Lloyd's Register Foundation

Foresight review of big data

Towards data-centric engineering

December 2014

Lloyd's Register Foundation
Report Series: No.2014.2

About the Lloyd's Register Foundation

Our vision

Our vision is to be known worldwide as a leading supporter of engineering-related research, training and education, which makes a real difference in improving the safety of the critical infrastructure on which modern society relies. In support of this, we promote scientific excellence and act as a catalyst working with others to achieve maximum impact.

The Lloyd's Register Foundation charitable mission

- To secure for the benefit of the community high technical standards of design, manufacture, construction, maintenance, operation and performance for the purpose of enhancing the safety of life and property at sea, on land and in the air.
- The advancement of public education including within the transportation industries and any other engineering and technological disciplines.

About the Lloyd's Register Foundation Report Series

The aim of this Report Series is to openly disseminate information about the work that is being supported by the Lloyd's Register Foundation. It is hoped that these reports will provide insights for the research community and also inform wider debate in society about the engineering safety-related challenges being investigated by the Foundation.

T +44 (0)20 7709 9166
E info@lrfoundation.org.uk

Contents

Executive summary

Computing power and the data it generates is growing exponentially. In another decade new companies and companies we already know will have global reputations established through data analytics businesses. This unprecedented growth in data will result from ubiquitous sensors, an 'Internet of Things', that will monitor and measure our machines, our businesses, our environment and us. Big data will be everywhere - large volumes of different types of data moving at speed through a digital ecosystem. To compete in this new landscape, companies and countries alike must learn to master big data, or they will find themselves out thought, out flanked and out dated.

The Lloyd's Register Foundation will need to understand and master big data. The new scale of data availability will change all the strategic sectors in which it supports work. They will be changed because data will feature in all aspects of the business life-cycle; from design to manufacturing, maintenance to decommissioning. Data will be used to predict and anticipate, plan and decide every aspect of the 21st century workplace. This Foresight Review of Big Data, commissioned by the Foundation, provides a deliberately broad view of the impact of big data. It describes technical, organisational, social, and legal implications of living in a world of data. It presents a framework for the Foundation to think about its future - a future based in data-centric engineering.

The report reviews large-scale data analysis and describes big data methods, techniques and solutions. It provides a number of accessible examples that indicate how and why big data is making a difference here and now, including in weather forecasting and the transportation and energy sectors.

There are many perspectives on big data that need to be appreciated by those seeking to innovate responsibly. Big data is an asset and needs to be understood within the value chain of an organisation. It imposes demands on infrastructure and on humans. Big data needs analytics; not only the techniques of statistics and machine learning, but also the human skills of insight and pattern recognition to find genuine meaning in the data. Big data can be complex to analyse because it comes in many varieties, shapes and sizes and may have been collected over different timescales. It can be uncertain, noisy, and incomplete. Collective responsibility and action by citizens, governments and businesses will be needed to realise the potential that big data offers.

The report asserts that to really embrace the opportunity that big data offers the Foundation, it needs to adopt a view that we term data-centric engineering. Data-centric engineering, recognising the value of data as an asset in itself, puts data considerations at the core of engineering design. It improves performance, safety, reliability and efficiency of assets, infrastructures and complex machines. From cradle to grave, design to decommissioning, big

data analytics will feature at all phases of the life-cycle of engineered systems, and will inform new developments as part of an iterative process. Analytics will create value from a wide range of data, informing not only asset and machine performance but linking these to the physical, economic, social, and human environments in which they sit.

We have produced a set of recommendations based on this view and the context already described.

The Foundation should take a prominent role in promoting data-centric engineering. Working with others it could provide a roadmap of what this approach requires; the technology development, policy making and business models that underpin it. The Foundation is well-placed to support the innovative methodology outlined in this report within a number of engineering domains.

At the heart of the proposed methodological approach for data-centric engineering is the responsible handling of decentralised data assets. The Foundation should catalyse entrepreneurship in the data ecosystem by supporting innovative ideas and business models in its strategic sectors.

The independent role of the Foundation also establishes its suitability as an authority informing regulation and standards. The Foundation can take the leading role in formulating the codes, standards, regulation and sector-specific terms of use for data. Together with clear requirements and guidelines for continuous capture of data provenance, this is the basis for accountable and trusted data-driven supply chains. The Foundation should support data certification services for the new class of data assets that will be central to safety and security in 21st century engineering. In complex engineering we also need certification for the analytical methods and predictive models that are applied to data. Working with organisations such as the Open Data Institute the Foundation should support consultancy and data certification products in this space.

The production of data catalogues and inventories, and the ability to find these assets on the web is an essential part of data engineering and the big data landscape. The schema.org initiative is a successful example of technically lightweight data integration at web-scale that supports data discovery. The Foundation should support data catalogues, data vocabularies and data discovery methods to support its strategic sectors. It should support the provision of reliable, long-term resource identifiers for data assets.

The Foundation has to constantly engage in horizon scanning to ensure it anticipates correctly developments that could dramatically change how computing and data analytics is performed in the future.

Foreword

Lloyd's Register is built on data. The first Register, issued in 1764, provided data on ship quality and analysts of the day used this data to understand and manage shipping risks. Modern infrastructures are far more complex: advanced construction, complex supply chains, networked operations and human interventions, all set within a fast evolving technological environment, provide huge challenges for those seeking to assure safety. Big data and advanced analytical tools will address these challenges.

From cradle to grave, design to decommissioning, big data analytics will feature at all phases of the life-cycle of engineered systems, and will inform new developments as part of an iterative process. Analytics will create value from a wide range of data, informing not only asset and machine performance but linking these to the physical, economic, social, and human environments in which they sit. This comprehensive view of data in engineered systems is why we have subtitled this report 'towards data-centric engineering'. It sets the agenda for the steps that need to be undertaken for a fundamental paradigm shift and describes the role of the Lloyd's Register Foundation in making this happen.

The future value of big data will only be realised if there is organisational and cultural change, accompanied by appropriate analytical tools, skills and practices. The UK Government Chief Scientific Adviser recently called for the creation of a 'National centre to promote advanced research and translational work in algorithms and the application of data science… to enable researchers from industry and academia to work together to undertake outstanding research with practical application'[1]. Such a centre could provide a useful focus to further develop the concepts in this report.

This report describes how big data can bring the societal benefits that are at the heart of the Lloyd's Register Foundation's mission, enhancing safety by fundamentally changing the design, manufacturing, maintenance and decommissioning processes for complex infrastructures and machinery. It will help the Foundation understand where it can make a distinctive contribution to the developments in big data, in pursuit of its charitable objectives, because life matters.

Professor Sir Nigel R. Shadbolt
Professor of Artificial Intelligence
University of Southampton and
Chairman of the Open Data Institute

Professor Richard Clegg
Managing Director
Lloyd's Register Foundation

[1] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/224953/13-923-age-of-algorithms-letter-to-prime-minister_1_.pdf


DATA-CENTRIC
ENGINEERING
WORKING
COMPLEX
COST
SET
STORAGE
TOOLS
MOVING
REQUIRING
SAFETY
BIG
MPP
DESCRIBING
RELATIONAL
LLOYD'S REGISTER FOUNDATION
EXAMPLES
PETABYTES
INDEXING
CITATION
TYPES
CURRENTLY
USE
DESKTOP
GARTNER
TARGET
SHARED
DATABASES
USED
DATA
CAPACITY
DISK
GROW
CREATED
SINCE
CAPTURE
ARCHIVES
ORGANIZATIONS
TENS
NETWORKS
TOLERABLE
SOFTWARE
SOCIAL
LARGER
DEFINING
NEEDED
HUNDREDS
FORMS
NOW
TERABYTES
CONTINUES
WITHIN
TECHNOLOGIES
SETS
MANAGEMENT
COMPLEXITY
BIOGEOCHEMICAL
TIME
INCLUDE
WORLD'S
CONNECTOMICS
ZETTABYTES
ELAPSED
INTERNET
PARALLEL
SENSOR
EVERY
DIFFICULTY
CASE
UBIQUITOUS
SOLID
MANAGE
SYSTEMS
CURRENT
INFORMATION
LOGS
RECORDS
SEARCH
USING
RECONSIDER
COMPUTING
ANALYTICS

Background

This report is the second in a series commissioned by the Lloyd's Register Foundation as part of its emerging technologies research theme. It looks forward at how developments in the area of big data might impact the safety and performance of the engineered assets and the infrastructures on which modern society relies.

The Lloyd's Register Foundation is a charity and owner of the Lloyd's Register Group Limited (LR). LR is a 254 year old organisation providing independent assurance and expert advice to companies operating high-risk, capitally intensive assets primarily in the energy, maritime and transportation sectors. It also serves a wide range of sectors with distributed assets and complex supply chains such as the food, healthcare, automotive and manufacturing sectors.

Building on the findings of this review, the Foundation will look to identify aspects of big data that might provide opportunities or threats to safety in line with its charitable objectives, and where the Foundation might focus its research and other grant giving to make a distinctive positive impact.

The Foundation is a charity with a global role. Reflecting this it assembled an international and cross-sectoral expert advisory panel which met in London in July 2014. This report contains the output and findings from that panel.

Expert Panel membership

Professor Sir Nigel R. Shadbolt (Chairman)
Professor of Artificial Intelligence
University of Southampton and
Chairman of the Open Data Institute, London

Professor Mandy Chessell
Fellow of the Royal Academy of Engineering

Professor Stefan Decker
Director Digital Enterprise Research Institute
National University of Ireland, Galway

Professor Tat-Seng Chua
KITHCT Chair Professor
Department of Computer Science
School of Computing
National University of Singapore

Professor Jim Hendler
Tetherless World Senior Constellation Professor
Director of Rensselaer Institute for Data Exploration and Applications
Department of Computer Science and Cognitive Science Department
Rensselaer Polytechnic Institute
USA

Dr. Markus Luczak-Roesch
Senior Research Fellow
University of Southampton

Professor Richard Clegg (Co-Chairman)
Managing Director
Lloyd's Register Foundation

Professor Honor Powrie
Engineering and the Environment
University of Southampton

Professor Richard Stobart
Caterpillar Innovation and Research Centre
Chair of Powertrain Systems
Loughborough University

Professor Jeremy Watson
Professor of Engineering Systems
University College London

Dr. Ruth Boumphrey
Head of Research Grants,
Lloyd's Register Foundation

Introduction to big data

Computing power and the data it generates is growing exponentially. In another decade, computer storage densities will be 1,000 times greater, the data generated will be measured in zettabytes, which is a thousand billion gigabytes. Companies we know well, and others we have not yet thought of or heard of, will be global players in the field of data analytics. This trend to generate more and more data will result from ubiquitous sensors, an Internet of Things[2] that will monitor and measure our machines, our businesses, our environment and us. Big data will be everywhere - large volumes of different types of data moving at speed through our digital ecosystem. Its veracity and quality may sometimes be in doubt but unless a company, an organisation or a nation state has learnt to master big data it will very rapidly find itself out thought, out flanked and out dated.

Big data is commonly characterised as having four dimensions: volume, velocity, variety and veracity. Data that is extraordinary in one or multiple of these dimensions - very large amounts, rapidly streamed, heterogeneous and/or uncertain – may be called big data.

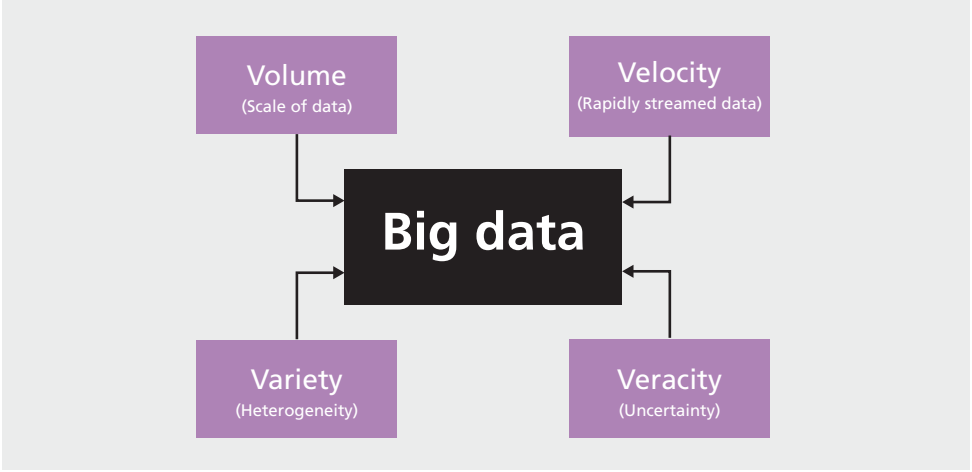


Figure 1: The four dimensions of big data.

Large data volumes are an underpinning feature of many activities. 'Big science' projects: like the Large Hadron Collider or the Square Kilometre Array; the data silos and structured data[3] of large web companies such as Google, Facebook or Twitter; the enterprise data held by companies like IBM or big retailers like Walmart; all generate terabytes of data quickly and routinely. Figure 2 illustrates relative amounts of data for such activities.

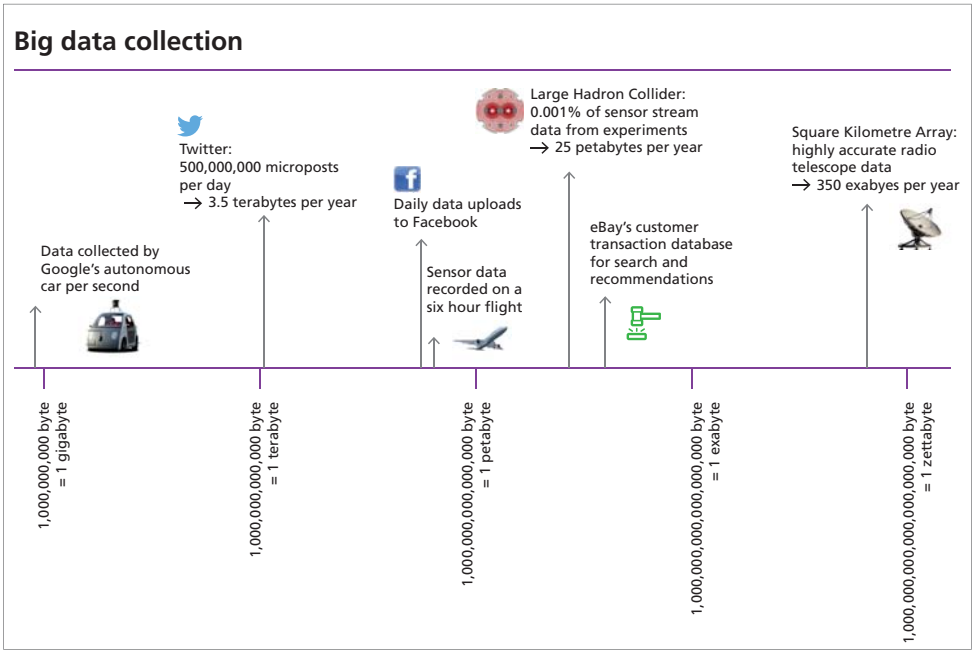


Figure 2: Examples of the scale of data collected

[2] https://www.gov.uk/government/collections/internet-of-things-review
[3] Structured data is typically regarded as data that has an associated formal schema, for example relational databases and data warehouses.

Big data and the new techniques in the way we distribute complex computing problems[4] underpin many businesses and services, providing opportunities to connect businesses to individuals in real time. Associated legal and policy frameworks are needed to support effective use and appropriate data protection and assurance and individual citizens may need to increase their understanding of their own rights and liabilities when data is collected, consumed or exposed[5]. But big data can bring huge benefits to citizens and the UN has identified its transformative potential to bring sustainable development to disadvantaged communities[6].

Increased use of big data has been underpinned by technological developments allowing us to better and faster generate, store, process, understand and visualise data. Technological advancements in automated sensing systems and user-generated inputs[7] are producing new data. Communications and computing technologies are underpinning development of near real-time applications, and policy interventions such as open standards and open data are supporting better access by those who can innovate.

Advanced sensor technologies are increasingly used on individual machines, machine components and across complex, interconnected machine systems producing a rich range and large volume of data. The analysis of the resulting big data has the potential to increase efficiency, reduce costs, improve reliability and productivity and enhance safety.

Big data underpins well known services such as the weather forecast, but has also enabled rapid growth of many novel applications, for example in the transport and energy sectors.

[4] New techniques allow distribution of complex computing problems to powerful new hardware and software solution, for example MapReduce and Hadoop.
[5] IBM. Ethics for big data and analytics, 2014, http://www.ibmbigdatahub.com/whitepaper/ethics-big-data-and-analytics
[6] A World that Counts: mobilising the data revolution for sustainable development, 2014 http://www.undatarevolution.org/report/
[7] For example smart meters and mobile devices.

Example 1: The digital oil field

Energy companies are increasingly integrating data and other asset information to improve workflow management, visualisation, monitoring, control, analytics and communications of their operations. Large seismic data sets, combined with powerful pattern recognition and rapid analysis, support exploration. Drilling data is streamed in real-time from the drill string and surface equipment throughout the drilling of a well. The industry is increasingly looking to automate these dangerous and expensive operations. Real-time analytics improve safety and efficiency, avoid equipment failures and provide geological information for critical decision support. Live data from producing and injecting wells across large oilfields allow the use of automated systems to quickly compensate for equipment failures or other incidents, balancing pressures to ensure maximum hydrocarbon recovery. Meanwhile techniques such as computational fluid dynamics (CFD) are used to improve production by modelling the complex downhole interactions between production equipment, the stressed reservoir rock and injected and produced fluids.

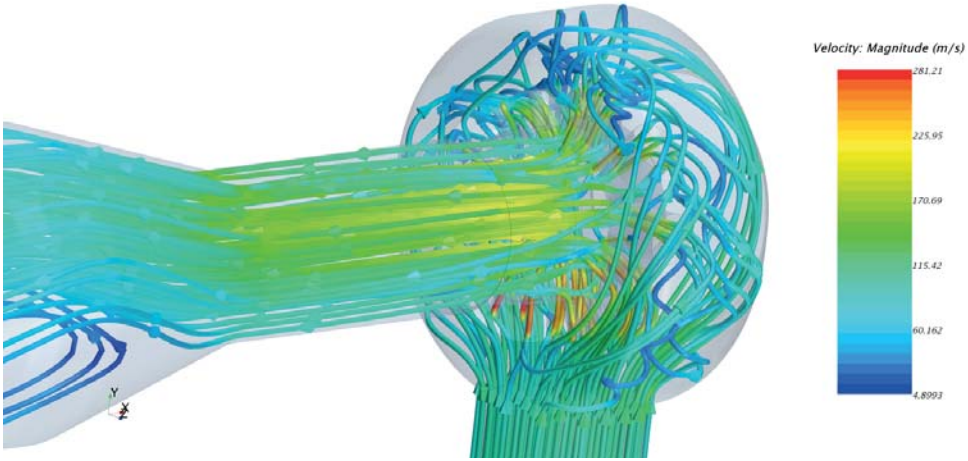
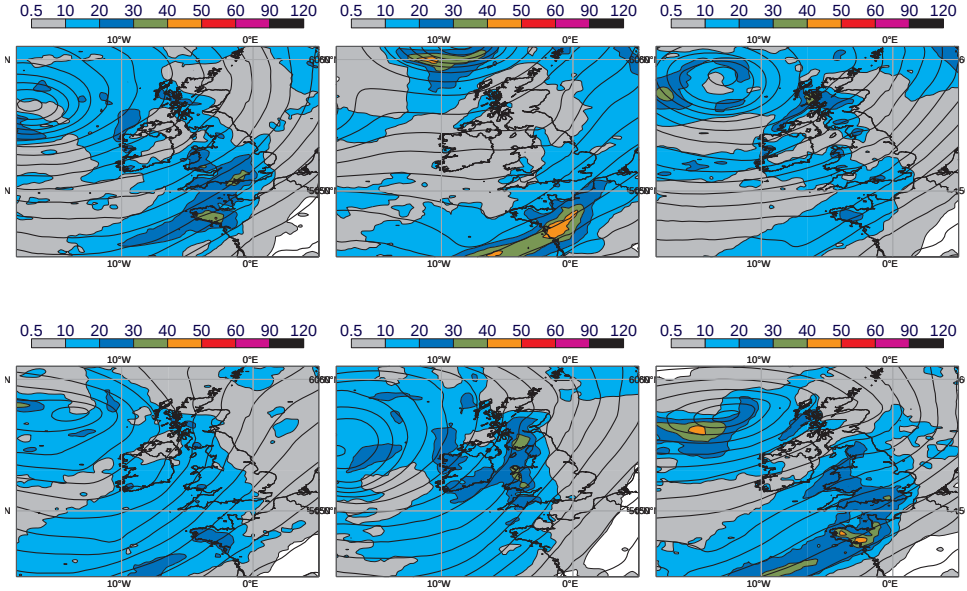


Figure 3: The image shows tracked particle velocities as gas carrying sand particles move from below up through the choke (a control on liquid flow) and on in to production system. The green represents higher velocities and the blue lower velocities. This modelling using CFD enables erosion risks in wells and production systems to be identified. Figure courtesy of LR Senergy.

Example 2: Weather forecasting

The daily weather forecast uses modelled and observed data from sensors, satellites, and in situ measurements, describing a large number of physical parameters such as barometric pressure, air and sea surface temperature, wind speed and direction, air moisture and terrain elevation. Very high performance computers are used to process huge amounts of data in near real time providing weather predictions and visualisations over different temporal and spatial scales. Advances in computational power and high resolution sensing have brought great improvements in forecast accuracy. Such improvements have brought huge benefits for individuals, businesses, and society, saving lives and property. From providing warnings for short-term events such as wind storms and floods to supporting long-term decisions on building and infrastructure design, the safety benefits of this big data application are unparalleled.



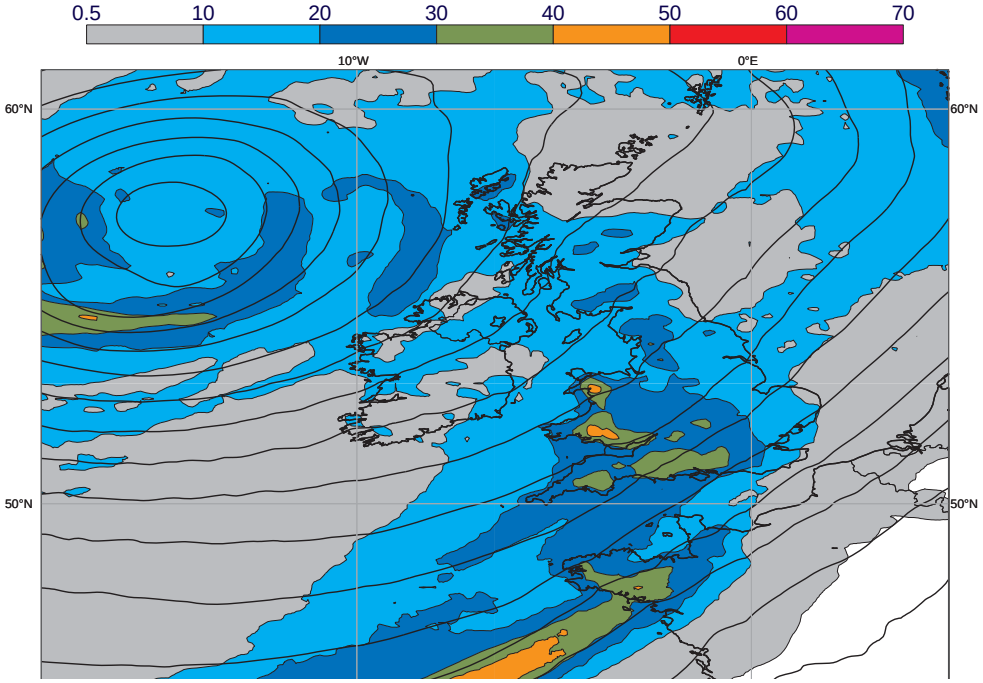


Figure 4: Ensemble forecasting uses large complex observed datasets and creates massive modelled datasets to predict the probability of different future weather events. The figure shows six out of fifty ensemble outputs, and the resulting 'most likely' forecast, predicted four days before the storms that hit the UK on 24 December (Christmas Eve) 2013, leaving 75,000 homes without power. The forecast shows pressure (contours) and precipitation (shades). Credit ECMWF.

Example 3: Transport

Big data methods open up new possibilities in transport sectors. For example, substantial data is already collected during the normal operation of rail rolling stock. Figure 5 illustrates the potential for data collection and analysis at different organisational scales of this system. Should all data be collected? How much local processing should be done? Can condition be assessed locally or by comparison between similar vehicles? Big data analytics can support systematic understanding to help in the development of new types of vehicle and fleet models, and management and design activities. Fleet data analytics also have applications in the road haulage, maritime and aviation sectors.

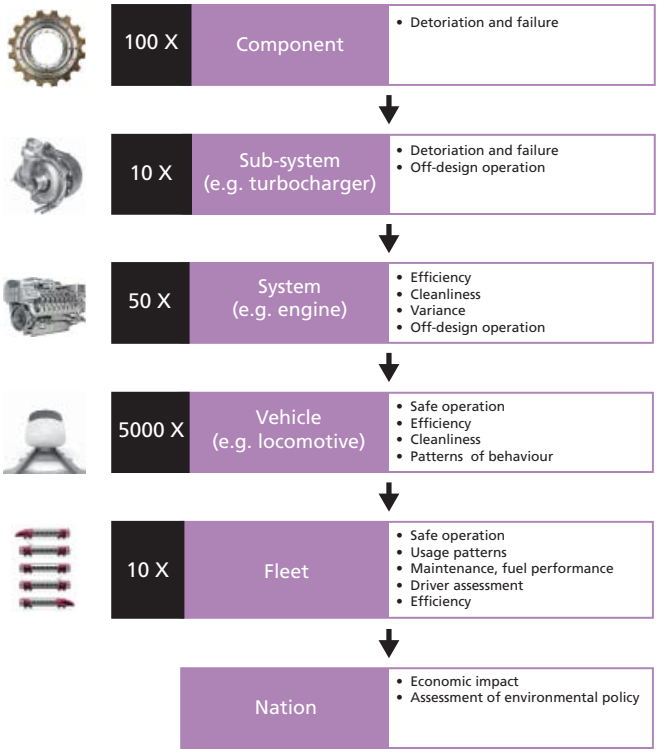


Figure 5: Data can be collected and analysed at multiple scales within a rail fleet. Credit: Richard Stobart.

Example 4: The smart grid
Smart grid technologies support intelligent management and analysis of the electrical supply based on massive amounts of sensed data, in near real time and at fine granularity. Smart grids draw on information such as behaviours of suppliers and consumers and aim to improve reliability, flexibility, and efficiency of the system. Big data is a ubiquitous theme in the research and development that has been undertaken to plan and realise this grand challenge for the power industry.

Perspectives on big data

There is no single perspective on big data. Big data is an asset and needs to be understood within the value chain of an organisation. It imposes infrastructure demands; hardware and software, human as well as physical resources. Big data needs analytics, not only the techniques of statistics and machine learning, but also the human skills of insight and pattern recognition to find genuine meaning in the data. Big data can be complex to analyse because it comes in many varieties, shapes and sizes and may have been collected over different timescales. It can be uncertain, noisy, and incomplete. Finally, to realise the potential that big data offers will require collective responsibility and action by citizens, governments and businesses. Appendix A provides further detail on these perspectives.

Data-centric engineering: implications for engineering-related disciplines

Data-centric engineering puts data considerations at the core of engineering design. It improves performance, safety, reliability and efficiency of assets, infrastructures and complex machines. From cradle to grave, design to decommissioning, big data analytics will feature at all phases of the life-cycle of engineered systems, and will inform new developments as part of an iterative process. Analytics will create value from a wide range of data, informing not only asset and machine performance but linking these to the physical, economic, social, and human environments in which they sit.

Data standards, generation, capture, annotation, storage, analysis, visualisation, security and ownership will increasingly become key parts of the modern engineering life cycle, significantly changing how design, manufacturing, maintenance and decommissioning of complex machinery and other assets will be carried out in the future. Data-centric engineering recognises that data is an asset and so there are two, co-crafted outputs from the data-centric engineering process, the physical asset and the digital asset.

Decentralised data ownership poses challenges to analysis and is a problem for some business models, but in engineering, data exchange, interoperation and analysis can lead to improved asset function and add enormous value. Intellectual property ownership will be a barrier to accountable, trusted, and controlled cross-company data infrastructures if mechanisms for rights management are not included at the design stage. But if these barriers can be overcome in engineering, it has the potential to bring positive disruptive change to other business models, sectors and society at large.

Condition-based maintenance

Advanced tagging technologies and smart materials will turn machines and vehicles into smart products with memories of each part's production and operation history. There will be a move away from fixed maintenance intervals, towards tailored predictive maintenance, reducing operator risk and providing better cost-efficiency. Predictive models will be based on in situ data collection and preferably performance data from multiple generations of design.

Big data and analytics can support maintenance planning and optimisation, and operational planning and deployment. For example machines such as ships and trains need to plan maintenance services at the right time and place. Routes might be adapted dynamically to do this effectively. Small errors or unexpected maintenance stops may have very large knock-on effects within the larger network, so it is important to consider how to build resilience into decision support systems.

Smart factories and autonomous machines

Future factories[8] will also be characterised by ubiquitous sensing and data-intense interaction with smart materials, products and machines. Planning from manufacturing to operational deployment will become highly interactive and individualised, facilitating changes in conditions along the supply chain. Smart materials and products will carry memories of their manufacturing history and plans for how they are meant to be processed next. Condition-based maintenance systems in the field will provide feedback into the production pipeline. The manufacturing of customised replacement parts will become responsive to actual requirements and allows for near real-time re-planning of production.

Advances in autonomous vehicles will impact machine-intense sectors such as mining. Humans will be taken out of the loop of actual operations and instead will become critical for rapid data-centric decision-making. This will change the skill set needed for operating complex machinery. And advanced data literacy will increasingly be a requirement in occupations operating complex machines and systems.

[8] http://www.news-sap.com/industry-4-0-two-examples-future-factory/

Data-enabled prosumers and the quantified worker

Individuals are increasingly generating data relating to individual behaviour and performance, a movement often termed the 'quantified self'. The further development and broader distribution of wearable sensors, smart watches and even smart glasses will impact how we live privately and how we work professionally, turning us into 'data-enabled prosumers'[9]. Sensing vital signs could increase worker security and well-being by decreasing the risk of critical failures due to fatigue or illness. User interactions with products and services will directly affect design and increase personalisation. Working environments will become much more organic by exploiting network effects of the individual demands of employees and varying requirements of reliable provisioning of services or delivery of products. The working individual, their employers and wider society could all benefit from this. Unprecedented flexibility, personalisation, and empowerment will revolutionise traditional work environments and change how work can be aligned with personal considerations such as caring responsibilities, ill health and ageing.

There are complex ethical and legal issues associated with such new approaches. Personal data can be used on behalf of workers to uncover workload and stress issues. Conversely employers can use this type of information to measure workforce productivity and identify underperformers. Sensitive brokering between employers, employees and unions will be required to deliver safety benefits in this field.

[9] Toffler, Alvin. The third wave. New York: Bantam books, 1981.

Big data: challenges and risks

Scale and uncertainty

One challenge from big data arises from heterogeneous modes of data sensing and sampling. Discrete data contrast with the continuous nature of the problems being tackled and many current mathematical and algorithmic approaches fail to deal with this. This is particularly prevalent in applied research, where the need to integrate results from the methodologies of disparate analytical fields often requires integrating levels of analysis that were not designed for co-analysis. Sparse data, even if sampled at very high rates, can also cause analytical problems such as coverage bias and may lead to important signals between samples being missed. New modes of sampling and approaches to modelling are ongoing challenges for big data and its underlying science. Similarly the problem of representing or modelling uncertainty in data requires handling random or statistical uncertainties, for example arising from measurement inaccuracies and data sampling problems, and also biases and uncertainties in data collection. To overcome such issues, which are important in the validation of data mining and machine learning, new models of abductive reasoning that attempt to generate and test hypotheses automatically need to be developed.

Data and cybercrime

Data misuse, unauthorised access to data and other data crimes are emerging issues. But more subtle misuses of data do not require access to secure data infrastructures, for example fake or abusive reviews, statements to discredit people and other kind of misleading information.

The European Union Agency for Network and Information Security (ENISA) published a comparative study on the cyber crisis management and the general crisis management in November 2014. The report lists six key recommendations:

1. Develop a common cyber crisis management glossary.
2. Gain further knowledge regarding cyber crisis management.
3. Initiate activities for enhancing the knowledge on cyber crisis management.
4. Support training and exercises in the field of cyber crisis management.
5. Support development and sharing of strategic cyber crisis management procedures.
6. Enhance information sharing and collaboration between private and public organisations.

The resilience of complex engineering supply chains against cyber attacks will be an important feature of cyber crisis management of the future. Cyber security is an important consideration for the engineering sector going forward.

Big data risks and effects

The two previous examples illustrate risks at the analytical and the infrastructural level. Table 1 summarises this and sets it in context with examples of industrial areas that are particularly affected by certain risks, although this cannot be exhaustive.

Big data reference	Description of potential risks	Example affected industrial area
Data integration technologies	• private data revealed by 'mosaic' of supposedly anonymised datasets • incompatible data standards and reference data causing correlation errors	• healthcare, transport, utilities • consumer services • world wide web
Correlation/causation difficulties	• data analytics without e.g. proper uncertainty quantification can lead to significant false positive results (i.e. implied causalities) • over attribution of propensities that are not necessities	• insurance and classification industries • healthcare • government and policy
Collection biases	• assuming data is predictive of a larger, unbiased cohort when it is not representative	• economic modelling • consumer prediction • business analytics
Autonomous machines	• vulnerability of machines to cyber attacks leading to unlawful control • terrorism • piracy	• transport • energy • marine • mining • aerospace
Data quality	• data insertion, updating or deleting by unauthorised individuals • obsolete or incomplete data sets	• all

Table 1: Potential big data related risks and affects.

Big data: implications for the Lloyd's Register Foundation

The Expert Panel highlighted the central and independent position of the Lloyd's Register Foundation from which it can provide leadership to address challenges, some of which are not yet fully appreciated in an age of data-centric engineering.

Current widely visible and successful examples of data-intensive solutions exploiting big data are driven by large web companies. But there is huge potential in tapping into developments and initiatives that have been successful on the open web. Revisiting or adapting them in the context of complex engineering disciplines is a very promising opportunity, even in sectors for which the transition from strictly linear models to highly interconnected network ecosystems are still ongoing or have not yet even started.

Catalysing data-centric engineering

The Expert Panel identified a central opportunity for the Foundation to drive the adoption of data-centric methodologies in complex engineering. It will become increasingly important to examine data infrastructures involving multiple stakeholders including the data publisher and the consumer. It will require multiple perspectives; methodological, technical, and legal points of view, and will require development of data-centric engineering codes and standards.

The Expert Panel identified the following technical and non-technical challenges relevant to this overall goal.

Technical challenges

Heterogeneous and multi-modal data
As a result of pervasive sensing at the machine, environment and worker level, every infrastructure has the potential to generate enormous amounts of data. This data is composed of user-generated content (e.g. traces of digital communication) and machine-generated content (e.g. data collected by sensors), complemented by structured data from external sources (e.g. open government data about environmental conditions or policies).

If a sensor delivers a data point then we have to consider if it is consistent with other data - perhaps the time periods are different, the units used are different, the types of modality sensed are different (text, audio, imagery), the way metadata (data about the data) is represented is different. The discovery, alignment and integration of this heterogeneous and multi-modal data is a critical step preceding its analysis.

Data needs to be transparently aligned at the schema level but also at the level of units and measurement precision. This becomes a challenging task in integrated systems that cross multiple sectors with a plethora of custom services and without comprehensive international standardisation. Initiatives on the world wide web, such as the linked data principles[10], show how carefully designed recommendations and best-practices for data publication and exchange help to lower the barriers for large-scale and cross-sectoral data integration.

These recommend such things as the adoption of persistent, globally unique identifiers for instance data, schema entities and relationships, the adoption of access to data via a common protocol, and a standard format as to how the accessed data should be represented and interlinked. They are the foundations for interoperability without the need for heavyweight, burdensome international service-level standardisation. It is an approach that has had dramatic success in the case of the internet and world wide web, informed by the abstract notion of dataspaces[11], the decentralisation of data ownership and control and shared effort to provide services for data discovery and integration between data publishers, data consumers and third parties.

There are a large and continuously growing amount of open data sources worldwide (e.g. open government data[12]), but not all open data sources are yet published to a high standard. For linked data the standards and guidelines are in place and there has been significant uptake and conformity. This demonstrates the impact that a light touch approach can have at a global scale and within a highly dynamic context. The Open Data Institute[13] is the blueprint for catalysing the further growth and sustainable establishment of an open data culture. It offers consultancy, development, startup incubation, training, and data certification for companies, governments, journalists and other stakeholders that work with open data.

[10] Heath, Tim, and Bizer, Christian. Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology 1.1 (2011): 1-136.
[11] Franklin, Michael; Halevy, Alon; Maier, David. From databases to dataspaces: a new abstraction for information management. ACM Sigmod Record 34.4 (2005): 27-33.
[12] http://data.gov.uk
[13] http://theodi.org
[14] http://schema.org
[15] http://videolectures.net/iswc2013_guha_tunnel/
[16] Uniform resource identifiers – a world wide web standard
[17] Tiropanis, Thanassis, et al. The Web Observatory: A Middle Layer for Broad Data. Big Data 2.3 (2014): 129-133; and Tiropanis, Thanassis, et al. The web science observatory. IEEE Intelligent Systems 28.2 (2013): 100-104.

The schema.org[14] initiative was jointly started by competing search engine providers in 2011. It set the standard for describing data on the web, first and foremost for eCommerce but increasingly other domains. eCommerce is a highly competitive sector where the stakeholders often seek to avoid comparability and are often reluctant to agree on service-level standards because this might threaten their business models. Nevertheless, the benefit of simplified discovery of product offers on the web has overcome the reservations and concerns of competing companies. Recent statistics indicate 15% of all web pages have schema.org mark-up[15].

Linked data uses identifiers[16] that work at web scale to connect islands of data together. There are good examples in areas such as news and media and eCommerce; examples from engineering disciplines are rare. An initiative comparable to shema.org has not been effectively developed. The Foundation, as an independent organisation with charitable aims, could provide leadership to launch such an initiative.

A recent direction in data analysis is web observatories[17]. The goal is to build a distributed infrastructure for the exchange and use of research data and analytical methods. The approach leverages web standards and simple best practices in order to achieve transparent data access and interoperability at very large scale. Web observatories expand this to the sharing and retrieval of analytics in order to increase the reproducibility of data-intensive research. The opportunity for the Foundation's strategic sectors is to adapt the web observatory concepts and best practices for trusted data exchange infrastructures in engineering.

Data analytics
The development of new methods and tools as well as technical capabilities for big data analytics is highly dynamic driven by academia and large companies. In everyday practice there are only a few examples of well-established models, such as the analytics underlying weather forecasting. While these models need to be adjusted as computational and sensing capabilities advance and the volume of processable data increases, applications do not have any reference models that have been proven to work reliably.

There is a significant potential for the Lloyd's Register Foundation to stimulate work with large datasets from the shipping and energy sectors, which are unavailable to researchers at the moment. Sectors such as meteorological services are equipped with public institutions with partial legal capacities (e.g. the Deutscher Wetterdienst[18]) and associated laws[19] to provide reliable information and research that meet the requirements of the economy and society. In the big data age many other sectors need similar capabilities. The Foundation is ideally placed to promote opportunities and to tie together national approaches at an international level.

Interconnected ecosystems
In most modern societies every single component is part of one or more complex interdependent networks of humans, machines, and environmental infrastructures. The production and delivery of food, from the harvesting of the raw ingredients to its consumption, exemplifies how the macro-scale ecosystem of our planet breaks down into many micro-scale ecosystems, systems that people, machines and goods traverse and interact with dynamically. Ubiquitous sensing enables unprecedented access to the digital traces of these ecosystems. Interdisciplinary work between theoreticians and practitioners in complex systems, computational modeling and various engineering-intensive sectors is required to understand the cascade effects of risks, benefits and liability.

[18] http://www.dwd.de/
[19] http://www.dwd.de/bvbw/generator/DWDWWW/Content/Oeffentlichkeit/KU/KUPK/Wir_ueber_uns/Gesetz_PDF_en,templateId=raw,property=publicationFile.pdf/Gesetz_PDF_en.pdf

The Lloyd's Register Foundation strategy targets the promotion of supply chain resilience. The previous examples show that big data enables and expands the view from linear interdependence to complex networked systems. Consequently, interdisciplinary research on the structure and behavioural effects in interconnected 'ecosystems' as well as non-linear value models will impact the Foundation's future business and strategic sectors.

Non-technical challenges

Data certification
Certification is a measure of how well a standard has been implemented. It provides consumers with reliability in terms of machine and service capabilities and allows operators and providers to be held accountable.

Data certification is concerned with metadata that describes what is in the data, who created it, for what purpose, what is the quality of data, and what value arises, among other possible objectives. This metadata can be very generic and coarse but it might also be necessary to record fine-grained continuous provenance[20] about how data is used and changed.

The W3C (World Wide Web Consortium) recommendation entitled PROV-DM: The Provenance Data Model defines provenance as 'a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing'[21]. Provenance is metadata that allows us to re-trace the entire history of a primary piece of data from its creation, through all the uses and changes it underwent up to the deletion of the data object. Depending on the level of detail to which provenance is represented, it is possible to infer which system stakeholder (e.g. a person or a system/system component) was involved in a particular activity that has been performed with or on the data. In order to develop accountable and trustworthy big data applications it is necessary to complement provenance with methods to express what can be done with data - so called data terms of use - and to enforce provenance rules at the right level.

A successful example of data certification is the open data certificate[22] developed by the Open Data Institute. Provided for data publishers and data consumers it describes what an open data asset is about, its intrinsic quality and how to access it. Information like availability, privacy, and licensing are assessed and result in the granting of a certificate on a tier from raw, to pilot, to standard, to expert. This four-tier scale reflects the characteristics of an open data source so users can decide how much to rely on it.

Big data not only involves data but also the analytical and predictive methods applied to it. Due to the critical and security-relevant impact these methods have, as in preventative maintenance for example, it is necessary to complement the certification of data with certification of the entire processing chain it runs through. In this scenario particular issues of accountability and data ownership arise when it comes to the latent information derived from inference based on multiple big data sources.

[20] Moreau, Luc, and Groth, Paul. Provenance: An Introduction to PROV. Synthesis Lectures on the Semantic Web: Theory and Technology 3.4 (2013): 1-129.
[21] http://www.w3.org/TR/prov-dm/
[22] http://certificates.theodi.org

Basic opportunities and questions for the Lloyd's Register Foundation relate to what data is certified and also for how long it should be curated. The Foundation's strategic sectors have not been provided with clear guidelines of how operational analytics can be assured, especially in energy and transport scenarios, in which big data is heavily implicated. Consequently the certification of data assets needs to be embedded in a holistic approach to certified data-centric supply chains. The Lloyd's Register Foundation could play a pivotal role in the adoption of data certification and the certification process itself.

The Foundation can also benefit from the momentum of the Open Data Institute initiative by supporting it in complex engineering sectors in order to promote the basic principle of data-centric engineering – 'design for data'.

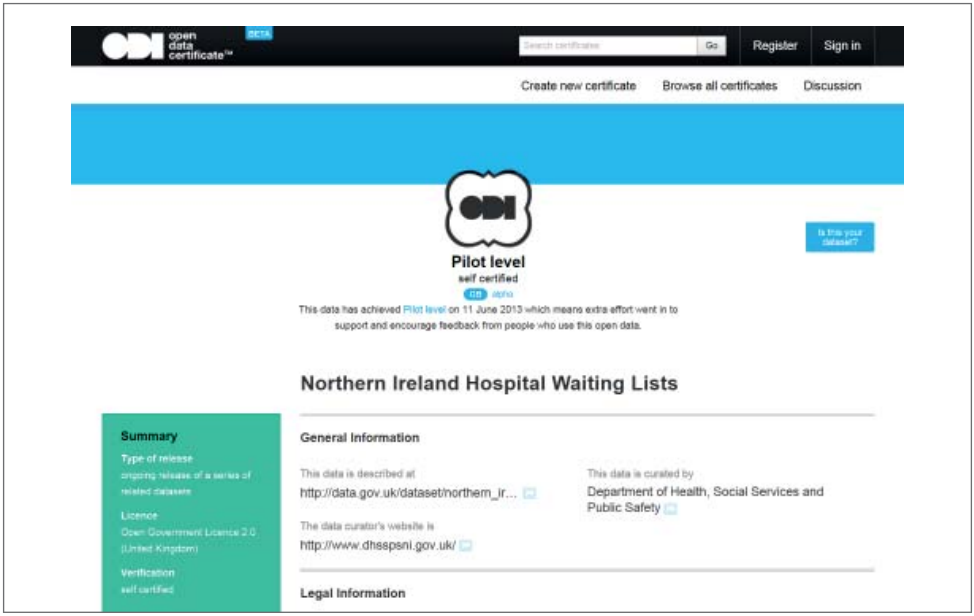


Figure 6: An open data certificate endorsed by the Open Data Institute.

Data code and standards

The market of data-centric applications is evolving fast as shown by start-ups, which develop new web and smartphone applications for communication, social networking, information sharing, self-tracking, health monitoring or smart homes. The business models are typically centred around the analysis of user data to advertise and recommend products for purchase. The most successful start-ups are often acquired by one of the large players on the web, extending the centralised repositories of personal data.

While these business models are promising in terms of short-term monetary success even for very small start-ups they can be a barrier for disruptive innovation. The technological nature of the protocols and standards that constitute the world wide web allow for building completely decentralised applications that put the user in control of their data and break with the way web applications are currently built. However, decentralised applications with decentralised data ownership pose even harder questions to big data analytics and are thus less lucrative for rapid acquisition and revenue generation.

A leadership position needs to be taken over social capital aspects of big data and in how big data business models are formulated, models that are not biased by individual interests of large web companies. The Lloyd's Register Foundation is in a perfect position with sectors that cannot adopt the web companies' business models directly due to regulatory issues and competition. This constitutes a significant potential for achieving market-readiness of decentralised web applications backed up with reliable policies and business models in these sectors. This will impact the way value on the web is generated.

Big data: a future timeline

In order to provide an overview of how big data will impact the core sectors relevant to the Lloyd's Register Foundation, the table in this section sets out likely future developments in the energy, transport and marine sectors. It also anticipates wider societal and safety implications.

Timescale	Short term
	0-5 years
Impact	**Energy:** • consumers assess and optimise energy consumption • demand management **Transport and marine:** • dynamic non-linear processes for predictive and preventative maintenance **Wider implications:** • connected data infrastructures
Applications	• stream processing • smart energy meters • prediction of machine failures • broad publication of high quality open data by governments as well as other public and private entities • business and economic modelling • supply chain modelling (and failure prediction) • environmental modelling • semantic data integration
Scientific underpinnings	• scalable and affordable in-memory data processing • application and fitting of statistical and machine learning models to large amounts of machine data • standards and best practices for linked data publication, quality assessment, repair and consumption • semantic modelling, ontologies, web technologies

Table 2: Big data timeline.

Medium term	Long term
5-10 years	10-20 years
Energy: • adaptive local energy generation **Transport and marine:** • autonomous underwater exploration **Wider implications:** • data enhanced science and engineering • national and international policies and guidelines to prevent and fight cybercrime	**Energy:** • smart grid optimisation **Transport and marine:** • autonomous transportation and delivery **Wider implications:** • personal health management
• question answering • unstructured data analytics • emergent open data ecosystems. • national information infrastructures • life logging • autonomous machines in controlled environments	• exabyte analytics at personal level • autonomous cars and drones • Internet of Things • smart city/ region/ nation information infrastructures
• smarter sensing • breaking the exaflop barrier • data-centric eScience infrastructures and processable publications • web observatories • coupling of data analytics with modelling and simulation • next generation human language technologies	• quantum computing • genomics coupled with large scale analytics • one-million fold increase in compute power and continued improvements in computation and memory • reactive machine learning in highly non-deterministic environments • significant artificial intelligence • data analytics applied to nano-scale components

Findings and recommendations

This report concludes that global trends in big data technology are going to have a major impact on the sectors and charitable aims supported by the Lloyd's Register Foundation. Within the next five to 10 years we are going to witness step changes in sensor technology, autonomous intelligent systems, computer science and algorithms for data analysis. The impact of these will be felt across all the sectors, assets and infrastructure of importance to the Foundation, and across the whole of the product and process life cycles.

Big data, in an engineering-related context, is going to bridge the gap from being able to monitor 'what is' to predicting 'what if' in near real time, creating value through potential enhancements in safety, reliability and performance of assets and infrastructure relevant to the Foundation.

Within this timeframe of the next five to 10 years, the Lloyd's Register Foundation has a unique and immediate opportunity to become a recognised UK and international player in the engineering applications of big data. In this report we have coined the phrase 'data-centric engineering' to describe this. Below are the Expert Panel's recommendations on where the Foundation could invest in data-centric engineering, focused on where it could make a distinctive difference and maximise its impact and value for the wider benefit of society.

This report sets the high-level strategic direction and priorities for the Foundation in the field of data-centric engineering, based on the collective opinion of the Expert Panel. Further work will be needed to decide the implementation details, down to the level of what individual projects to invest in and priorities. The global reach of the Lloyd's Register Foundation is a great advantage and it benefits from working with the best research individuals and teams internationally in the field. The Foundation's approach, exemplified in other research areas, is to establish a research hub that can be networked with international research centres possessing complimentary interests and capabilities.

The Expert Panel's main recommendations are summarised in the figure on the next page. It shows the four main areas relevant to the Foundation where it could take action and invest in data-centric engineering to make a distinctive impact.

Priority action areas			
Technology road mapping	Design for data	Codes and standards	Data analytics
Horizon scanning	Complex independent networks of humans, machines and the enviroment	Data codes and standards: quality, security, integrity	Development of algorithms and modelling tools
	Human sensing - the quantified worker	Open data principles applied to machine data	Data visualisation and simulation
	Data collection - knowing what to instrument and measure	Data certification	Handling of decentralised data sets
	Supply chain resilience: the 'interconnected ecosystem'		Data integration - support data catalogues and data discovery methods
	Data-driven intelligent systems		

Figure 7: Priority action areas

In summary, the four main action areas are:

Technology roadmapping

There is a clear opportunity for the Foundation, working with other funding bodies, the research community, and stakeholders, to jointly construct a technology roadmap for data-centric engineering. This would serve as a tool to help forecast technology developments and plan and co-ordinate efforts, would be a valuable contribution by the Foundation and provide a framework to promote collaboration. Appendix B provides further thoughts on potential collaboration partners.

Design for data

Nowadays it is conventional practice to base engineering design on principles such as 'design for decommissioning' and 'design for maintenance'. On a similar basis it is expected that 'design for data' will become important in the future, in recognition that embedded sensors, intelligent systems, and data management will form part of the integral design, which needs planning at the outset rather than as an add-on. As part of 'design for data', consideration will need to be given to factors such as what to measure, where to place sensors, choice of sensors, sensor development, system integration, interoperability, scalability, computer system design, human interface etc. This is where the science of big data meets engineering.

Codes, standards and data sharing

As more and more data is generated, collected, transmitted, stored and manipulated by engineering systems, there is a need for assurance of that data. Potentially, some very big decisions will be made on the back of such data, creating the need for codes and standards to certify such factors as data quality, traceability, security and integrity. It is also the view of the Expert Panel that there could be a role for the Lloyd's Register Foundation in catalysing the evolution of an open data culture applied to data-centric engineering. This would entail competitors finding economic, environmental, safety and social value in sharing proprietary engineering-related data, which is currently held behind firewalls, to address global issues.

Data analytics

This is concerned with the development of algorithms and mathematical models for data analysis. It is data analytics that will enable the value of big data to be realised by enabling data scientists, predictive modellers and mathematicians to analyse large volumes of sensor and other types of data. This will help make more informed decisions leading to enhancements in safety, reliability and performance of assets and infrastructure relevant to the Foundation.

Appendix A: Broader perspectives on big data in engineering

There is no single perspective on big data. This appendix provides further consideration of the wide range of perspectives that will influence how successfully big data will impact in the engineering-related sectors.

Big data is an asset

Data gathered at each step in a supply chain can add value to the life cycle of a product or process. Tracing components through the supply chain can help feed back data for design purposes as well as for warranty and liability. Trends in usage patterns can become apparent through data acquired in the supply chain.

Taking this into account and making supply chains ready for data storage, interoperability, reuse and analysis leads to a novel perspective - one that we term 'data-centric engineering'. It is a perspective that is prepared for the economic and societal opportunities and challenges that arise in a networked world saturated with data.

Big data in complex engineering scenarios is characterised by a multiplicity of sources with different degrees of precision, veracity and completeness across various dimensions. The rates and volume of machine acquired data stand in stark contrast to the sparsity of manually recorded observations and results from periodic inspection. With manual inspection, timescales can be very different and often data is logged intermittently. This will continue to be a challenge when monitoring complex machinery. For example, manual inspection is expensive, involves downtime and so this will necessarily be performed at periodic or discontinuous intervals, yet condition information is pivotal to correlating the condition of a machine with the data we analyse. Many of the machines we will want to bring into the world of big data will be fitted with unsatisfactory (by today's standards) 'legacy' systems. Economics will dictate these are unlikely to be upgraded, therefore data sparsity and variability are real-world and on-going challenges for big data practices. Whatever big data technology we develop will need to deal with these scenarios in the near to mid-term. Retrofitting increasingly powerful and cheap networked sensor systems to legacy machinery may be one solution[23].

[23] Modern Machine Shop. Data-Driven Manufacturing Moves Ahead at Mazak, http://www.mmsonline.com/articles/data-driven-manufacturing-moves-ahead-at-mazak

Big data is infrastructure

Big data is not just about the data itself but also about the techniques, methods and tools that are used in its analysis. We see continuous improvements in affordable high-performance computing and large-scale storage capabilities leading to a variety of commercial products that offer distributed computation and storage suitable for a variety of analytical tasks that can be run in batch mode. It is noteworthy that new methods and tools are being continuously developed by open source communities as well as large companies. They are mainly intended to overcome the limitations of current methods when it comes to parallel, in-memory operations. The types of operations needed in real-time stream processing for example. This leads to new trends, new opportunities and a very dynamic market that has to be analysed carefully. Not least because of the danger of costly vendor lock in. Once very large amounts of data have been migrated into particular systems it is only possible to migrate it at high-cost. One reason being that cloud providers charge for data transfer in and out. This is a particularly pressing problem when prices are subject to change, so carefully designed contracts are needed to circumvent such uncertainties.

In big science projects even current supercomputing approaches are at their limits for the largest scientific computing problems such as simulations of the human brain. The fastest computers are currently able to perform 33 quadrillion floating point operations per second (petaflops) but scaling currently comes at a very large cost - not least the cost of energy. A 2011 study estimated that the all data centres worldwide account for 1.1 to 1.5% of world energy consumption[24]. The burgeoning costs of high volume computing is a principal reason why IT companies make large investments into the energy efficiency of their data centre infrastructures.

Innovative cluster and processor architectures are the topic of a number of research initiatives worldwide (e.g. the DEEP and DEEP-ER projects funded by the European Commission) and it is estimated that the exaflop barrier will be overcome by 2020. But physical supercomputer architectures developed in research labs to achieve this goal face real competition from the cloud solutions that evolve from the needs of IT companies as part of their daily business. Initiatives like the Open Compute Project (OCP)[25] show the technical challenges being tackled. The OCP approach in which ideas, specifications and technologies are shared openly is a compelling blueprint for what is needed to maximise innovation in scalable computing.

Hardware architectures will continue to evolve to meet the needs of big data applications. Some analytics make particular architectural demands. In the real-time application of high performance digital systems for example, various multicore devices are required, each addressing a different aspect of the application. Complex algorithms are supported by a digital

signal processing core whose internal architecture is matched to certain types of arithmetic operation[26].

The Controller Area Network (CAN, also known as CANbus or CAN bus)[27] is a microcontroller network specified almost 25 years ago. Originally specified as an in-vehicle network for passenger cars, CAN is implemented by many microcontroller systems today in transport, manufacturing, construction, agriculture, healthcare, communication, retail and finance, entertainment as well as science. CAN is a central part of the 'legacy' infrastructure we face in complex engineering. MILCAN is an adaptation of CAN for military vehicle applications[28].

In complex engineering scenarios the transmission of data can be a critical bottleneck within the data infrastructure. Ships might be in operation at sea for many months, for example. The ability to transmit data via satellite might be disrupted intermittently due to bad weather conditions. This requires big data architectures that provide a certain level of processing and storage capabilities on the vessel. Local sampling can reduce the amount of data that needs to be sent immediately for near real-time processing, as is common practice in aviation already. The overall amount of captured data is transmitted when the plane touches down or the ship reaches port again.

Big data is analytics

In most big data scenarios the existence of data often predates the existence of any model of the actual causal processes at work. Analytical methods are therefore needed to derive or infer patterns of behaviour and models that account for this behaviour. Data mining, descriptive and inferential statistics and machine learning are the most important methods in the big data toolbox to find patterns and help construct models. Data mining methods help to find regular patterns and similar items in data. Descriptive statistics are necessary to describe general properties of samples of data. Inferential statistics are then used to estimate population parameters from sample statistics. With machine learning it is possible to learn the best model that describes the relation between a number of variables.

[24] Koomey, Jonathan. Growth in data center electricity use 2005 to 2010. A report by Analytical Press, completed at the request of The New York Times (2011).
[25] http://www.opencompute.org/
[26] Wiggins, Andrea, and Crowston, Kevin. Developing a conceptual model of virtual organisations for citizen science. International Journal of Organisational Design and Engineering 1.1 (2010): 148-162.
[27] http://www.can-cia.org/
[28] http://www.can-cia.org/fileadmin/cia/files/icc/8/majoewsky.pdf

One ultimate aim of these methods is to make predictive statements about the system the data relates to. Access to the technologies comes at very low costs. Statistical and scientific programming languages such as R[29] or Julia[30] are freely available. However, machine learning has limitations. There is often the need for labelled data provided by experts and engineering know how if we are ultimately to get the results we desire.

For complex machinery, a general picture of big data collection and analysis involving the human in the loop at various stages is illustrated by the example in figure 8. There are usually sensors on all critical systems of machines which provide measurements relating to the machine's health or condition. Typically, the sensor measurements are processed to some degree at the machine level. This data may then be passed off the machine, which can be automatic via satellite or wireless technology. This could also be done manually via a PC link, data stick or similar.

This information is then brought together in one place and can be held alongside other relevant information such as maintenance logs, operational history, design/build/manufacture data and many more information sources. Here the machine's data can be further analysed to generate alerts relating to its health, which can then be used to direct maintenance actions and aid other operational decisions. Fault diagnosis is a major goal of alert generation, although it is not always immediately achievable with complex machinery. The data can also be used to provide a fleet-wide view of the machines. Sometimes alerts are only generated at the machine level (e.g. in cars). However collecting all the relevant data in one place enables much more to be understood about the machines, both individually and collectively.

The process of alerting and diagnosing from data, directing and getting feedback from maintenance is iterative and may also be used to inform design and equipment improvements. Finally, the increased knowledge itself becomes part of the big data.

Even as analytical methods advance continuously and promise increasing automation there will always be grand challenges that are impossible to tackle in a purely computational fashion. The research challenges addressed by a citizen science approach are one example. Here, for example, we find humans recruited and trained to use their exquisitely powerful visual systems to classify galaxies, identify wildlife in video footage and the like. This form of human machine collaboration is another powerful form of data-driven problem solving that needs the computing power of the human cognitive system. The quality assessment and repair of open data falls into that category.

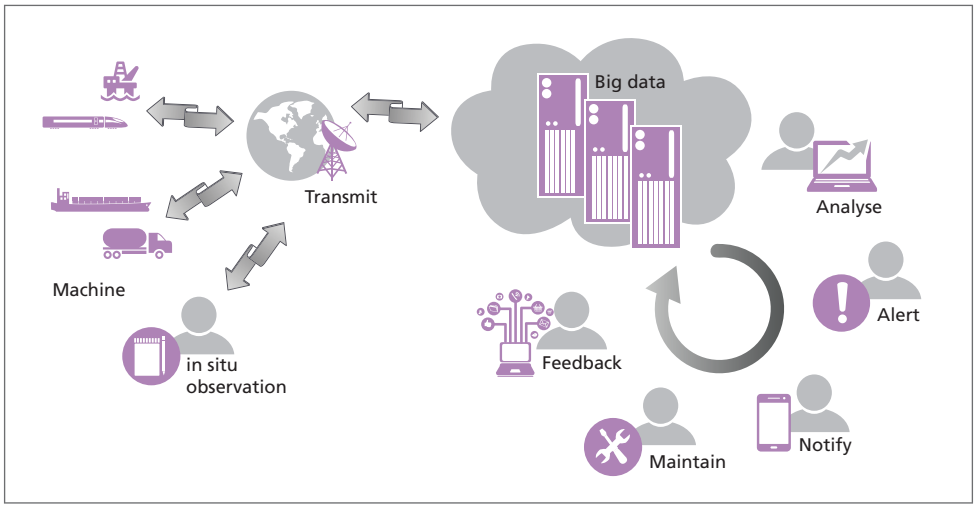


Figure 8: Big data collection and analysis: Humans in the loop at various stages of the big data life cycle. Credit: Honor Powrie.

The big data life cycle is likely to remain a hybrid system, coupling state-of-the-art statistics and machine learning with human-based computation as shown in figure 8. These kinds of socio-technical systems are generally known as social machines[31]. Research on social machines focuses on the patterns of success and failure of existing instances of such systems. It also researches how best to align humans and machines in large-scale, data-driven collaborations.

[29] Team, R. Core. R: A language and environment for statistical computing. (2012), http://cran.case.edu/web/packages/dplR/vignettes/timeseries-dplR.pdf, retrieved 08-11-2014

[30] Bezanson, Jeff; Karpinski, Stefan; Shah, Viral B; Edelman, Alan. (2014) Julia: A fresh approach to numerical computing. http://arxiv.org/abs/1411.1607, retrieved 08-11-2014

[31] Shadbolt, Nigel R., et al. Towards a classification framework for social machines. Proceedings of the 22nd international conference on world wide web companion. International World Wide Web Conferences Steering Committee, 2013; and Hendler, Jim, and Berners-Lee, Tim. From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. Artificial Intelligence 174.2 (2010): 156-161.

Big data is complex

Whenever data is analysed it is necessary to carefully consider the difference between causation and correlation. A task that becomes even more complex the larger the amount of data and the higher its dimensionality (number of variables). The impact of system dynamics on cause and effect is important. In engineering applications even very slow dynamics are crucial to the analysis and help inform the analytics approach. A variety of modelling techniques can be used, such as Petri nets, bond graphs and sometimes simply chains of cause and effect.

Complex machinery, typified by engines used in transport applications, are characterised by a wide range of dynamics. Sudden and catastrophic failure of a component such as an actuator or sensor will cause an instantaneous change in the system behaviour while a slow process of wear or the accumulation of deposits in flow paths (referred to as fouling) produces a correspondingly slow change, but one that can be masked or confounded by other slow changes as illustrated by figure 9.

In modern internal combustion engine systems, the widespread use of electronic systems introduces the potential for the kind of sudden failure associated with electrical systems. Loss of a power connection to a component means that its function will simply cease within a timescale of the order of tenths of seconds. Often this situation is compensated by a second (redundant) system that is either switched in as soon as the failure is detected, or is continually monitoring and takes over once failure is detected. Usually in lower costs systems, the new status is quickly detected and a 'limp-home' mode invoked.

Depending on the failure mode and also how effectively a component is being monitored the breakage of a mechanical component can occur after a period of deterioration, but the effect is noticeable very quickly. The failure of a fuel supply or the internal components of an engine cylinder will lead to a sudden loss of engine performance and a change in its vibration patterns. Such changes will be seen within a few engine cycles and in time scales of the order of tenths of seconds.

Fouling of flow paths within the engine and blockage of heat exchangers can take a prolonged period, and would be managed through normal maintenance procedures, however exceptional conditions can cause a build up over hours, but a significant change in system behaviour may only become apparent over a period of minutes prior to a total loss of flow.

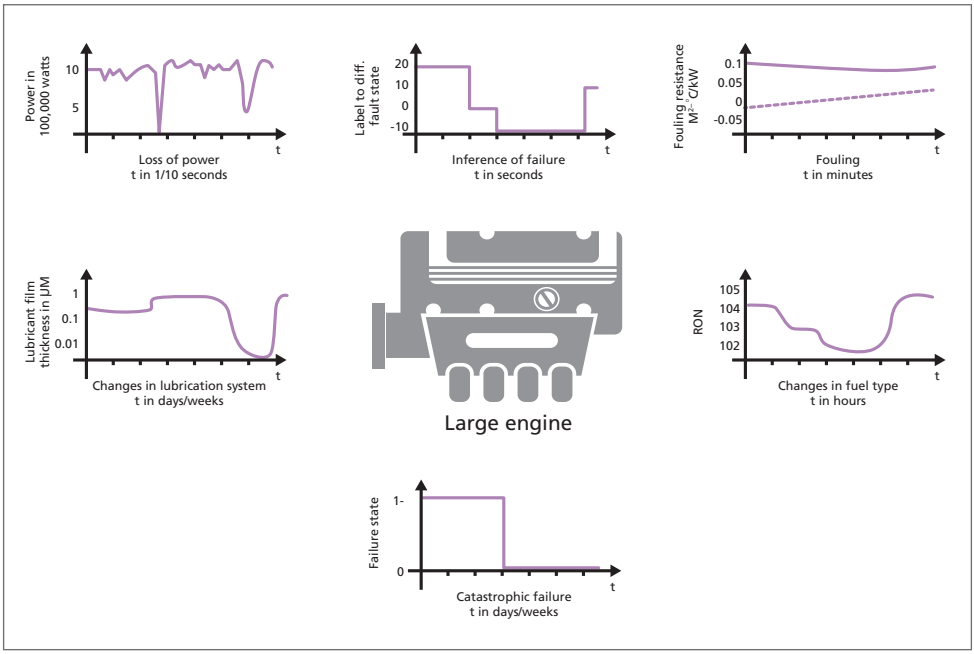


Figure 9: Temporal variability in big data. Credit: Markus Luczak-Roesch.

Changes in the lubrication system are long term and often due to the accumulation of planned modifications of the oil during its serviceable life. Additives are essentially consumed during the operating life of the oil, but such changes are often masked by other effects such as the solution of combustion products in the oil. These changes occur over hundreds of operating hours, and could be detected in large fleet applications by a comparison between individual engines within the fleet or by clustering different classes of behaviour. In diesel engines, the loss of fuel into the lubricating oil due to a mechanical failure could have an effect over days or weeks.

Another type of timescale in engine systems is the one invoked by changes in fuel. Fuel supplies change suddenly – for example between refuelling events. The new fuel may be different and will produce a different behaviour. The benefit of fleet wide assessment of data enables a benchmark performance to be established against which this kind of fault can be assessed. It also allows the record to be held for future classification of performance. In a different domain, natural gas fuelled engines can be subject to this kind of change. The fuel supplied can vary substantially in quality according to the different gases blended and supplied into the gas supply system. In this kind of engine, the change is instantaneous with the new fuel.

What this very real engineering example illustrates is that our diagnostic, analytic and predictive methods need to be sensitive to a very wide range of temporal patterns many of which will have complex and interdependent underlying origins.

Big data is broad

To understand the full picture of big data it is important to account for further varieties of data from open to closed, personal to non-personal, task specific to 'exhaust' data, real time to offline, or structured to semistructured. This perspective, sometimes called broad data, emphasises additional challenges that are associated with the web-scale ecosystem in which most modern data-centric applications reside.

Open data is one important part of this picture. Since 2009, both the US and UK governments have been making increasing amounts of non-personal public sector data available as open data. This is machine readable data that has a licence that says it is open - freely available for others to use subject usually only to the requirement of attribution. Increasing numbers of governments, regions, cities and organisations both public and private have followed suit.

In 2012 the European Commission also highlighted the importance of open access to publicly funded research and governmental data. Open data was described as the fuel for effective research and innovation. The quantity and quality of open data is set to further increase as more public and private organisations publish and consume it. The Open Data Institute (ODI)[32] is demonstrating the wide range of value creation that flows from the production and consumption of open data.

With the emergence of effective data discovery methods, enabled by the uptake of standard vocabularies such as schema.org and the improving quality of open data portals like data.gov.uk, we are seeing open data assume a fundamental role in the data landscape. Open data will become an important complementary resource for businesses, which in turn need advanced technologies, methods and tools to work with data from such sources, including data search and discovery, rapid, ad hoc data integration as well as policies for data use, reuse and combination.

This emphasises the need for data infrastructures, which we argue will need to be powered by semantic technologies. These are needed to provide computational approaches that will allow researchers and practitioners to search-for and discover data resources, rapidly integrate large-scale data collections from heterogeneously collected resources or multiple data sets, and compare these results across datasets to allow generation and validation of hypotheses. Designing better, automated tools that will allow us to find and reuse data that is currently unknown, is necessary to turn data analytics from art to science. Allowing for cross-dataset validation, reproducibility studies on data-driven results, and the concomitant citation of data products enables recognition for those who curate and share important data resources.

This trend towards open data and open standards is also echoed in specific sector developments. For example, over the past decade the Operations and Maintenance Information Open System Alliance, MIMOSA[33], has been dedicated to the development and promotion of open information standards for operations and maintenance in manufacturing, fleet, and facility environments. The MIMOSA Open Systems Architecture for Condition Based Maintenance (OSA-CBM) defines the data interface standards for the different layers of condition-based monitoring systems and is widely used in complex engineering. Big data infrastructures in engineering have to account for such systems in operation and integrate with them.

Another example for emerging open standards in a broad sector is the Building Information Modelling (BIM) framework which provides a digital model of a building in order to make it an openly sharable artefact. This promises more effective services for the architecture, engineering and construction (AEC) industries when working on joint projects[34]. BIM is seen as key to the delivery of efficient services over the course of an entire project life cycle.

[32] http://theodi.org
[33] http://www.mimosa.org/
[34] Azhar, Salman. Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry. Leadership and Management in Engineering 11.3 (2011): 241-252.

Big data is a collective responsibility

Data exchange between businesses, governments and citizens happens with increasing rapidity. Despite the examples from the last section many of these exchanges are currently ad hoc. For individuals data exchange is not always optional but often obligatory. Whether it is the workplace of an employee, a governmental agency consulted by a citizen or when services or goods are purchased, data is acquired from the person. The result is that very large amounts of personal data travel between different jurisdictions. Much of this is not well regulated and may breach confidentiality and ownership rights. Certainly there is an asymmetry between the harvesters of this data such as the large social network or recommender sites and the individuals themselves. The current uncertainty and flux in this space is one area of concern as big data methods and approaches are deployed. It is an area that needs agreed international policies and potentially regulation.

The generation of personal data in interactions between citizens and various stakeholders is ubiquitous. This makes it necessary to develop broad data literacy amongst citizens. As was successfully done for paper records and contracts, this needs to be guided by clear legislation. But also the individual is required to adopt the conditions of the digital age and adjust behaviour accordingly.

Sensitivity over the ownership and use of data is not just the prerogative of individuals. In engineering systems, suppliers of equipment are naturally very careful to protect their intellectual property – including the design – of mechanical aspects, hardware and software. For many suppliers and OEMs (original equipment manufacturers), data underpins a significant part of their revenue and value stream. There will be a concern about opening this up to others. This makes the exchange of certain data difficult. Independent brokers might be needed that do not seek to reverse-engineer design information from data.

Appendix B: Big data: partnerships for the Lloyd's Register Foundation

Big data is a central topic for today's research agendas. While it is not possible to draw a complete picture of the very large and dynamically changing global funding landscape, the panel identified a number of potential strategic partners for the Lloyd's Register Foundation, and medium-term programmes where complementary funding is viewed as useful.

International bodies focused on codes and standards

ISO, IEC, NIST and IEEE are standards organisations that have recently initiated individual and joint study groups on big data. Central to these groups are questions around the unification of what defines big data and what characterises big data solutions. These organisations also work on accelerating the deployment of robust big data solutions by looking at standardisation challenges and how big data integrates with currently deployed and standardised architectures and technologies.

The maritime sector is one of the Foundation's strategic application areas. In this sector the International Maritime Organization (IMO), as a specialised agency of the United Nations (UN), works for the safety and security of shipping and the prevention of marine pollution by ships. The IMO is responsible for regulation and standards on an international basis for shipping with the goal to reduce risks resulting from compromising on safety, security and environmental performance and to encourage innovation. As such it is a partner to leverage certification of data-centric engineering approaches to increase supply chain resilience in shipping.

The UN released its data revolution report on 7 November 2014 entitled 'A World That Counts: Mobilising the Data Revolution for Sustainable Development'[35] . The report is a call to action listing five key recommendations:

1. Develop a global consensus on principles and standards.
2. Share technology and innovations for the common good.
3. New resources for capacity development.
4. Leadership for co-ordination and mobilisation.
5. Exploit some quick wins on sustainable development goals (SDG) data.

Engineering-intense industries have a key position for the sustainable future of our planet. This report describes the technological and methodological agenda of data-centric engineering. The Lloyd's Register Foundation could be a key partner for the UN as an independent and well-established organisation in this field.

[35] United Nations (2014). A World That Counts: Mobilising The Data Revolution for Sustainable Development, http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf, retrieved 07-11-2014.

This report has emphasised the impact of the Open Data Institute on the global open data culture. A significant potential lies in a strategic partnership between the ODI and the Lloyd's Register Foundation in order to enforce a design-for-data culture in engineering disciplines.

International funding landscape

In order to give a flavour of the current financial support for big data research and development we consider the funding landscape in Europe, the United States and Asia.

Europe
The Excellent Science pillar of the Horizon 2020 work programme is basically targeted at high risk research of outstanding individuals or groups in academia. With the European Research Council (ERC) and Future and Emerging Technologies (FET) instruments research is funded that has the potential to shift boundaries significantly. This is relevant to big data research as it is the appropriate programme to develop and evaluate disruptive new computational theories and methods or processor architectures. This pillar also supports universities to establish data science curricula as part of the Marie Sklodowska-Curie Support Actions and the building of computational capabilities for big data under the Research Infrastructures scheme. The second pillar - Industrial Leadership - supports collaborative big data research and development. Consortia consisting of partners from academia and industry are supported in the Leadership in Enabling and Industrial Technologies (LEIT) and Innovation in SME's funding lines. The third pillar is about Europe's Societal Challenges. It is noteworthy that information and communication technology (ICT) topics in general and big data in particular are not only covered by specific funding schemes. They are regarded as key technologies to solve grand challenges, such as the mobility of citizens or goods, energy supply and climate change or food security, health and wellbeing.

Beside the pan-European level the member states of the European Union provide national funding for big data research. While most national programmes are quite generic it is noteworthy that the German Ministry of Research and Education invests up to 200 million Euro in Industry 4.0, the fourth industrial revolution characterised by smart factories and products enabling highly flexible and personalised supply chains that are embedded in complex value ecosystems. So called cyber-physical systems are declared to be a key technology, embedded ICT systems that are highly interconnected with each other and remote services on the internet. Big data, as it is characterised in this report, is a natural component in Industry 4.0 scenarios.

United States

In contrast to the German funding focus the United States put much more emphasis on social media analytics complemented by database research. This results from the success of silicon valley industry around the key players Google, Facebook, LinkedIn and Twitter. As security is a common driver for US investments in research, the NSF (National Science Foundation) is supporting work on cyberinfrastructures, and privacy in a big data context is a big governmental concern and widely addressed. Energy funding schemes support advances in exascale computing. There is no indicator for a focus on complex engineering and also the materials genome project[36] does not have any big data components funded.

Asia

In Asia a number of big data centres are currently funded with a strong focus on database and infrastructure issues such as large-scale data storage and processing, multimedia and social media content analysis. Recently, one can observe a shift towards funding social media analytics. In Asia, 200 million US dollars will be invested in big data topics beginning in 2016.

[36] Materials Genome Initiative, US Federal government http://www.whitehouse.gov/mgi

Further reading

Accenture: Big success from big data, http://www.accenture.com/us-en/Pages/insight-big-success-big-data.aspx

BARC Institute: Big data survey Europe, http://www.pmone.com/fileadmin/user_upload/doc/study/BARC_BIG_DATA_SURVEY_EN_final.pdf

Economist Intelligence Unit: Big data - Lessons from the Leaders, http://www.economistinsights.com/sites/default/files/downloads/EIU_SAS_BigData_4.pdf

Economist Intelligence Unit: Big data - Harnessing a game-changing asset, http://www.sas.com/resources/asset/SAS_BigData_final.pdf

Economist Intelligence Unit for PWC: Gut and gigabtye. Capitalising on the art & science in decision making, http://www.pwc.com/gx/en/issues/data-and-analytics/big-decisions-survey/assets/big-decisions2014.pdf

GE: Digital Resource Productivity Ecomagination, the Industrial Internet, and the Global Resource Challenge, http://www.ge.com/sites/default/files/ge_digital_resource_productivity_whitepaper.pdf

Mayer-Schönberger, Viktor, and Cukier, Kenneth. Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, 2013.

McKinsey Global Institute: Big data: The next frontier for innovation, competition, and productivity, http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx

McKinsey Global Institute: Open data: Unlocking innovation and performance with liquid information, http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights/Business%20Technology/Open%20data%20Unlocking%20innovation%20and%20performance%20with%20liquid%20information/MGI_Open_data_FullReport_Oct2013.ashx

Wikibon: Big data, http://wikibon.org/wiki/v/Big_Data