, 31(3):1–22, 2013. doi:10.1145/2491245. Available from: http://research.google.com/archive/spanner-osdi2012.pdf. [p. 38]

[86] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Series in Telecommunications and Signal Processing. Wiley-Interscience, 2006. 2nd edition. [p. 23]

[87] M. Craglia, F. Ostermann, and L. Spinsanti. Digital Earth from vision to practice: Making sense of citizen-generated content. *International Journal of Digital Earth*, 5(5):398–416, 2012. doi:10.1080/17538947.2012.712273. Available from: http://www.tandfonline.com/doi/full/10.1080/17538947.2012.712273. [p. 27]

[88] M. Craglia and L. Shanley. Data democracy – Increased supply of geospatial information and expanded participatory processes in the production of data. *International Journal of Digital Earth*, 2015. doi:10.1080/17538947.2015.1008214. Available from: http://www.tandfonline.com/doi/pdf/10.1080/17538947.2015.1008214. [pp. 14 and 18]

[89] K. Crawford. The anxieties of Big Data. The New Inquiry, May 2014. Available from: http://thenewinquiry.com/essays/the-anxieties-of-big-data/. [p. 23]

[90] A. Croll. Big Data is our generation's civil rights issue, and we don't know it: What the data is must be linked to how it can be used. Radar O'Reilly, August 2012. Available from: http://radar.oreilly.com/2012/08/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it.html. [p. 30]

[91] P.J.H. Daas, M.J. Puts, B. Buelens, and P.A.M. van den Hurk. Big Data and official statistics. In *Proc. of New Techniques and Technologies for Statistics*, 2013. Working paper for the UNECE. Available from: http://www.cros-portal.eu/sites/default/files//NTTS2013fullPaper_76.pdf. [pp. 25, 26, and 41]

[92] G. De Francisci Morales. SAMOA: A platform for mining Big Data streams. In *Proc. International Conference on World Wide Web Companion*, pages 777–778, 2013. Available from: http://melmeric.files.wordpress.com/2013/04/samoa-a-platform-for-mining-big-data-streams.pdf. [p. 40]

[93] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. *Communications of ACM*, 51(1):107–113, 2008. Available from: http://research.google.com/archive/mapreduce-osdi04.pdf. [p. 38]

[94] J. Dean and S. Ghemawat. MapReduce: A flexible data processing tool. *Communications of ACM*, 53(1):72–77, 2010. doi:10.1145/1629175.1629198. Available from: http://dl.acm.org/ft_gateway.cfm?id=1629198&ftid=693562&dwn=1&CFID=598592163&CFTOKEN=62105266. [p. 38]

[95] K. Diethelm. The limits of reproducibility in numerical simulation. *IEEE Computing in Science Engineering*, 14(1):64–72, 2012. doi:10.1109/MCSE.2011.21. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5719578. [p. 49]

[96] Content & Technology Directorate General for Communications Networks. Digital Science in Horizon 2020 – Concept paper of the Digital Science vision, 2013. Available from: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=2124. [p. 48]

[97] H. Do, B. Doherty, A. Annunziato, and M. Atkinson. An early warning system in support of humanitarian emergency preparedness. Technical Report JRC69918, European Commission – Joint Research Centre, 2012. Available from: http://skp.jrc.cec.eu.int/skp/scientific_outputs/scientificOutput/download.do?documentId=59988. [p. 27]

[98] C. Doulkeridis and K. Nørvåg. A survey of large-scale analytical query processing in MapReduce. *The VLDB Journal*, 23(3):355–380, 2014. doi:10.1007/s00778-013-0319-9. Available from: https://www.idi.ntnu.no/~noervaag/papers/VLDBJ2013_MapReduceSurvey.pdf. [p. 39]

[99] S. Dustdar, Y. Guo, B. Satzger, and H.-L. Truong. Principles of elastic processes. *IEEE Internet Computing*, 15(5):66–71, 2011. doi:10.1109/MIC.2011.121. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6015579. [p. 34]

[100] Editor The Economist. Technology: The data deluge – Data, data everywhere. The Economist, February 2010. Available from: http://www.economist.com/node/15557443. [p. 13]

[101] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G. Fox. Twister: A runtime for iterative MapReduce. In *Proc. ACM International Symposium on High Performance Distributed Computing*, pages 810–818, 2010. doi:10.1145/1851476.1851593. Available from: http://www.iterativemapreduce.org/hpdc-camera-ready-submission.pdf. [p. 38]

[102] J. Ekanayake, S. Pallickara, and G. Fox. MapReduce for data intensive scientific analyses. In *IEEE International Conference on eScience*, pages 277–284, 2008. doi:10.1109/eScience.2008.59. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4736768. [p. 38]

[103] B. El-Darwiche, V. Koch, D. Meer, R.T. Shehadi, and W. Tohme. Big Data maturity: An action plan for policymakers and executives. In *The Global Information Technology Report 2014* [58], chapter 1.3, pages 43–52. Available from: http://www3.weforum.org/docs/GITR/2014/GITR_Chapter1.3_2014.pdf. [p. 15]

[104] A. Ene, S. Im, and B. Moseley. Fast clustering using MapReduce. In *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 681–689, 2011. doi:10.1145/2020408.2020515. Available from: http://www.cs.princeton.edu/~aene/papers/MapReduce-Clustering-KDD.pdf. [p. 40]

[105] J. Fan, F. Han, and H. Liu. Challenges of Big Data analysis. *National Science Review*, 2014. doi:10.1093/nsr/nwt032. Available from: http://nsr.oxfordjournals.org/content/early/2014/02/06/nsr.nwt032.full.pdf. [pp. 29, 30, and 41]

[106] U. Fayyad, G . Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 82–88, 1996. doi:10.1007/s10618-007-0067-9. Available from: http://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf. [p. 45]

[107] J.-D. Fekete. Visual analytics infrastructures: From data management to exploration. *IEEE Computer*, 46(7):22–29, 2013. doi:10.1109/MC.2013.120. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6488679. [p. 40]

[108] A.R. Ferguson, J.L. Nielson, M.H. Cragin, A.E. Bandrowski, and M.E. Martone. Big Data from small data: data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience*, 17(11):1442–1448, 2014. doi:10.1038/nn.3838. Available from: http://www.researchgate.net/profile/Anita_Bandrowski/publication/267625305_Big_data_from_small_data_data-sharing_in_the_%27long_tail%27_of_neuroscience/links/5453e7b80cf26d5090a5537b. [p. 34]

[109] P. Fiorenza, K. Long, J. Ribeira, C. Moeger, and A. Krzmarzick. Fighting waste, fraud and abuse through analytics. In *Unlocking the Power of Government Analytics* [112], pages 25–27. Available from: http://govloop.com/blogs/4001-5000/4845-AnalyticsGuide2013_final.pdf. [p. 23]

[110] P. Fiorenza, K. Long, J. Ribeira, C. Moeger, and A. Krzmarzick. Increasing transparency initiatives through analytics. In *Unlocking the Power of Government Analytics* [112], pages 21–23. Available from: http://govloop.com/blogs/4001-5000/4845-AnalyticsGuide2013_final.pdf. [p. 17]

[111] P. Fiorenza, K. Long, J. Ribeira, C. Moeger, and A. Krzmarzick. Ten steps to leveraging analytics in the public sector. In *Unlocking the Power of Government Analytics* [112], pages 29–32. Available from: http://govloop.com/blogs/4001-5000/4845-AnalyticsGuide2013_final.pdf. [pp. iv and 44]

[112] P. Fiorenza, K. Long, J. Ribeira, C. Moeger, and A. Krzmarzick. Unlocking the power of government analytics. Technical report, GovLoop, February 2013. Available from: http://govloop.com/blogs/4001-5000/4845-AnalyticsGuide2013_final.pdf. [pp. 19 and 69]

[113] J. Freire, P. Bonnet, and D. Shasha. Computational reproducibility: State-of-the-art, challenges, and database research opportunities. In *Proc. ACM SIGMOD International Conference on Management of data*, pages 593–596, 2012. Available from: http://vgc.poly.edu/~juliana/pub/freire-sigmod2012.pdf. [p. 49]

[114] A. Frias-Martinez, C. Soguero, M. Josephidou, and E. Frias-Martinez. Forecasting socioeconomic trends with cell phone records. In *Proc. Symposium on Computing for Development*, 2013. doi:10.1145/2442882.2442902. Available from: http://www.vanessafriasmartinez.org/uploads/dev13.pdf. [p. 27]

[115] V. Frias-Martinez and E. Frias-Martinez. Enhancing public policy decision making using large-scale cell phone data. United Nations Global Pulse Blog, September 2012. Available from: http://www.unglobalpulse.org/publicpolicyandcellphonedata. [p. 22]

[116] K. Fung. The pending marriage of Big Data and statistics. *Significance*, 10(4):22–25, 2013. doi:10.1111/j.1740-9713.2013.00679.x. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2013.00679.x/pdf. [p. 47]

[117] J. Gantz, D. Reinsel, and C. Arend. The digital universe in 2020: Big Data, bigger digital shadows, and biggest growths in the far East – Western Europe. Technical report, International Data Corporation, 2013. Available from: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-western-europe.pdf. [p. 13]

[118] M. Gascó-Hernández, editor. *Open Government – Opportunities and Challenges for Public Governance*. Public Administration and Information Technology. Springer, 2014. doi:10.1007/978-1-4614-9563-5. [pp. 16 and 78]

[119] D. Gayo-Avello. "I wanted to predict elections with Twitter and all I got was this lousy paper" – A balanced survey on election prediction using Twitter data, 2012. arXiv:1204.6441. Available from: http://arxiv.org/pdf/1204.6441v1.pdf. [p. 28]

[120] D. Gayo-Avello. No, you cannot predict elections with Twitter. *IEEE Internet Computing*, 16(6):91–94, 2012. doi:10.1109/MIC.2012.137, arXiv:1204.6441. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6355554. [p. 28]

[121] F. Genova, H. Hanahoe, L. Laaksonen, C. Morais-Pires, P. Wittenburg, and J. Wood. The data harvest: How sharing research data can yield knowledge, jobs and growth. Technical report, Research Data Alliance, 2014. Available from: http://europe.rd-alliance.org/sites/default/files/report/TheDataHarvestReport_%20Final.pdf. [p. 21]

[122] B. Gentile. The new factors of production and the rise of data-driven applications. Forbes, October 2011. Available from: www.forbes.com/sites/ciocentral/2011/10/31/the-new-factors-of-production-and-the-rise-of-data-driven-applications. [pp. 19 and 21]

[123] P.W. Gething and A.J. Tatem. Can mobile phone data improve emergency response to natural disasters? *PLoS Medicine*, 8(8), 2011. doi:10.1371/journal.pmed.1001085. Available from: http://www.plosmedicine.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pmed.1001085&representation=PDF. [pp. 22 and 27]

[124] S. Ghemawat, H. Gobioff, and S. Leung. The Google File System. In *Proc. ACM Symposium on Operating Systems Principles*, pages 29–43, 2003. Available from: http://research.google.com/archive/gfs-sosp2003.pdf. [p. 39]

[125] P. Gill, N. Jain, and N. Nagappan. Understanding network failures in data centers: Measurement, analysis, and implications. *ACM SIGCOMM Computer Communication Review*, 41(4):350–361, 2011. doi:10.1145/2043164.2018477. Available from: http://research.microsoft.com/en-us/um/people/navendu/papers/sigcomm11netwiser.pdf. [p. 37]

[126] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, pages 1012–1014, 2009. doi:10.1038/nature07634. Available from: www.nature.com/nature/journal/v457/n7232/pdf/nature07634.pdf. [p. 27]

[127] K. Goda and M. Kitsuregawa. The history of storage systems. *Proceedings of the IEEE*, 100:1433–1440, 2012. doi:10.1109/JPROC.2012.2189787. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6182574. [p. 38]

[128] M. Gokhale, J. Cohen, A. Yoo, and W.M. Miller. Hardware technologies for high-performance data-intensive computing. *IEEE Computer*, 41(4):60–68, 2008. doi:10.1109/MC.2008.125. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4488252. [p. 38]

[129] A. Goodman, A. Pepe, A.W. Blocker, C.L. Borgman, K. Cranmer, C. Merce, R. Di Stefano, Y. Gil, P. Groth, M. Hedstrom, D.W. Hogg, V. Kashyap, A. Mahabal, A. Siemiginowska, and A. Slavkovic. Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology*, 10(4):e1003542, 2014. doi:10.1371/journal.pcbi.1003542. Available from: http://www.ploscollections.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pcbi.1003542&representation=PDF. [p. 52]

[130] N. Grinberg, M. Naaman, B. Shaw, and G. Lotan. Extracting diurnal patterns of real world activity from social media. In *Proc. AAAI International Conference on Web and Social Media*, pages 205–214, 2013. Available from: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6087/6359. [p. 28]

[131] IBM Software Group. Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics. Technical report, IBM Corporation, October 2013. Available from: http://public.dhe.ibm.com/common/ssi/ecm/en/tiw14162usen/TIW14162USEN.PDF. [p. 24]

[132] Information Society Technologies Advisory Group. Software technologies: The missing Key Enabling Technology – Toward a strategic agenda for software technologies in Europe. Technical report, European Commission, 2014. Available from: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=6770. [pp. 40, 53, and 54]

[133] United Nations Secretary-General's Independent Expert Advisory Group. A World that counts: Mobilising the data revolution for sustainable development. Technical report, United Nations Secretary-General, 2014. Available from: http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf. [pp. 16, 19, 21, and 44]

[134] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7):1645–1660, 2013. doi:10.1016/j.future.2013.01.010. Available from: http://www.sciencedirect.com/science/article/pii/S0167739X13000241. [p. 15]

[135] B. Habegger, O. Hasan, L. Brunie, N. Bennani, H. Kosch, and E. Damiani. Personalization vs. privacy in Big Data analysis. *International Journal of Big Data*, 1(1):25–35, 2014. Available from: http://www.hipore.com/ijbd/2014/IJBD-Vol1-No1-2014-pp25-35-Habegge.pdf. [pp. 26 and 35]

[136] G. Halevi. *Research Trends – Special Issue on Big Data*, volume 30. Elsevier, 2012. Available from: http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf. [pp. 19, 72, and 74]

[137] G. Halevi and H. Moed. The evolution of Big Data as a research and scientific topic. In *Research Trends* [136], chapter 1, pages 3–6. Available from: http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf. [p. 42]

[138] A. Hall, O. Bachmann, R. Büssow, S. Gănceanu, and M. Nunkesser. Processing a trillion cells per mouse click. *Proceedings of the VLDB Endowment*, 5(11):1436–1446, 2012. doi:10.14778/2350229.2350259. Available from: http://vldb.org/pvldb/vol5/p1436_alexanderhall_vldb2012.pdf. [p. 39]

[139] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Series in Data Management Systems. Morgan Kaufmann, 2nd edition edition, 2006. [pp. iii, 45, 46, and 72]

[140] J. Han and M. Kamber. Mining object, spatial, multimedia, text, and web data. In *Data Mining: Concepts and Techniques* [139], chapter 10, pages 591–648. Available from: http://web.engr.illinois.edu/~hanj/cs512/bk2chaps/chapter_10.pdf. [p. 40]

[141] J. Han and M. Kamber. Mining stream, time-series, and sequence data. In *Data Mining: Concepts and Techniques* [139], chapter 8, pages 467–534. Available from: http://web.engr.illinois.edu/~hanj/cs512/bk2chaps/chapter_8.pdf. [p. 40]

[142] T. Harford. Big Data: Are we making a big mistake? FT Magazine, March 2014. Available from: http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz3G7xuuxwk. [p. 28]

[143] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning – Data Mining, Inference, and Prediction,*. Series in Statistics. Springer, 2009. 2nd edition. Available from: http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf. [pp. iii and 28]

[144] K.A. Hawick, P.D. Coddington, and H.A. James. Distributed frameworks and parallel algorithms for processing large-scale geographic data. *Parallel Computing*, 29(10):1297–1333, 2003. doi:10.1016/j.parco.2003.04.001. Available from: http://www.cs.brandeis.edu/~dilant/WebPage_TA160/science1.pdf. [p. 40]

[145] Q. He, F. Zhuang, J. Li, and Z. Shi. Parallel implementation of classification algorithms based on MapReduce. In J. Yu, S. Greco, P. Lingras, G. Wang, and A. Skowron, editors, *Rough Set and Knowledge Technology*, volume 6401 of *Lecture Notes in Computer Science*, pages 655–662. Springer, 2010. doi:10.1007/978-3-642-16248-0_89. Available from: http://link.springer.com/content/pdf/10.1007%2F978-3-642-16248-0_89.pdf. [p. 40]

[146] J. Heer, J.D. Mackinlay, C. Stolte, and M. Agrawala. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1189–1196, 2008. doi:10.1109/TVCG.2008.137. Available from: http://laurenwilcox.net/papers/2008-GraphicalHistories-InfoVis.pdf. [p. 40]

[147] J. Hellerstein, C. Ré, F. Schoppmann, D. Wang, E. Fraktin, A. Gorajek, K. Ng, C. Welton, X. Feng, and A. Li, K. Kumar. The MADlib analytics library or MAD skills, the SQL. *Proceedings of the VLDB Endowment*, 5(12):1700–1711, 2012. Source available from http://madlib.net. doi:10.14778/2367502.2367510. Available from: http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-38.pdf. [p. 40]

[148] J.V. Henderson, A. Storeygard, and D.N. Weil. Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028, 2012. doi:10.1257/aer.102.2.994. Available from: http://www.econ.brown.edu/faculty/David_Weil/Henderson%20Storeygard%20Weil%20AER%20April%202012.pdf. [p. 26]

[149] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft Research, 2009. Available from: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf. [pp. 42, 62, and 76]

[150] M. Hilbert and P. Lopez. The World's technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65, 2011. doi:10.1126/science.1200970. Available from: http://www.sciencemag.org/content/332/6025/60.full.pdf. [pp. 13 and 44]

[151] H. Hu, Y. Wen, T.-S. Chua, and X. Li. Toward scalable systems for Big Data analytics: A technology tutorial. *IEEE Access*, 2:652–687, 2014. doi:10.1109/ACCESS.2014.2332453. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6842585. [pp. 13, 29, 30, 36, 37, 38, 39, 40, 53, and 56]

[152] B. Hulliger, R. Lehtonen, and Münnich R. Analysis of the future research needs for official statistics. Methodologies & working papers, EuroStat, 2012. doi:10.2785/19629. Available from: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-12-026/EN/KS-RA-12-026-EN.PDF. [pp. 14 and 25]

[153] D.C. Ince, L. Hatton, and J. Graham-Cumming. The case for open computer programs. *Nature*, 482:485–488, 2011. doi:10.1038/nature10836. Available from: http://www.nature.com/nature/journal/v482/n7386/pdf/nature10836.pdf. [pp. iii, 48, 49, and 50]

[154] J.P.A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124:696–701, 2005. doi:10.1371/journal.pmed.0020124. Available from: http://www.plosmedicine.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pmed.0020124&representation=PDF. [p. 44]

[155] M. Isard and Y. Yu. Distributed data-parallel computing using a high-level programming language. In *Proc. ACM SIGMOD International Conference on Management of data*, pages 987–994, 2009. Available from: http://research.microsoft.com/pubs/102137/sigmod09.pdf. [p. 38]

[156] C.C. Jaeger, P. Jansson, S. van del Leeuw, M. Resch, J.D. Tabara, and R. Dum. GSS: Towards a research program for Global Systems Science – A summary of the orientation paper. Technical report, Global Systems Science, 2013. Available from: http://global-systems-science.eu/gss/sites/default/files/GSS%20synthesis%20paper.pdf. [pp. iii, 22, 24, and 51]

[157] J.C. Jason and S. Acharya. Towards a trusted HDFS storage platform: Mitigating threats to Hadoop infrastructures using hardware-accelerated encryption with TPM-rooted key protection. *Journal of Information Security and Applications*, 19(3):224–244, 2014. doi:10.1016/j.jisa.2014.03.003. Available from: http://www.sciencedirect.com/science/article/pii/S2214212614000155. [p. 37]

[158] K. Jeffery and L. Shubert. Complete computing: Toward information, incentive and intention – Research priorities in cloud computing, in the context of software and services, taking into account Internet of Things, Future Internet and Big Data. Technical report, Directorate General of Communications Networks, Content & Technology, 2014. See also 2012 report 'Advances in Clouds'. Available from: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=6775. [pp. 21, 31, 32, 33, and 38]

[159] D. Kahneman and A. Tversky. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430–454, 1972. doi:10.1016/0010-0285(72)90016-3. Available from: http://datacolada.org/wp-content/uploads/2014/08/Kahneman-Tversky-1972.pdf. [p. 28]

[160] K. Kambatla, G. Kollias, V. Kumar, and A. Grama. Trends in Big Data analytics. *Journal of Parallel and Distributed Computing*, 74(7):2561 – 2573, 2014. Special Issue on Perspectives on Parallel and Distributed Processing. doi:10.1016/j.jpdc.2014.01.003. Available from: http://www.sciencedirect.com/science/article/pii/S0743731514000057. [pp. 36 and 45]

[161] U. Kang, D.H. Chau, and C. Faloutsos. Pegasus: Mining billion-scale graphs in the cloud. In *Proc. International Conference on Acoustics Speech and Signal Processing*, pages 5341–5344, 2012. doi:10.1109/ICASSP.2012.6289127. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6289127. [p. 40]

[162] M. Karlberg and M. Skaliotis. Big Data for official statistics – Strategies and some initial European applications. In *Proc. of Conference of European Statisticians*, 2013. Working paper for the UNECE. Available from: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2013/mgt1/WP30.pdf. [p. 25]

[163] D.S. Katz and G. Allen. Computational & data science, infrastructure, & interdisciplinary research on university campus. In *Research Trends* [136], chapter 5, pages 13–16. Available from: http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf. [p. 41]

[164] J. Kehrer and H. Helwig. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):495–513, 2013. doi:10.1109/TVCG.2012.110. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6185547. [p. 40]

[165] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering the Information Age – Solving Problems with Visual Analytics*. Eurographics Association, 2010. Available from: http://www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf. [pp. iii, 40, and 41]

[166] D.A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In Simoff et al. [243], pages 76–90. doi:10.1007/978-3-540-71080-6_6. Available from: https://crawford.anu.edu.au/public_policy_community/content/doc/2008_Keim_Visual_analytics.pdf. [p. 40]

[167] D.A. Keim, F. Mansmann, and J. Thomas. Visual analytics: How much visualization and how much analytics? 11(2):5–8, 2009. doi:10.1145/1809400.1809403. Available from: http://www.hiit.fi/vakd09/vakdsi09keim_final.pdf. [p. 40]

[168] S. Kelling, W. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker. Data-intensive science: A new paradigm for biodiversity studies. *BioScience*, pages 613–620, 2009. doi:10.1525/bio.2009.59.7.12. Available from: http://bioscience.oxfordjournals.org/content/59/7/613.full.pdf. [p. 42]

[169] J. Kelly. Big data vendor revenue and market forecast, 2013-2017. Wikibon, February 2014. Available from: http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017. [p. 21]

[170] I. Kerr and J. Earle. Prediction, preemption, presumption: How Big Data threatens big picture privacy. *Stanford Law Review Online*, 66:65–72, 2012. Available from: http://www.stanfordlawreview.org/sites/default/files/online/topics/66_StanLRevOnline_65_KerrEarle.pdf. [p. 36]

[171] R. Kirkpatrick. A possible role for crowdsourcing at the United Nations? United Nations Global Pulse Blog, November 2010. Available from: http://www.unglobalpulse.org/blog/possible-role-crowdsourcing-united-nations. [pp. 24 and 47]

[172] R. Kitchin. Big Data, new epistemologies and paradigm shifts. *Big Data and Society*, 1, 2014. doi:10.1177/2053951714528481. Available from: http://bds.sagepub.com/content/1/1/2053951714528481.full.pdf. [pp. iii, 18, 42, 44, 45, and 47]

[173] R. Kitchin and T.P. Lauriault. Small data, data infrastructures and Big Data. Technical report, The Programmable City, 2014. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2376148. [pp. 29, 42, and 47]

[174] J. Kolodziej, S.U. Khan, L. Wang, M. Kisiel-Dorohinicki, S.A. Madani, E. Niewiadomska-Szynkiewicz, A.Y. Zomaya, and C.-Z. Xu. Security, energy, and performance-aware resource allocation mechanisms for computational grids. *Future Generation Computer Systems*, 31:77–92, 2014. Special Section: Advances in Computer Supported Collaboration: Systems and Technologies. doi:10.1016/j.future.2012.09.009. Available from: http://www.sciencedirect.com/science/article/pii/S0167739X12001823. [p. 37]

[175] D. Koop, E. Santos, P. Mates, H.T. Vo, P. Bonnet, B. Bauer, B. Surer, M. Troyer, D.N. Williams, J.E. Tohline, J. Freire, and C.T. Silva. A provenance-based infrastructure to support the life cycle of executable papers. *Procedia Computer Science*, 4:648–657, 2011. Proc. International Conference on Computational Science. doi:10.1016/j.procs.

2011.04.068. Available from: http://www.sciencedirect.com/science/article/pii/S1877050911001268. [p. 48]

[176] R.T. Kouzes, G.A. Anderson, S.T. Elbert, I. Gorton, and D.K. Gracio. The changing paradigm of data-intensive computing. *IEEE Computer*, 42(1):26–34, 2009. doi:http://doi.ieeecomputersociety.org/10.1109/MC.2009.26. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4755152. [pp. 29, 36, 37, and 42]

[177] T. Kraska, A. Talwalkar, J. Duchi, R. Griffith, M. Franklin, and M.I. Jordan. MLbase: A distributed machine learning system. In *Conference on Innovative Data Systems Research*, 2013. Available from: http://www.cs.berkeley.edu/~ameet/mlbase_cidr.pdf. [p. 40]

[178] N. Kroes. The data gold rush, March 2014. SPEECH/14/229. Available from: http://europa.eu/rapid/press-release_SPEECH-14-229_en.htm. [p. 20]

[179] V. Kumar, H. Andrade, B. Gedik, and K.-L. Wu. DEDUCE: At the intersection of MapReduce and stream processing. In *Proc. International Conference on Extending Database Technology*, pages 657–662, 2010. doi:10.1145/1739041.1739120. Available from: http://www.icdt.tu-dortmund.de/proceedings/edbticdt2010proc/edbt/papers/p0657-Kumar.pdf. [p. 39]

[180] R. Lämmel. Google's MapReduce programming model – Revisited. *Science of Computer Programming*, 70(1):1–30, 2008. doi:10.1016/j.scico.2007.07.001. Available from: http://www.sciencedirect.com/science/article/pii/S0167642307001281. [p. 38]

[181] S. Landefeld. Uses of Big Data for official statistics: Privacy, incentives, statistical challenges, and other issues. In *Proc. of International Conference on Big Data for Official Statistics*, 2014. Available from: http://unstats.un.org/unsd/trade/events/2014/Beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf. [pp. 26 and 36]

[182] H.E. Landemore. Why the many are smarter than the few and why it matters. *Journal of Public Deliberation*, 8(1):7, 2012. Available from: http://www.publicdeliberation.net/jpd/vol8/iss1/art7. [p. 18]

[183] D. Laney. 3D data management: Controlling data volume, velocity, and variety. Technical report, Application Delivery Strategies, 2001. Available from: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. [pp. 14 and 36]

[184] J. Larus and D. Gannon. Multicore computing and scientific discovery. In Hey et al. [149]. Available from: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part3_larus_gannon.pdf. [p. 40]

[185] S. LaValle, E. Lasser, R. Shockley, M.S. Hopkins, and N. Krushwitz. Big Data, analytics, and the path from insight to value. *MIT Sloan Management Review*, 52(2):21–31, 2011. Available from: http://www.ibm.com/smarterplanet/global/files/in_idea_smarter_computing_to_big-data-analytics_and_path_from_insights-to-value.pdf. [pp. iii, 16, 22, 29, 30, and 45]

[186] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of Google Flu: Traps in Big Data analysis. *Science*, 343(6176):1203–1205, 2014. doi:10.1126/science.1248506. Available from: http://www.sciencemag.org/content/343/6176/1203.full.pdf. [pp. 23 and 29]

[187] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009. doi:10.1126/science.1167742. Available from: http://www.sciencemag.org/content/323/5915/721.full.pdf. [p. 41]

[188] T.M. Lenard and P.H. Rubin. The Big Data revolution: Privacy considerations. Technical report, Technology Policy Institute, 2013. Available from: http://www.techpolicyinstitute.org/files/lenard_rubin_thebigdatarevolutionprivacyconsiderations.pdf. [p. 36]

[189] J. Leskovec, A. Rajaraman, and J.D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2011. Available from: http://www.mmds.org/. [p. 41]

[190] E. Letouzé. Mining the web for digital signals: Lessons from public health research. United Nations Global Pulse Blog, 2011. Available from: http://www.unglobalpulse.org/node/14534. [pp. 23 and 26]

[191] J. Lin. Scalable language processing algorithms for the masses: A case study in computing word co-occurrence matrices with MapReduce. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pages 419–428, 2008. Available from: http://www.aclweb.org/anthology/D08-1044. [p. 40]

[192] J. Lin and C. Dyer. *Data-Intensive Text Processing with MapReduce*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers, 2010. Available from: https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf. [p. 40]

[193] J. Lin and D. Ryaboy. Scaling Big Data mining infrastructure: The Twitter experience. *ACM SIGKDD Explorations Newsletter*, 14(2):6–19, 2012. doi:10.1145/2481244.2481247. Available from: http://www.kdd.org/sites/default/files/issues/14-2-2012-12/V14-02-02-Lin.pdf. [pp. 34 and 40]

[194] S. Lohr. The age of Big Data. The New York Times, February 2012. Available from: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0. [p. 13]

[195] G. Lopez and W. St. Amand. Discovering global socio economic trends hidden in Big Data. United Nations Global Pulse Blog, July 2012. Available from: http://www.unglobalpulse.org/discoveringtrendsinbigdata-CHguestpost. [p. 26]

[196] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J.M. Hellerstein. Distributed graphLab: A framework for machine learning and data mining in the cloud. *Proceedings of the VLDB Endowment*, 5(8):716–727, 2012. doi:10.14778/2212351.2212354. Available from: http://graphlab.org/files/vldb2012-low-gonzalez-kyrola-bickson-guestrin-hellerstein.pdf. [p. 41]

[197] D.J. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003. [pp. 23 and 28]

[198] G. Malewicz, M. Austern, A. Bik, J. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. Pregel: A system for large-scale graph processing. In *Proc. ACM SIGMOD International Conference on Management of data*, 2010. doi:10.1145/1807167.1807184. Available from: http://kowshik.github.io/JPregel/pregel_paper.pdf. [p. 38]

[199] J. Manyika, M. Chui, J. Bughin, B. Brown, R. Dobbs, C. Roxburgh, and A.H. Byers. Big Data: The next frontier for innovation, competition and productivity. Technical report, McKinsey Global Institute, 2011. Available from: http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx. [pp. 13, 14, 16, 21, and 36]

[200] N. Marz and J. Warren. *Big Data – Principles and Best Practices of Scalable Realtime Data Systems*. Manning Publications, 2012. Available from: http://www.manning.com/marz/. [p. 39]

[201] C.A. Mattman. Cultivating a research agenda for data science. *Journal of Big Data*, 1(6), 2014. doi:doi:10.1186/2196-1115-1-6. Available from: http://www.journalofbigdata.com/content/pdf/2196-1115-1-6.pdf. [pp. 41 and 42]

[202] V. Mayer-Schönberger and C. Cukier. *Big Data: A Revolution That Will Transform How We Live Work and Think*. John Murray, London, UK, 2013. Watch interview at http://www.booktv.org/Program/14393/Big+Data+A+Revolution+That+Will+Transform+How+We+Live+Work+and+Think.aspx. [pp. 19, 35, and 42]

[203] S. Melnik, A. Gubarev, J.J. Long, G. Romer, S. Shivakumar, M. Tolton, and T. Vassilakis. Dremel: Interactive analysis of web-scale datasets. *Proceedings of the VLDB Endowment*, 3(1):330–339, 2010. doi:10.14778/1920841.1920886. Available from: http://research.google.com/pubs/archive/36632.pdf. [p. 39]

[204] R. Mikut and M. Reischl. Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):431–443, 2011. doi:10.1002/widm.24. Available from: http://zakki.dosen.narotama.ac.id/files/2012/02/Data-Mining-Tool-Reviews-March-2011.pdf. [p. 35]

[205] G. Misuraca, D. Broster, and C. Centeno. Digital Europe 2030: Designing scenarios for ICT in future governance and policy making. *Government Information Quarterly*, 29:121–131, 2012. doi:10.1016/j.giq.2011.08.006. Available from: http://www.sciencedirect.com/science/article/pii/S0740624X11000724. [pp. 17, 18, 19, 24, and 25]

[206] G. Misuraca, C. Codagnone, and P. Rossef. From practice to theory and back to practice: Reflexivity in measurement and evaluation for evidence-based policy making in the information society. *Government Information Quarterly*, 30:S68–S82, 2013. doi:10.1016/j.giq.2012.07.011. Available from: http://www.sciencedirect.com/science/article/pii/S0740624X12001530. [pp. 17 and 18]

[207] G. Misuraca, F. Mureddu, and D. Osimo. Policy-making 2.0: Unleashing the power of Big Data for public governance. In Gascó-Hernández [118], pages 171–188. doi:10.1007/978-1-4614-9563-5_11. [pp. 16, 18, 21, and 22]

[208] Editor Nature. How to encourage the right behaviour. *Nature*, 416(6876), 2002. doi:10.1038/416001b. Available from: http://www.nature.com/nature/journal/v416/n6876/pdf/416001b.pdf. [p. iii]

[209] Editor Nature. Community cleverness required. *Nature*, 455(7209), 2008. doi:10.1038/455001a. Available from: http://www.nature.com/nature/journal/v455/n7209/pdf/455001a.pdf. [p. i]

[210] Editor Nature. Challenges in irreproducible research. *Nature*, 2014. Special & supplement archive. Available from: http://www.nature.com/nature/focus/reproducibility/index.html. [p. 50]

[211] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed stream computing platform. In *Proc. IEEE International Conference on Data Mining*, pages 170–177, 2010. doi:10.1109/ICDMW.2010.172. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5693297. [p. 39]

[212] G. Newman, A. Wiggins, A. Crall, E. Graham, S. Newman, and K. and Crowston. The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, 10(6):298–304, 2012. doi:10.1890/110294. Available from: http://www.esajournals.org/doi/pdf/10.1890/110294. [p. 18]

[213] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig Latin: A not-so-foreign language for data processing. In *Proc. ACM SIGMOD International Conference on Management of data*, pages 1099–1110, 2008. doi:10.1145/1376616.1376726. Available from: http://infolab.stanford.edu/~olston/publications/sigmod08.pdf. [p. 39]

[214] T. O'Reilly. What is Web 2.0 – Design patterns and business models for the next generation of software. O'Reilly Network, September 2005. Retrieved 15/10/14. Available from: http://oreilly.com/web2/archive/what-is-web-20.html. [p. 13]

[215] P. Pääkkönen and D. Pakkala. Reference architecture and classification of technologies, products and services for Big Data systems. *Big Data Research*, 2015. doi:10.1016/j.bdr.2015.01.001. Available from: http://www.sciencedirect.com/science/article/pii/S2214579615000027. [pp. 36, 39, 40, and 53]

[216] A. Pavlo, E. Paulson, A. Rasin, D.J. Abadi, D.J. DeWitt, S. Madden, and M. Stonebraker. A comparison of approaches to large-scale data analysis. In *Proc. ACM SIGMOD International Conference on Management of data*, pages 165–178, 2009. doi:10.1145/1559845.1559865. Available from: http://www3.nd.edu/~dthain/courses/cse40771/spring2010/benchmarks-sigmod09.pdf. [p. 37]

[217] R.D. Peng. Reproducible research in computational science. *Science*, 6060(334):1226–1227, 2011. doi:10.1126/science.1213847. Available from: http://www.sciencemag.org/content/334/6060/1226.full.pdf. [pp. 49 and 50]

[218] A. Pentland. Big Data: Balancing the risks and rewards of data-driven public policy. In *The Global Information Technology Report 2014* [58], chapter 1.4, pages 53–60. Available from: http://www3.weforum.org/docs/GITR/2014/GITR_Chapter1.4_2014.pdf. [pp. iii, 16, 24, and 47]

[219] R. Pepper and J. Garrity. The Internet of Everything: How the network unleashes the benefits of Big Data. In *The Global Information Technology Report 2014* [58], chapter 1.7, pages 73–80. Available from: http://www3.weforum.org/docs/GITR/2014/GITR_Chapter1.2_2014.pdf. [pp. 15 and 25]

[220] C.L. Philip Chen and C.-Y. Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275:314–347, 2014. doi:10.1016/j.ins.2014.01.015. Available from: http://www.sciencedirect.com/science/article/pii/S0020025514000346. [pp. 18, 19, 33, 38, 40, and 45]

[221] I. Polato, R. Ré, A. Goldman, and F. Kon. A comprehensive view of Hadoop research – A systematic literature review. *Journal of Network and Computer Applications*, 46:1–25, 2014. doi:10.1016/j.jnca.2014.07.022. Available from: http://www.sciencedirect.com/science/article/pii/S1084804514001635. [p. 39]

[222] J. Polonetsk and O. Tene. Privacy and Big Data: Making ends meet. *Stanford Law Review Online*, 66:25–33, 2013. Available from: http://www.stanfordlawreview.org/sites/default/files/online/topics/PolonetskyTene.pdf. [p. 36]

[223] T. Preis, H.S. Moat, and H.E. Stanley. Quantifying trading behavior in financial markets using Google trends. *Nature Scientific Reports*, 3(1684), 2013. doi:10.1038/srep01684. Available from: http://www.nature.com/srep/2013/130425/srep01684/pdf/srep01684.pdf. [p. 27]

[224] T. Preis, H.S. Moat, H.E. Stanley, and S.R Bishop. Quantifying the advantage of looking forward. *Nature Scientific Reports*, 2:350, 2012. doi:10.1038/srep00350. Available from: http://www.nature.com/srep/2012/120405/srep00350/pdf/srep00350.pdf. [p. 26]

[225] P. Priolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. International Conference on Intelligence Analysis*, 2005. Available from: https://www.e-education.psu.edu/drupal6/files/sgam/Sense_Making_206_Camera_Ready_Paper.pdf. [p. 23]

[226] L. Probst, E. Monfardini, L. Frideres, D. Clarke, S.and Demetri, and A. Schnabel, L.and Kauffmann. Big Data – Analytics & decision making. Technical report, Directorate General for Enterprise & Industry, 2013. Business Innovation Observatory. Available from: http://ec.europa.eu/enterprise/policies/innovation/policy/business-innovation-observatory/files/case-studies/08-bid-analytics-decision-making_en.pdf. [pp. 16, 21, and 25]

[227] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Characterizing and modeling the dynamics of online popularity. *Pattern Recognition Letters*, 105:158701, 2010. doi:10.1103/PhysRevLett.105.158701. Available from: http://journals.aps.org/prl/pdf/10.1103/PhysRevLett.105.158701. [p. 27]

[228] C. Reimsbach-Kounatze. The proliferation of "Big Data" and implications for official statistics and statistical agencies: A preliminary analysis. Technical Report 245, Organisation for Economic Co-operation and Development, 2015. OECD Digital Economy Papers. doi:10.1787/5js7t9wqzvg8-en. Available from: http://www.oecd-ilibrary.org/science-and-technology/the-proliferation-of-big-data-and-implications-for-official-statistics-and-statistical-agencies_5js7t9wqzvg8-en. [pp. 25 and 29]

[229] N.M. Richards and J.H. King. Three paradoxes of Big Data. *Stanford Law Review Online*, 66:41–46, 2013. Available from: http://www.stanfordlawreview.org/sites/default/files/online/topics/66_StanLRevOnline_41_RichardsKing.pdf. [pp. 18, 26, and 36]

[230] M. Rios and J. Lin. Distilling massive amounts of data into simple visualizations: Twitter case studies. In *Workshop on Social Media Visualization*, 2012. Available from: http://www.umiacs.umd.edu/~jimmylin/publications/Rios_Lin_2012.pdf. [p. 40]

[231] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc. World Wide Web Conference*, pages 851–860, 2010. doi:10.1145/1772690.1772777. Available from: http://www.ymatsuo.com/papers/www2010.pdf. [p. 27]

[232] G.K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig. Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10):e1003285, 2013. doi:10.1371/journal.pcbi.1003285. Available from: http://www.ploscompbiol.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pcbi.1003285&representation=PDF. [p. 51]

[233] M. Scarfi. Social media and the Big Data explosion. Forbes, June 2012. Available from: http://www.forbes.com/sites/onmarketing/2012/06/28/social-media-and-the-big-data-explosion/. [p. 14]

[234] E.E. Schadt, M.D. Linderman, J. Sorenson, L. Lee, and G.P. Nolan. Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, 11:647–657, 2010. doi:10.1038/nrg2857. Available from: http://binf.gmu.edu/vaisman/binf630/nrg10-schadt.pdf. [p. 40]

[235] S. Schelter. Scaling data mining in massively parallel dataflow systems. In *Proc. ACM SIGMOD International Conference on Management of data*, pages 11–15, 2014. doi:10.1145/2602622.2602631. Available from: http://ssc.io/wp-content/uploads/2011/12/phdt33-schelter.pdf. [p. 40]

[236] H. Schoen, D. Gayo-Avello, P.T. Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor. The power of prediction with social media. *Internet Research*, 23(5):528–543, 2013. doi:10.1108/IntR-06-2013-0115. Available from: https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/sowi_lehrstuehle/politikwissenschaften_2/MANUSKRIPTE_FEB/Schoen_et_al._2013_Predicting_with_Social_Media.pdf. [p. 47]

[237] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, 2010. doi:10.1109/TVCG.2010.179. Available from: http://vis.stanford.edu/files/2010-Narrative-InfoVis.pdf. [p. 40]

[238] S. Sehgal, M. Erdelyi, A. Merzky, and S. Jha. Understanding application-level interoperability: Scaling-out MapReduce over high-performance grids and clouds. *Future Generation Computer Systems*, 27(5):590–599, 2011. doi:10.1016/j.future.2010.11.001. Available from: http://www.sciencedirect.com/science/article/pii/S0167739X10002116. [p. 40]

[239] J. Shaw. Why 'Big Data' is a big deal. Harvard Magazine, March-April 2014. Available from: http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal. [p. 14]

[240] N. Shoresh and B. Wong. Points of view: Data exploration. *Nature Methods*, 9(5), 2012. doi:10.1038/nmeth.1829. Available from: http://www.nature.com/nmeth/journal/v9/n1/pdf/nmeth.1829.pdf. [pp. 46 and 47]

[241] J.E. Short, R.E. Bohn, and C. Baru. How Much Information? 2010 – Report on Enterprise Server Information. Technical report, Global Information Industry Center, 2011. Available from: http://hmi.ucsd.edu/pdf/HMI_2010_EnterpriseReport_Jan_2011.pdf. [p. 13]

[242] N. Silver. *The Signal and the Noise: Why So Many Predictions Fail – but Some Don't.* Penguin Press HC, 2012. [p. 23]

[243] S.J. Simoff, M.H. Böhlen, and A. Mazeika, editors. *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, number 4404 in Lecture Notes in Computer Science. Springer, 2008. doi:10.1007/978-3-540-71080-6. Available from: http://download.springer.com/static/pdf/184/bok%253A978-3-540-71080-6.pdf?auth66=1416217451_c045fdd70240c25e6f3e2cba74213a8d&ext=.pdf. [pp. iii, 35, and 75]

[244] D. Singh and C.K. Reddy. A survey on platforms for Big Data analytics. *Journal of Big Data*, 2(8), 2014. doi:doi:10.1186/s40537-014-0008-6. Available from: http://www.journalofbigdata.com/content/pdf/s40537-014-0008-6.pdf. [pp. 36 and 40]

[245] A. Sorokine, A. Myers, C. Liu, P. Coleman, E. Bright, A. Rose, P. Nugent, and B. Bhaduri. Tackling Big Data: Strategies for parallelizing and porting geoprocessing algorithms to high-performance computational environments. In *Proc. International Conference on Geographic Information Science*, 2012. Available from: http://www.giscience.org/past/2012/proceedings/abstracts/giscience2012_paper_130.pdf. [p. 40]

[246] E.R. Sparks, A. Talwalkar, V. Smith, J. Kottalam, X. Pan, J. Gonzalez, M.J. Franklin, M.I. Jordan, and T. Kraska. MLI: An API for distributed machine learning. In *Proc. IEEE International Conference on Data Mining*, pages 1187–1192, 2013. doi:10.1109/ICDM.2013.158. Available from: http://www.cs.berkeley.edu/~xinghao/papers/sparks-etal.icdm2013mli1.pdf. [p. 40]

[247] S. Sreenivasan. Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords. *Nature Scientific Reports*, 3:2758, 2013. doi:10.1038/srep02758. Available from: http://www.nature.com/srep/2013/130926/srep02758/pdf/srep02758.pdf. [p. 27]

[248] S.N. Srirama, P. Jakovits, and E. Vainikko. Adapting scientific computing problems to clouds using MapReduce. *Future Generation Computer Systems*, 28(1):184–192, 2012. doi:10.1016/j.future.2011.05.025. Available from: http://www.sciencedirect.com/science/article/pii/S0167739X11001075. [p. 40]

[249] M. Stonebraker, U. Çetintemel, and S. Zdonik. The 8 requirements of real-time stream processing. In *Sigmod Rec*, volume 34/4, pages 42–47, 2005. doi:10.1145/1107499.1107504. Available from: http://cs.brown.edu/~ugur/8rulesSigRec.pdf. [p. 39]

[250] A. Stopczynski, R. Pietri, A. Pentland, D. Lazer, and S. Lehmann. Privacy in sensor-driven human data collection: A guide for practitioners, March 2014. arXiv:1403.5299. Available from: http://arxiv.org/pdf/1403.5299v1.pdf. [p. 26]

[251] W.J. Sutherland, D. Spiegelhalter, and M. Burgman. Policy: Twenty tips for interpreting scientific claims. *Nature*, 503:335–337, 2013. doi:10.1038/503335a. Available from: http://www.nature.com/polopoly_fs/1.14183!/menu/main/topColumns/topLeftColumn/pdf/503335a.pdf. [pp. 22, 23, 26, 28, and 29]

[252] C. Sweeney, L. Liu, S. Arietta, and J. Lawrence. *HIPI: A Hadoop Image Processing Interface for Image-Based MapReduce Tasks*. PhD thesis, University of Virginia, 2011. Available from: http://cs.ucsb.edu/~cmsweeney/papers/undergrad_thesis.pdf. [p. 41]

[253] A.S. Szalay. Extreme data-intensive scientific computing. *Computing In Science & Engineering*, 13(6):34–41, 2011. doi:10.1109/MCSE.2011.74. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5959140. [p. 37]

[254] O. Tene and J. Polonetsk. Privacy in the age of Big Data: A time for big decisions. *Stanford Law Review Online*, 64:63–69, 2012. Available from: http://www.stanfordlawreview.org/sites/default/files/online/topics/64-SLRO-63_1.pdf. [p. 36]

[255] C. Thanos. Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures – Final roadmap. Technical report, GRDI2020 Consortium, 2011. Available from: http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf. [pp. 18, 31, 32, 33, 40, 42, and 51]

[256] A. Thusoo, J.S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy. Hive: A warehousing solution over a Map-Reduce framework. *Proceedings of the VLDB Endowment*, 2(2):1626–1629, 2009. doi:10.14778/1687553.1687609. Available from: http://www.vldb.org/pvldb/2/vldb09-938.pdf. [p. 39]

[257] B. Tierney, E. Kissel, M. Swany, and E. Pouyoul. Efficient data transfer protocols for Big Data. In *IEEE International Conference on E-Science*, 2012. doi:10.1109/eScience.2012.6404462. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6404462. [p. 37]

[258] A. Tubke, K. Ducatel, and P. Gavivan, J.P. Moncada-Paternó-Castello. Strategic policy intelligence: Current trends, the state of play and perspectives – S&T intelligence for policy-making processes. Technical Report EUR 20137, European Commission, JRC-IPTS, European Science and Technology Observatory, 2001. Available from: http://ftp.jrc.es/EURdoc/eur20137en.pdf. [pp. 17 and 24]

[259] J.W. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962. Available from: http://www.jstor.org/stable/2237638. [pp. i and 46]

[260] J.W. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25, 1980. doi:10.2307/2682991. Available from: http://www.jstor.org/stable/2682991. [pp. iii and 46]

[261] M. Valenstein. The promise of large, longitudinal data sets. *Psychiatric Services*, 64(6):503, 2013. doi:10.1176/appi.ps.201300134. Available from: http://ps.psychiatryonline.org/doi/pdf/10.1176/appi.ps.201300134. [p. 40]

[262] J. van den Hoven, K. Jacob, L. Nielsen, F. Roure, L. Rudze, J. Stilgoe, K. Blind, A.-L. Guske, and M. Riera. Options for strengthening responsible research and innovation – Report of the expert group on the state of art in Europe on responsible research and innovation. Technical

Report EUR25766, Directorate General for Research & Innovation Science in Society, 2013. doi:10.2777/46253. Available from: http://ec.europa.eu/research/swafs/pdf/pub_public_engagement/options-for-strengthening_en.pdf. [pp. 49 and 50]

[263] R. Villars, Olofson C., and M. Eastwood. Big Data: What it is and why you should care. Technical report, International Data Corporation, 2011. Available from: http://sites.amd.com/sa/Documents/IDC_AMD_Big_Data_Whitepaper.pdf. [p. 14]

[264] Y. Wang, G. Norcie, and L.F. Cranor. Who is concerned about what? A study of American, Chinese and Indian users privacy concerns on social network sites. In *Proc. International Conference on Trust & Trustworthy Computing*, 2011. doi:10.1007/978-3-642-21599-5_11. Available from: http://www.cs.cmu.edu/~yangwan1/papers/TRUST2011-%C2%AD-AuthorCopy.pdf. [p. 36]

[265] J.S. Ward and A. Barker. Undefined by data: A survey of Big Data definitions, September 2013. arXiv:1309.5821. Available from: http://arxiv.org/pdf/1309.5821v1.pdf. [pp. 13 and 14]

[266] A. Wesolowski, C.O. Buckee, L. Bengtsson, E. Wetter, X. Lu, and A.J. Tatem. Commentary: Containing the Ebola outbreak – The potential and challenge of mobile network data. *PLoS Currents Outbreaks*, 2014. doi:10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e. Available from: http://currents.plos.org/outbreaks/article/containing-the-ebola-outbreak-the-potential-and-challenge-of-mobile-network-data/pdf/. [p. 27]

[267] W.F. Whyte. Advancing scientific knowledge through participatory action research. *Sociological Forum*, 4(3):367–385, 1989. Available from: http://www.jstor.org/stable/684609. [p. i]

[268] J. Williams. Introducing the concept of Web 3.0. Tweak And Trick, May 2012. Available from: http://www.tweakandtrick.com/2012/05/web-30.html. [p. 13]

[269] A. Wöhrer, P. Brezany, I. Janciak, and E. Mehofer. Modeling and optimizing large-scale data flows. *Future Generation Computer Systems*, 31:12–27, 2014. doi:10.1016/j.future.2013.10.004. Available from: http://www.sciencedirect.com/science/article/pii/S0167739X13002148. [p. 41]

[270] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding. Data mining with Big Data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):97–107, 2014. doi:10.1109/TKDE.2013.109. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6547630. [pp. 40, 41, and 47]

[271] Z. Wu and O.B. Chin. From Big Data to data science: A multi-disciplinary perspective. *Big Data Research*, 1:1, 2014. Special Issue on Scalable Computing for Big Data. doi:10.1016/j.bdr.2014.08.002. Available from: http://www.sciencedirect.com/science/article/pii/S2214579614000082. [p. 41]

[272] C. Yiu. The Big Data opportunity – Making government faster, smarter and more personal. Technical report, Policy Exchange, 2011. Available from: http://www.policyexchange.org.uk/images/publications/the%20big%20data%20opportunity.pdf. [pp. 16, 17, 21, 24, and 47]

[273] Y. Zhao and J. Wu. Dache: A data aware caching for Big Data applications using the MapReduce framework. In *Proc. IEEE INFOCOM*, pages 35–39, 2013. doi:10.1109/INFCOM.2013.6566730. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6566730. [p. 38]

[274] P.C. Zikopoulos, D. deRoos, K. Parasuraman, T. Deutsch, D. Corrigan, and J. Giles. *Harness the Power of Big Data: The IBM Big Data Platform*. McGraw-Hill, 2013. Available from: ftp://public.dhe.ibm.com/software/pdf/at/SWP10/Harness_the_Power_of_Big_Data.pdf. [pp. 16, 33, and 45]

[275] C. Zins. Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, 58(4):479–493, 2007. doi:10.1002/asi.20508. Available from: http://onlinelibrary.wiley.com/doi/10.1002/asi.20508/pdf. [pp. 21, 22, 23, and 45]

[276] A. Zuiderwijk and M. Janssen. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31(1):17–29, 2014. doi:10.1016/j.giq.2013.04.003. Available from: http://www.sciencedirect.com/science/article/pii/S0740624X13001202. [pp. 17 and 34]

List of Illustrations

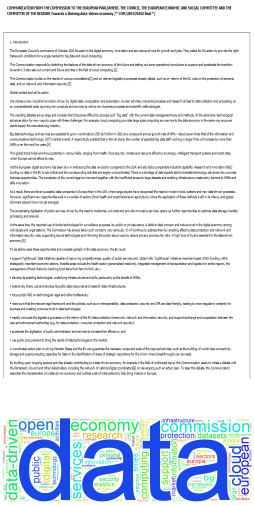
Figures

Tables

List of Acronyms

BD	Big Data
BDVA	Big Data Value Association
CAPS	Collective Awareness Platforms for Sustainability and Social Innovation
CC	Cloud Computing
CPU	Central Processing Unit
CS	Citizen Science
COTS	Commercial-off-the-shelf
DFS	Distributed File Systems
DG CONNECT	Directorate General of Communications Networks, Content & Technology
DG ESTAT	Directorate General Eurostat
DISC	Data-Intensive Scalable Computing
DISD	Data-Intensive Scientific Discovery
DS	Digital Science
EC	European Commission
EDA	Exploratory Data Analysis
EU	European Union
GDP	Gross Domestic Product
GPU	Graphics Processing Unit
GSS	Global Systems Science
HPC	High Performance Computing
ICT	Information and Communication Technologies
IEAG	Independent Expert Advisory Group
ISTAG	Information Society Technologies Advisory Group
I/O	Input/Output
IoT	Internet of Things
IPR	Intellectual Property Rights
ISA	Interoperability Solutions for European Public Administrations
KDDM	Knowledge Discovery and Data Mining
MPP	Massively Parallel Processing
MR	MapReduce
NESSI	Networked European Software and Services Initiative
OA	Open Access
OD	Open Data
OECD	Organisation for Economic Co-operation and Development
PSI	Public Sector Information
SQL	Structured Query Language
SWOT	Strengths, Weaknesses, Opportunities and Threats
UN	United Nations

DIY: generate the image of the cover as the word cloud of the European Commission communication *"Towards a thriving data-driven economy"* [30] (text available, in English, at http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52014DC0442&from=EN) using these Python lines of code[143]:




```
import urllib
import bs4
import operator

import kinki
from kinki.semantics import nlp, tags

LANG = 'en'
URL_EURLEX = 'http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=
    CELEX:52014DC0442&from=' + LANG.upper()

soup = bs4.BeautifulSoup(urllib.urlopen(URL_EURLEX).read(), "lxml")
text = soup.text.lower()
text = text.encode('ascii', errors='ignore')

STOP_WORDS = nlp.NLP.stop_words(lang=LANG) + nlp.MYSQL_EN_STOPWORDS
CLOUD_IMAGE = 'data_driven_eurlex_cloud.png'
LEN_SHORT_WORDS = 3

d_items = dict(nlp.NLP.freqdist(text,"([a-z]+(?:[-'][a-z]+)?)").items())
[d_items.pop(s) for s in d_items.keys() if s in STOP_WORDS or len(s)<
    LEN_SHORT_WORDS]
items = sorted(d_items.items(), key=operator.itemgetter(1), reverse=True)

cloudtags = tags.Cloud.create_cloud(items[:tags.DEF_MAX_TAGS],cloud=
    CLOUD_IMAGE)
```

[143] See Part II of this report.

European Commission
EUR 27094 EN – Joint Research Centre – Institute for Environment and Sustainability

Title: Collaborative research-grade software for crowd-sourced data exploration: from context to practice – Part I: Guidelines for scientific evidence provision for policy support based on Big Data and open technologies

Author(s): Jacopo Grazzini and Francesco Pantisano

Luxembourg: Publications Office of the European Union

2014 – 92 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1831-9424 (online)

ISBN 978-92-79-45377-9 (PDF)

doi: 10.2788/329540

LB-NA-27094-EN-N

JRC Mission

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new methods, tools and standards, and sharing its know-how with the Member States, the scientific community and international partners.

Serving society
Stimulating innovation
Supporting legislation

doi: 10.2788/329540

ISBN 978-92-79-45377-9

Publications Office