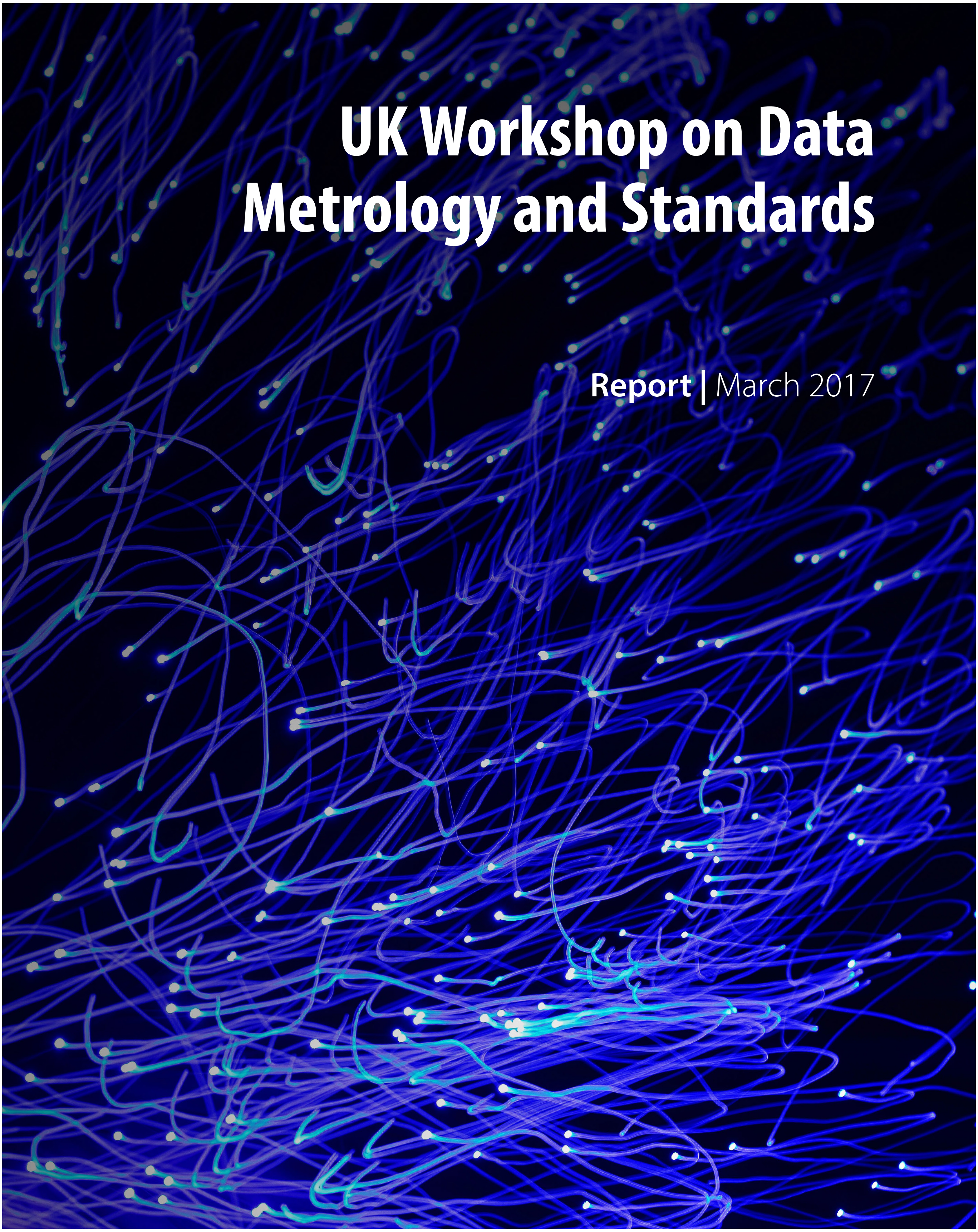
NPL
National Physical Laboratory



# UK Workshop on Data Metrology and Standards

**Report** | March 2017

University of Huddersfield
Inspiring tomorrow's professionals

University of Cambridge

IfM Management Technology Policy

The National Physical Laboratory and partners at the University of Huddersfield and University of Cambridge commissioned this report.

Authors:

**Imoh Ilevbare (PhD) and Nicky Athanassopoulou (PhD)**

IfM Education and Consultancy Services
Institute for Manufacturing
University of Cambridge
www.ifm.eng.cam.ac.uk

**Jenny Wooldridge (PhD)**

National Physical Laboratory
www.npl.co.uk

## EXECUTIVE SUMMARY

Data is a growing part of everyday life, and a key driver for the prosperity and security of the UK. Huge growth in the number of web enabled devices is driving a digital revolution, through the development of systems exploiting the Internet of Things (IoT), cloud computing and industrial automation (4th industrial revolution). Large increases in economic output are forecast with the adoption of these technologies, due to associated growth in productivity, the emergence of new markets, and product and service innovation. In the world of metrology, measurements are moving to be on-line, always on and always calibrated.

For both consumers and industry there are clear risks in terms of data privacy and security in the cloud; a balance is required between increasing the value of information through interconnecting systems and processes, and a need to protect privacy and intellectual property. The value of data is also dependent on quality and the appropriate use of information derived from online systems. To understand the ***trustworthiness of information*** to make business critical, or safety critical decisions, is to understand the ***accuracy and precision of data***, the provenance of data, and the propagation of uncertainty through data processing algorithms and data curation processes (data drift).

For more than a century, the National Physical Laboratory (NPL) has developed and maintained the nation's primary measurement standards. Good measurement improves productivity and quality; the ability to ***quantify quality assurance*** in services and products underpins consumer confidence and trade and is vital to innovation. The development of an effective data infrastructure is necessary to support innovation and increase productivity and growth across the UK. These are key elements within the pillars of the UK Government's Industrial Strategy green paper. In response to this challenge, NPL is expanding its core mission from physical, chemical and biological metrology, and establishing a ***data research initiative*** dedicated to supporting industry in the rapidly accelerating reliance on data.

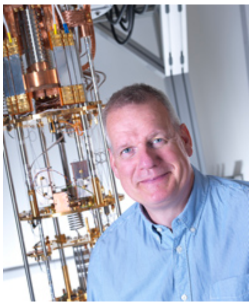
Under its remit as the UK's National Measurement Institute (NMI), NPL will create the measurement framework required for traceability in data systems. Quality assurance enables ***confidence in the intelligent and effective use of data***, increasing the value of information and ensuring the legal standing of decisions made on data analytics. NPL's expertise in the rigour of analysis in physical measurement can be applied to digital systems to meet these goals, through the provision of data standards and verified data processing methodologies, generating unbroken chains of data flow with quantifiable uncertainties at each step.

As part of the kick-off for this initiative NPL, along with partner organisations the University of Cambridge and the University of Huddersfield, organised a UK Workshop on Data Metrology and Standards on 5 December 2016, engaging UK industrial users of data to identify data measurement challenges and explore research project ideas. The most pressing ***industry challenges*** identified during the workshop were:

A. Decision making from multiple sources of information, how data quality can assure high quality information
B. Quantification of data quality to assure high quality information and decision making
C. Trustworthy real-time data and information – quality indicators of Artificial Intelligence (AI) algorithm and the data it produces
D. Standards for archival, metadata and searching of data
E. Sensor technology – standardisation of sensor metadata, storage of sensor datasets, encryption of data to individual sensors and validation and governance of data from sensor to analytics

F. Reliable methods for combining data streams with different characteristics (data type, uncertainty, etc.)
G. Methods for propagating uncertainties through data curation methods and data analytics
H. Training of UK data scientists to meet current and future industry needs
I. Management, use and learning from historical, legacy or available data
J. Improved provenance of measurements, data and databases (and IoT)
K. Ethics of data collection and use on a large scale
L. Machine learning for data processing and analytics
M. Certification of trusted algorithms

The aim of the new data initiative at NPL is to be a business focused partner, providing pre-competitive and bespoke research, and developing standards with enduring value and use. NPL will continue to develop the ideas generated during the workshop, to connect and collaborate with new partners to ensure that quality assurance is embedded into digital systems to the benefit of all users of data.



Dr JT Janssen
Director of Research, National Physical Laboratory (NPL)

## TABLE OF CONTENTS

## LIST OF FIGURES



## LIST OF TABLES

## WORKSHOP DETAILS

### DATE

Monday 5 December 2016, 9.00am – 5.00pm

### VENUE

**The Hauser Forum**

3 Charles Babbage Road

Cambridge

CB3 0GT

### FACILITATORS

**Dr Imoh Ilevbare**, **Dr Nicky Athanassopoulou**, **Dr Michèle Routely** and **Mr Rob Munro**

IfM Education and Consultancy Services Ltd.

Institute for Manufacturing

17 Charles Babbage Road

Cambridge

CB3 0FS.

*A list of delegates is provided in the Appendix.*

## WORKSHOP GOALS

### BACKGROUND

The National Physical Laboratory (NPL) is the UK's National Measurement Institute (NMI). It develops, maintains and applies the nation's measurement standards and solutions. These standards and solutions provide the measurement capability that underpins the UK's prosperity and quality of life.

NPL is establishing a data research initiative dedicated to supporting industry in its rapidly growing reliance on data and the digital economy. Under its remit as the UK's NMI, its activities will include creating the measurement framework required for traceability in data systems and providing data standards and verified data processing methodologies. These activities are required to deliver confidence in the intelligent and effective use of data, increase the value of data and ensure the legal standing of decisions make on big data analytics. Within this initiative, NPL will partner with industry in carrying out pre-competitive and bespoke research.

As part of the kick-off for the initiative, NPL (together with partners at the University of Cambridge and University of Huddersfield) organised a *UK Workshop on Data Metrology and Standards* on 5 December 2016 to engage UK industrial users of data in identifying data measurement challenges over the short-, medium- and long-term. Workshop delegates developed and explored research project ideas to address the challenges.

### AIMS

The workshop's specific aims were to:

A. Engage UK industrial users of data
B. Capture present and future industry needs and challenges regarding the development and use of data analytics and data systems
C. Identify, develop and prioritise project ideas to respond to the needs and challenges
D. Scope and explore top priority projects in greater detail, identifying, for each one, development steps and expected milestones, resources requirements, enablers (e.g. funding mechanisms) and anticipated risks

The workshop process generally followed the Institute for Manufacturing's (University of Cambridge) S-Plan roadmapping process and framework, which allowed contribution, alignment and examination of multiple strategic perspectives by the workshop delegates. These perspectives covered: (1) Industry needs and challenges; (2) Project ideas; (3) Technologies and Capabilities. They extended over three time periods: the short-term (2017 to 2018), the medium-term (2019 to 2021), and the long-term (2022 and beyond).

The workshop had a total of 89 participants from 44 different organisations (including eight universities and NPL).

## METHODOLOGY

The roadmapping methodology followed for the workshop consisted of three parts: scoping and design; data gathering and planning; and the workshop.

### SCOPING AND DESIGN

During the scoping and design phase the following activities took place:

- Confirmation, based on input from the NPL steering group, the aims and scope of the workshop
- Discussion and design of the workshop process. The process was designed based on S-Plan framework developed by IfM over several years.[1, 2, 3] The framework was configured to support the NPL objectives, in aligning research activities with industry needs and challenges, and support decision-making and action.
- Design and customisation of templates to be used during the workshop as well as for pre-workshop (e.g. data gathering) activities;
- Agreement on the factors for comparing and prioritising project ideas
- Agreement on the detailed workshop agenda

### DATA GATHERING AND PLANNING

During this phase, the following activities took place:

- Delegates from each participating organisation were sent a briefing document and a request to prepare their perspectives (on industry needs and challenges and project ideas) ahead of the workshop
- Consolidation of participant perspectives (e.g. to identify obvious overlapping perspectives across participants) to derive a more manageable number of issues for the workshop to focus on

### WORKSHOP

The workshop brought together a total of 89 participants from 44 different organisations, and had the following agenda:

- Registration and coffee
- Welcome and overview of the new data metrology and standards partners: presentations by Peter Thompson (CEO, NPL), Paul Alexander (Chair of Cambridge Big Data Strategic Research Initiative, University of Cambridge, and Andrew Ball (Pro Vice-Chancellor for Research and Enterprise, University of Huddersfield)
- Introduction to workshop process by workshop facilitators

---

[1] http://www3.eng.cam.ac.uk/research_db/publications/rp108

[2] Phaal, R., Farrukh, C. J. P. and Probert, D. R. (2004), "Customizing Roadmapping", *Research Technology Management*, 47 (2), pp. 26-37

[3] Phaal, R., Farrukh, C. J. P. and Probert, D. R. (2007), "Strategic Roadmapping; A workshop-based approach for identifying and exploring innovation issues and opportunities", *Engineering Management Journal*, 19(1), pp. 16-24

- Overview of NPL's three key science areas: presentation by Alistair Forbes (Data Metrology & Standards Science Area Leader, NPL)
- Data Management Initiatives at NIST: a presentation by Bob Hanisch (Director, Office of Data and Informatics, NIST)
- Presentations by each organisation of their perspectives on data metrology and standards needs and challenges, and their project ideas to address them
- Prioritisation of needs and challenges by all delegates
- Prioritisation of project ideas using a list of pre-determined factors by all delegates
- Funding project ideas: presentations by JT Janssen (Head of Science, NPL), and Jonathan Mitchener and Nigel Rix (Innovate UK)
- Exploration of priority project in small groups
- Small group feedback of explored ideas



## INDUSTRY NEEDS AND CHALLENGES

Each participating organisation[4] contributed its perspectives on important industry needs and challenges. These perspectives were collected and consolidated before the workshop. They were then reviewed during the workshop by all participants, whereby a few additional perspectives were added resulting in the list of 56 *industry needs and challenges*, as presented in Table 1. Subsequently, each organisation (through its representative(s)) was asked to identify six industry needs and challenges that it considered most important.

Table 1 shows the industry needs and challenges, listed according to the total number of 'votes' each received across all the participants. This list provides an indication of priorities with 39 of the 56 needs and challenges identified as being important by any of the participants. Two-thirds of all the votes went to only the first 13 needs and challenges.

A list of the remaining 17 needs and challenges (not identified as important, and not shown in Table 1) is provided in the Appendix.

### Table 1 - Priority industry needs and challenges

| Industry needs and challenges | | Timescale | Votes |
|---|---|---|---|
| 1 | **Decision making from multiple sources of information. How can data quality assure high quality information?** | MT-LT | 12 |
| 2 | **Trustworthy real-time data and information; quality indicators of AI algorithm and the data it produces** | ST | 12 |
| 3 | **Quantify data quality to assure high quality information and decision making** | ST-LT | 11 |
| 4 | **Standards for archival, metadata and searching of data** | ST-MT | 10 |
| 5 | **Sensor technology: standardisation of sensor metadata, storage of sensor data sets, encryption of data to individual sensors and validation and governance of the data from sensor to analytics system** | ST-MT | 10 |
| 6 | **Reliable methods for combining data streams with different characteristics (data type, uncertainty etc.)** | ST-LT | 10 |
| 7 | **Methods for propagating uncertainties through data curation methods and data analytics** | ST-MT | 10 |
| 8 | **Training of UK data scientists to meet current and future industry needs** | ST-MT | 9 |
| 9 | **Management, use and learning from historical, legacy or available data** | ST-LT | 9 |
| 10 | **Improved provenance of measurements, data and databases (and IoT)** | ST-LT | 9 |
| 11 | **Ethics of data collection and use on a large scale** | MT | 9 |
| 12 | **Machine learning for data processing and analytics** | ST-MT | 8 |
| 13 | **Certification of trusted algorithms** | ST | 8 |
| 14 | **Confidentiality, Integrity and Availability of data and software in a Cloud** | LT | 6 |
| 15 | **High-speed algorithms for analytics on the fly, and real time uncertainty quantification** | MT | 5 |
| 16 | **Raise awareness in STEM education of the need for metadata to support measurement data** | ST | 5 |

[4] Where multiple departments were represented from the same university, for the purposes of this workshop, each department was treated as a separate organisation.

| | | | |
|---|---|---|---|
| **17** | **Research study and application of data science to data-driven materials design** | ST-LT | 4 |
| **18** | **Agnostic/platform independent algorithms and data security assurance** | MT | 4 |
| **19** | **Quantifying data drift and its effect on data quality** | ST | 4 |
| **20** | **Open disease biology/target validation e.g. 'omics data sets/images** | LT | 4 |
| **21** | **Constructing a secure software environment for the measuring instruments software** | MT | 4 |
| **22** | **Drive toward probabilistic engineering** | ST | 3 |
| **23** | **Education of legislators/policy-makers on the benefits of big data** | ST | 3 |
| **24** | **IP in an age of distributed digital manufacturing** | LT | 3 |
| **25** | **More companies create value through the use of Artificial Intelligence** | MT | 3 |
| **26** | **Visualisation of multiple image types to enable hybrid images; visualisation/display of metadata** | ST-MT | 3 |
| **27** | **GUI interfaces to sophisticated (and context appropriate) optimisation for cognitively limited (human)** | MT | 3 |
| **28** | **Confidence in online identity verification (Digital Economy Bill)** | ST-MT | 2 |
| **29** | **Over reliance on bulk collection. Match collection strategies to intelligent requirements** | ST | 2 |
| **30** | **Maximise effective use of skills through increased use of automation in data analytics** | MT | 2 |
| **31** | **Publicise good metrology practice for specifying, developing and operating cyber-physical systems** | MT | 2 |
| **32** | **Approaches to provide Integrity of Actuation over the internet that confirms faithful physical motion following a remote command** | LT | 2 |
| **33** | **Fully integrated data driven enterprise** | LT | 1 |
| **34** | **Data analytics in finance presents a huge legislative challenge** | MT | 1 |
| **35** | **Measuring/qualifying non-standard data sets such as images, video and/or social media streams** | ST | 1 |
| **36** | **Understand environmental crime by garnering insights into causal factors** | ST | 1 |
| **37** | **Create networks of quantitative data; not just integrating data** | MT | 1 |
| **38** | **Technology: recent surge in computation and data has lead to fast growth in algorithmic capabilities** | ST | 1 |
| **39** | **Whilst technology is rapidly progressing, legislation progresses much slower** | ST | 1 |

## IDENTIFICATION AND PRIORITISATION OF PROJECT IDEAS

Each participating organisation proposed project ideas. These were collected and consolidated before the workshop. In total, 45 different ideas were contributed across the following categories:

- Standardisation projects (11)
- Pre-competitive projects (19)
- Commercial projects (10)
- Other (5)

To identify priority projects ideas, the 45 proposed project ideas were assessed using two different criteria: opportunity and feasibility. **Opportunity** – the magnitude of the opportunity that could plausibly be opened up by virtue of the project's success and **Feasibility** – the ability or preparedness of NPL and its collaborating partners to deliver the project successfully. The specific factors underpinning 'opportunity' and 'feasibility' were selected prior to the workshop through deliberation with the Workshop Steering Group. These factors are presented in Table 2.

### Table 2 - Opportunity and Feasibility factors used to assess the different project ideas

| **Opportunity** | | **Feasibility** | |
|---|---|---|---|
| **Projected impact** | Potential value of new technology in terms of social and economic factors | **Alignment to NPL research themes** | How well does the project align with the themes:<br>- Measuring and transmitting data<br>- Storing and retrieving data<br>- Data analytics |
| **Market size** | Size of potential market, or number of potential adoptions, reasonably available | **Technical challenge** | How confident are we that the proposed technology solution is technically feasible? |
| **Synergy opportunities** | Possible additional benefits to other projects or activities; or the possibility of new opportunities in combination | **Differentiation** | What is the added value generated by quantified uncertainties and verified quality assurance processes? |

There were two steps to the prioritisation process. Firstly, each organisation present was asked to review the 45 projects and its representative(s) were asked to identify (and 'vote' for) the six projects that most aligned with the three opportunity factors. This created a shortlist of 23 projects (22 projects did not receive any opportunity votes).

Thereafter, the participants were asked to consider only the shortlist of 23 projects from the previous step. Each organisation was asked to identify (and vote for) up to four projects that most satisfied the feasibility factors. This further narrowed the shortlist to 21 projects (two projects did not receive any feasibility votes). The shortlist is presented in Table 3.

### Table 3 - Shortlisted projects

| Projects | | Category | Opportunity votes | Feasibility votes |
|---|---|---|---|---|
| 1 | **Develop standards (and optimisation models) for data quality (incl. accuracy, confidence and fidelity)** | Standardisation | 25 | 13 |
| 2 | **Develop data (and metadata) provenance standards and requirements** | Standardisation | 24 | 16 |
| 3 | **Next-generation integration algorithms and methodologies for multiple data sources** | Pre-competitive | 15 | 15 |
| 4 | **Methods and statistics to estimate uncertainty (and develop applications) for spatial-temporal models** | Pre-competitive | 13 | 13 |
| 5 | **Applying HPC, Big Data and cognitive systems for decision support in chemistry, materials, life science and engineering discovery** | Pre-competitive | 12 | 12 |
| 6 | **Develop standards for data security** | Standardisation | 10 | 10 |
| 7 | **Best practice techniques/algorithms for analysis and modelling of sensor data (incl. data compression for storage of previously recorded sensor data)** | Pre-competitive | 7 | 7 |
| 8 | **Machine augmented learning and knowledge extraction from scientific documents** | Pre-competitive | 7 | 7 |
| 9 | **Curation and annotation of very large datasets available for public and commercial usage** | Other | 6 | 3 |
| 10 | **Integrated optimisation of supply chain** | Pre-competitive | 6 | 2 |
| 11 | **Develop tools and standards for sensor precision and calibration over internet** | Standardisation | 6 | 2 |
| 12 | **Improved data quality through advances in measurement and simulation capability** | Pre-competitive | 5 | 2 |
| 13 | **Develop methodology/metrics to track latency across deployment scenarios and technologies, in order to identify 'hot' and 'cold' areas of the system** | Pre-competitive | 5 | 2 |
| 14 | **Storing and analysing data in a cloud and enable services for the manufacturer and market surveillance** | Pre-competitive | 5 | 2 |
| 15 | **Develop risk prediction and analysis models using multiple data sources/types** | Pre-competitive | 5 | 1 |

| | | | | |
|---|---|---|---|---|
| 16 | **Developing next-generation toolsets for data analytics** | Commercial | 4 | 7 |
| 17 | **Extension of NPL Time and ID verification, alongside development of fraud/malpractice detection algorithms** | Pre-competitive | 4 | 5 |
| 18 | **Prototyping IoT in the lab for context awareness.** | Commercial | 4 | 2 |
| 19 | **Develop standards (including ethics and pre-harvest/reconnaissance processes) and homogenous tools/techniques for data collection (and use on large scale)** | Standardisation | 4 | 2 |
| 20 | **Determine new models of data storage, access and distribution that can allow a new distributed economy to thrive under existing restrictions - or rewrite legislation** | Pre-competitive | 4 | 1 |
| 21 | **Statistical modelling for estimation of interactions beyond 'omics layers and for identification of key molecules, biomarkers, drug targets using trans-omics data** | Commercial | 2 | 3 |

The shortlisted projects were spread across the categories as follows:

- Standardisation projects (5)
- Pre-competitive projects (12)
- Commercial projects (3)
- Other (1)

The 21 projects were transferred onto a 2X2 matrix based on their opportunity votes and feasibility votes, with opportunity shown on the vertical axis and feasibility on the horizontal axis (see figure 1 below). This facilitated decision-making and selection of the most appropriate projects to further explore during the workshop.

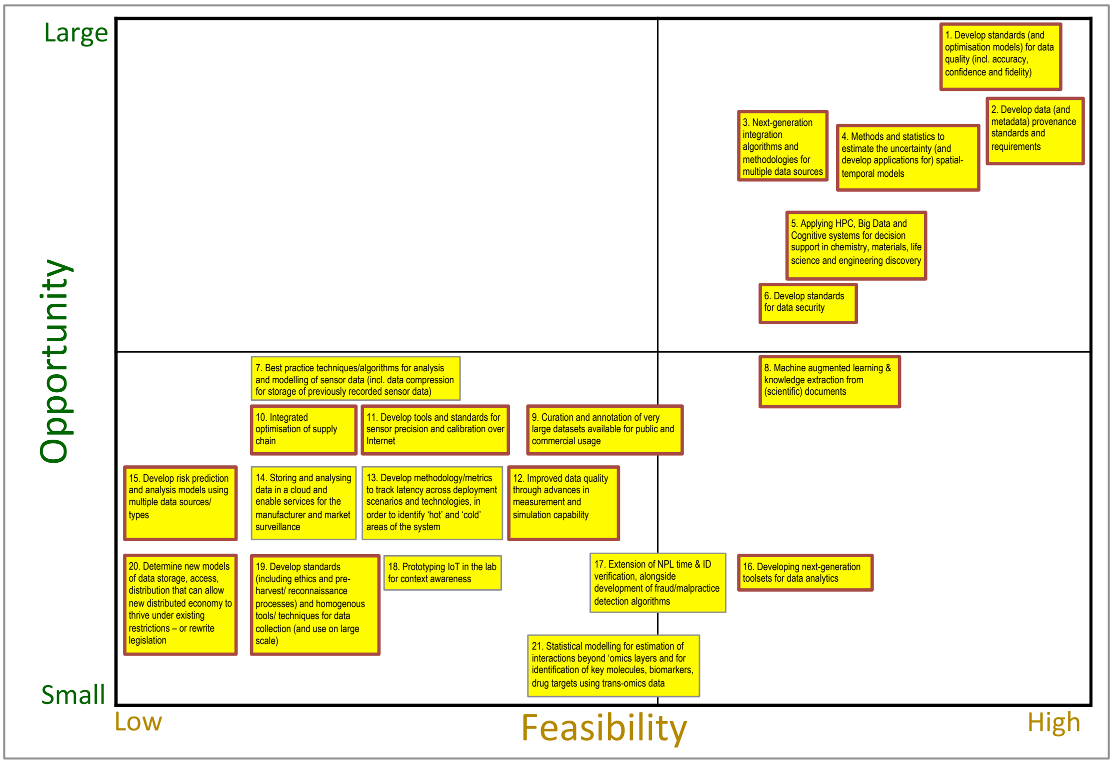



**Figure 1 - Opportunity-feasibility chart showing shortlist projects (and the priorities selected during the workshop (shown with red borders))**

Through their discussions, the workshop steering group selected 15 projects to take forward in the workshop (each is highlighted in Figure 1 with a red border).

These were:

- Develop standards (and optimisation models) for data quality (incl. accuracy, confidence and fidelity)
- Develop data (and metadata) provenance standards and requirements
- Next-generation integration algorithms and methodologies for multiple data sources
- Methods and statistics to estimate the uncertainty (and develop applications) for spatial-temporal models (& Best practice techniques/algorithms for analysis and modelling of sensor data (including data compression for storage of previously recorded sensor data))
- Applying HPC, Big Data and Cognitive systems for decision support in chemistry, materials, life science and engineering discovery
- Develop standards for data security
- Machine augmented learning and knowledge extraction from scientific documents
- Curation and annotation of very large datasets available for public and commercial usage
- Integrated optimisation of supply chain
- Develop tools and standards for sensor precision and calibration over internet
- Improved data quality through advances in measurement and simulation capability
- Develop risk prediction and analysis models using multiple data sources/types
- Developing next-generation toolsets for data analytics
- Develop standards (including ethics and pre-harvest/reconnaissance processes) and homogenous tools/techniques for data collection (and use on large scale)
- Determine new models of data storage, access and distribution that can allow a new distributed economy to thrive under existing restrictions – or rewrite legislation

## SUMMARY DATA METROLOGY AND STANDARDS LANDSCAPE

A summary of the output derived from the foregoing workshop process and discussion is shown in Figure 3, where the industry needs and challenges, the proposed projects and cross-cutting technologies and capabilities are shown across time.

Technologies and capabilities (within NPL's existing data science research areas of *Measuring and Transmitting Data*, *Storing and Retrieving Data*, and *Data Analytics*) that will contribute across the projects were also identified:

A. Comprehensive uncertainty quantification in data integrity/provenance
B. Standardisation in metadata for sensor network systems (including data provenance assurance, records and automation of calibration, the effect of data curation methods)
C. Development of a training and skills plan to ensure the available of appropriate resources to drive growth and innovation in data companies
D. Standards and safety protocols for the next generation of AI and machine written software
E. Modelling systems evolving over time (time series/tipping points/change point analysis, spatio-temporal systems, quality assured dynamic maps, data assimilation for environment monitoring applications)
F. Development of digital calibration certificates
G. Measuring and annotating data quality/fidelity in real-time. Established methodologies to adjust data veracity in real-time to meet the need for which data is being gathered
H. Uncertainty quantification for imaging systems (uncertainty methodologies in quantitative imaging, compressed sensing, sparse reconstruction, high level features, etc.)
I. Algorithms for model discovery from multiple data streams (e.g. robust PCA, tensor decompositions)
J. Comprehensive uncertainty quantification in algorithm/computation (software standardisation and certification)
K. Verified lineage of data and governance of the data from sensor to system

| | Priority Projects | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Technologies and Capabilities | Develop standards (and optimisation models) for data quality (incl. accuracy, confidence and fidelity) | Develop data (and metadata) provenance standards and requirements | Next-generation integration algorithms and methodologies for multiple data sources | Methods and statistics to estimate the uncertainty (and develop applications for) spatial-temporal models | Applying HPC, Big Data and cognitive systems for decision support in chemistry, materials, life science and engineering discovery | Develop standards for data security | Machine augmented learning & knowledge extraction from scientific documents | Integrated optimisation of supply chain | Develop tools and standards for sensor precision and calibration over Internet | Curation and annotation of very large datasets available for public and commercial usage | Improved data quality through advances in measurement and simulation capability | Develop risk prediction and analysis models using multiple data sources/types | Develop standards (including ethics and pre-harvest/reconnaissance processes) and homogenous tools/techniques for data collection (and use on large scale) | Developing next-generation toolsets for data analytics | Determine new models of data storage, access and distribution that can allow new distributed economy to thrive under existing restrictions - or rewrite legislation | Total number linkages |
| Capability for comprehensive uncertainty quantification in data integrity / provenance | ■ | ■ | ■ | ■ | | ■ | | | ■ | ■ | | ■ | | | | 8 |
| Development of training and skills plan to ensure the availability of appropriate resources to drive growth and innovation in data companies | ■ | ■ | | | ■ | | | ■ | | ■ | | | ■ | | ■ | 7 |
| Standardistaion in metadata for sensor network systems (including data provenance assurance, records and automation of calibration, the effect of data curation methods) | | ■ | | | | ■ | | ■ | ■ | ■ | | | ■ | | | 6 |
| Standards and safety protocols for the next generation of AI and machine written software | | | | | ■ | | ■ | | ■ | | | | ■ | ■ | | 5 |
| Capability for modelling systems evolving over time (time series/tipping point/change point analysis,. Spatio-temporal systems, quality assured dynamic maps, data assimilation for environment monitoring applications) | | | | ■ | | | | | | | | ■ | ■ | ■ | | 4 |
| Development of digital calibration certificates | | ■ | | | | | | ■ | ■ | | ■ | | | | | 4 |
| Measuring and annotating data quality/fidelity in real time. Established methodologies to adjust data veracity in real time to meet the need for which the data is being gathered. | ■ | | ■ | | | | | | | ■ | | ■ | | | | 4 |
| Capability for uncertainty quantification for imaging systems (uncertainty methologies in quantitative imaging, compressed sensing, sparse reconstruction, high-level feature extraction/classification, and sensor networks as irregular imaging systems) | ■ | | | ■ | | | | | | | ■ | | | | | 3 |
| Algorithms for model discovery from multiple data streams, e.g. sparse reconstruction algorithms, robust PCA, tensor decompositions. | | | ■ | | | | | | | | | ■ | | | | 2 |
| Capability for comprehensive uncertainty quantification in algorithms / computation (software standardisation and certification) | ■ | | | ■ | | | | | | | | | | | | 2 |
| Verified lineage of data and governance of the data from sensor to system | | | | | | | | ■ | | | | | | | | 1 |

**Figure 2 - Crosscutting linkages between NPL (and partner) capabilities and priority projects**

| NPL Data Metrology and Standards Workshop 2016 Summary Landscape | | Short term 2018 | Medium term 2021 | Long term 2022+ |
|---|---|---|---|---|
| Industry Needs and Challenges | STEEPLE | Drive toward probabilistic engineering | Data analytics in finance presents a huge legislative challenge | |
| | | Whilst technology is rapidly progressing, legislation progresses much slower | More companies create value through the use of Artificial Intelligence | |
| | | Technology: recent surge in computation and data has led to fast growth in algorithmic capabilities | | |
| | Confidence in data | Standards for archival, metadata and searching of data | | |
| | | Improved provenance of measurements, data and databases (& IoT) | | |
| | | Reliable methods for combining data streams with different characteristics (data type, uncertainty etc.) | | |
| | | Over reliance on bulk collection. Match collection strategies to intelligent requirements | High-speed algorithms for analytics on the fly, and real time unceratinty quantification | Confidentiality, Integrity and Availability of data and software in a Cloud |
| | | Certification of trusted algorithms | | |
| | | Measuring /qualifying non-standard data sets such as images, video and/or social media streams | | |
| | | Trustworthy real-time data and information. Quality indicators of AI algorithm and the data it produces. | | |
| | | Methods for propagating uncertainties through data curation methods and data analytics | | |
| | Effective use of data | | Decision making from multiple sources of information. How can data quality assure high quality information? | |
| | | | | Fully integrated data driven enterprise |
| | | Research study and application of data science to data-driven materials design | | |
| | | | Agnostic / platform independent algorithms and data security assurance | Approaches to provide Integrity of Actuation over the internet that confirms faithful physical motion following a remote command |
| | | Confidence in online identity verification (Digital Economy Bill) | | |
| | | | Visualisation of multiple image types to enable hybrid images; visualisation/display of metadata | |
| | Measurement | Quantify data quality to assure high quality information and decision making | | |
| | | Quantifying data drift and its effect on data quality | Constructing a secure software environment for the measuring instruments software | |
| | Skills and Capabilities | Training of UK data scientists to meet current and future industry needs | | Open disease biology/target validation e.g. 'omics data sets/images |
| | | Machine learning for data processing and analytics | | |
| | | Education of legislators/policy-makers on the benefits of big data | Maximise effective use of skills through increased use of automation in data analytics | |
| | | Raise awareness in STEM education of the need for metadata to support measurement data | Publicise good metrology practice for specifying, developing and operating cyber-physical systems | |
| | Other Needs | Sensor technology: standardisation of sensor metadata, storage of sensor data sets, encryption of data to individual sensors and validation and governance of the data from sensor to analytics system | | |
| | | Management, use and learning from historical, legacy or available data | | |
| | | Understand environmental crime by garnering insights into causal factors | Ethics of data collection and use on a large scale | IP in an age of distributed digital manufacturing |
| | | Create networks of quantitative data; not just integrating data | | |
| | | GUI interfaces to sophisticated (and context appropriate) optimisation for cognitively limited (human) | | |
| Projects | Pre-competitive projects | Next-generation integration algorithms and methodologies for multiple data sources | | |
| | | Applying HPC, Big Data and Cognitive systems for decision support in chemistry, materials, life science and engineering discovery | | |
| | | Best practice techniques/algorithms for analysis and modelling of sensor data (incl. data compression for storage of previously recorded sensor data) | | |
| | | Improved data quality through advances in measurement and simulation capability | Integrated optimisation of supply chain | |
| | | | Extension of NPL Time & ID verification, alongside development of fraud/malpractice detection algorithms | Machine augmented learning & knowledge extraction from scientific documents |
| | | | Develop risk prediction and analysis models using multiple data sources/types | Determine new models of data storage, access, distribution that can allow new distributed economy to thrive under existing restrictions - or rewrite legislation |
| | | | Methods and statistics to estimate the uncertainty (and develop applications for) spatial-temporal models | |
| | | | Develop methodology/metrics to track latency across deployment scenarios and technologies, in order to identify 'hot' and 'cold' areas of the system | Storing and analysing data in a cloud and enable services for the manufacturer and market surveillance |
| | Commercial projects | Developing next-generation toolsets for data analytics | Statistical modelling for estimation of interactions beyond omics layers and for identification of key molecules, biomarkers, drug targets using trans-omics data | |
| | | | Prototyping IoT in the lab for context awareness. | |
| | Standardisation projects | Develop standards (and optimisation models) for data quality (incl. accuracy, confidence and fidelity) | | |
| | | Develop data (and metadata) provenance standards and requirements | | |
| | | Develop standards for data security | Develop standards (including ethics and pre-harvest/reconnaissance processes) and homogenous tools/ techniques for data collection (and use on large scale) | |
| | | Develop standards for sensor Precision and Calibration over Internet | | |
| | Other ideas | | Curation and annotation of very large datasets available for public and commercial usage | |
| Technologies and Capabilities | Measuring and transmitting data | Capability for modelling systems evolving over time (time series/tipping point/change point analysis,. Spatio-temporal systems, quality assured dynamic maps, data assimilation for environment monitoring applications) | | |
| | | Verified lineage of data and governance of the data from sensor to system | | |
| | | Standardistaion in metadata for sensor network systems (including data provenance assurance, records and automation of calibration, the effect of data curation methods) | | |
| | Storing and retrieving data | Capability for comprehensive uncertainty quantification in data integrity / provenance | | |
| | Data analytics | Algorithms for model discovery from multiple data streams, e.g. sparse reconstruction algorithms, robust PCA, tensor decompositions. | | |
| | | Capability for comprehensive uncertainty quantification in algorithms / computation (software standardisation and certification) | | |
| | | | Measuring and annotating data quality/fidelity in real time. Established methodologies to adjust data veracity in real time to meet the need for which the data is being gathered. | |
| | Partner capabilities & resources | Development of digital calibration certificates | | |
| | Other requirements | | Capability for uncertainty quantification for imaging systems (uncertainty methodologies in quantitative imaging, compressed sensing, sparse reconstruction, high-level feature extraction/classification, and sensor networks as irregular imaging systems) | |
| | | | | Standards and safety protocols for the next generation of AI and machine written software |
| | | | Development of training and skills plan to ensure the availability of appropriate resources to drive growth and innovation in data companies | |

**Figure 3 - Summary of workshop output of priority industry needs and challenges, proposed projects and crosscutting technologies and capabilities**

## ROADMAPS FOR PRIORITY PROJECTS

High-level roadmaps and summaries for the 15 priority projects are presented in this section. Each roadmap is introduced using the verbal summary given (during the workshop) by the group that developed it. The high-level roadmaps include the following fields:

A. Description of the project including the industry needs and challenges it directly addresses
B. The scope and boundaries of the project, explicitly indicating aspects that are included and excluded
C. Necessary research and technology development as well as important milestones that will indicate progress
D. Resources required for research and technology development including the funding mechanisms that may be relied upon over the lifetime of the project
E. Enablers and risks that may support or hinder progress
F. Immediate next steps to jumpstart project delivery

## A. DEVELOP STANDARDS (AND OPTIMISATION MODELS) FOR DATA QUALITY (INCLUDING ACCURACY, CONFIDENCE AND FIDELITY)

The vision here would be to have a unit for data quality as a real anchor in the industry. Such an output would enable businesses to win orders, because of their use of appropriate data quality tools and standards. It will also improve their productivity and create a competitive economy.

To achieve this, the sector needs to be working towards a framework, tools and standards to enable interoperability, assurance, trust and efficient use of data.

This needs to be an international, collaborative project, making the most of metadata tools and standards currently available. Also, it will need to develop publicly available specifications and best practice that can be fed into the international standardisation frameworks.

The first step would be to define the data quality characteristics and metrics. This will be followed by the development of tools and standards tested in real user scenarios to increase the level of trust and confidence.

Project: ***Develop Standards (and Optimisation Models) for Data Quality (including Accuracy, Confidence and Fidelity)***

| | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>To work toward an international framework for tools and standards for data quality, for interoperability assurance and efficiency<br><br>Industry needs/ challenges it addresses:<br>All industries, using data economically | Scope<br>**What's IN:**<br>Publicly available specification in international highly collaborative metadata standards; Publicly Available Specification (best practice); Metadata tools.<br><br>**What's OUT:**<br>Only enabling, not mining data; Data analysis only for demonstration. | **Desired future**:<br>A new SI unit for data quality; Businesses win orders because of data quality tools and standards; UK industry productivity and more competitive economy. |

| | ***Short term (+1 year)*** 2018 | ***Medium term (+3 years)*** 2020 | ***Long term (+5 years)*** 2022 | ***NEXT STEPS*** |
|---|---|---|---|---|
| **Required research/ technology development** | Current coverage of data quality standards (identify gaps & analogous work);<br>Partnering/collaborate (governance);<br>Data quality definition (characteristics, metrics);<br>Partners? – BSI, - Inspire | Testing in user cases;<br>Design & Develop data quality tools;<br>Global sensing & satellite centre (use case for EO data);<br>Data Fidelity Centre (metadata standards, quality, quantify quality for use);<br>Software for internal consistency of data. | Longer term framework;<br>A consensus on data quality metrics;<br>Machine learning (deep automation) of tools and standards;<br>National data hub in each country - collaborate & negotiate | *Immediate next steps:*<br><br>**Define data quality characteristics and metrics** |
| **Milestones** | Consortia building and network: IBM, Cisco, Microsoft etc.; Evidence capture; **Go NoGo** | Characteristics building; Internationalise project; Pilot on use cases | Community acceptance & verification; Fully deploy & feedback | |
| **Resource requirements** (people, equipment; prototyping, etc.) | People heavy;<br>Industry partners and buy-in;<br>Marketing (communicate!) | Back HPC & stor local (capital);<br>New and emerging technologies;<br>Front end industry scale | Standard as a service;<br>Flexibility of deployment of the tools & standards. | |
| **Funding mechanisms** | Universities; BEIS – DFC; I-UK; Commercial sector | Other NMIs (internationalise); Regional funding; DIT (was UKTI); IP/licence income | Self-sustaining model | |
| **Other enablers** | BSI; IoT alliance AIOTI standardization (EC project);<br>Digital Catapult; Open data Institute; IoT world alliance; Turing Institute | HVMC; Existing data service centres; PAASS IoT sensors; Other NMI (e.g. NIST, PTB) may have own alliances | Government | |
| **Risks** (& risk appraisal) | Risk of overlap with other organisations **(critical risk)**<br>Removal of funding from EU **(critical risk)**<br>Data quality standard is too generic in open community **(critical risk)**<br>[Government] level of interest reduces **(high risk)** | Industry level of interest reduces **(critical risk)**<br>Vested interest lack of collaboration **(medium risk)** | Lack of trust **(medium risk)** | |

**Project A – Develop standards (and optimisation models) for data quality**

## B. DEVELOP DATA (AND METADATA) PROVENANCE STANDARDS AND REQUIREMENTS

This project was explored from the perspective of information and data transfer management in complex supply chains. For example, in the automotive industry, where there are long supply chains with complex data and material flows into an individual business, there needs to be trust end-to-end. This can be achieved if data is future-proofed, has quantifiable trust levels and the solutions implemented are global. This will enable fast and nearly effortless decisions to be made by management and operators. There needs to be community-driven standards for these solutions to be implemented and used widely.

*Project:* ***Develop data (and metadata) provenance standards and requirements***

| | | | | |
|---|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>Enable automated data trust management across complex supply chains<br><br>Industry needs/ challenges it addresses:<br>Be able to use data in ***x*** years' time; quantify data quality to assure high quality information and decision making; standards for archival, metadata and searching of data; management/use/learning from legacy and available data; improved provenance of measurements, data and databases; integrity of data and software | Scope<br>**What's IN:**<br>Scale-free (small to big data); Methodology to enforce the standard; Testing, evolution and specialisation of W3C provenance; Open source tooling to accelerate adoption<br><br>**What's OUT:**<br>Domain-specific standards | **Desired future:**<br>Data & embedded trust levels for decision making; Future-proofing data | |
| | ***Short term (+1 year)*** *2018* | ***Medium term (+3 years)*** *2020* | ***Long term (+5 years)*** *2022* | ***NEXT STEPS*** |
| **Required research/ technology development** | Survey of metadata & provenance models; Survey of metadata management tools & technology; Study cases from specific industries (food, precious stones)<br>TARGET: SOTA (State-of-the-Art) in provenance management; Case study | Propagating trust over provenance graphs (inputs to outputs); System and user space instrumentation for provenance collection; Communicating trust to data consumers<br>TARGET: Guidelines; Demonstrate quantifiable value for provenance | From provenance to quantifiable trust levels<br>TARGET: Tools and services for provenance exploitation and analysis | *Immediate next steps:*<br><br>**Scoping study (National Centres)** |
| **Milestones** | Identify enterprise stakeholders; Demo case study basic provenance collection | | Stable standard with prototype tooling | |
| **Resource requirements** **(people, equipment; prototyping, etc.)** | Multiple industry sectors;<br>Private and public consortium – research phase;<br>International partners/global views | | | |
| **Funding mechanisms** | RCUK, EC Funding, Joint UK/US | | | |
| **Other enablers** | Heavy hitters on board | | | |
| **Risks** | No enterprise stakeholders **(medium risk)**<br>Big diverse membership => risk of divergence/high entropy **(medium risk)**<br>Remains "local" or too small scale **(high risk)**<br>No take up of standards **(high risk)**<br>Technology change leads to standards not applicable **(low risk)** | | | |

### Project B - Develop data (and metadata) provenance standards and requirements

## C. NEXT GENERATION INTEGRATION ALGORITHMS AND METHODOLOGIES FOR MULTIPLE DATA SOURCES

This project is about linking datasets from different data sources to add value to those datasets, to learn about the system or process, and to support intelligent decision-making.

The vision is to end up with a framework, a workflow, or a software system that would help non-experts combine data from different sources, together with some domain specific implementations. The research required would progress from initially physical systems and datasets that are largely in-house, to augmenting those with external data in the medium-term, and looking at how to incorporate and treat social and ecological data in the long-term.

This project requires collaboration between data owners and data generators as well as data analysts, software engineers and academia. The initial steps would be to build some relevant collaboration and explore the most appropriate funding mechanisms to support these.

*Project:* ***Next Generation Integration Algorithms and Methodologies for Multiple Data Sources***

| | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>Linking datasets from different sources (to add value, to learn about a system/process); Data analytics for intelligent decision support; Using data for one quantity as a surrogate for another; Data fusion and mining.<br><br>Industry needs/ challenges it addresses:<br>Quantify data quality to assure high quality decision making; how can data quality assure high quality information; reliable methods for combining data streams with different characteristics; machine learning for data processing and analytics; certification of trusted algorithms; visualisation of multiple image types to enable hybrid images. | Scope<br>**What's IN:**<br>Unknown data accuracies; - Different data accuracies, - Different data sampling; Different quantities (maybe different scales or semantics); Originating from several parties.<br><br>**What's OUT:**<br>Designing metadata for interoperability. | **Desired future**:<br>Framework and workflow to support non-experts;<br>Domain specific implementations. |

| | *Short term (+1 year)* 2018 | | *Medium term (+3 years)* 2020 | | *Long term (+5 years)* 2022 | | *NEXT STEPS* |
|---|---|---|---|---|---|---|---|
| **Required research/ technology development** | Physical systems and processes; In-house; Specify domain-specific problems; Understanding casualties and correlations; Machine learning I. | TARGET<br>Build repository of datasets; Model development (library) | Augmenting with external data; Machine learning II | TARGET<br>Add tools to repository | Social & ecological systems & processes; Machine learning III | TARGET<br>Validated by expert evaluation and peer review | *Immediate next steps:*<br>**Invite collaborators**<br>**Explore funding**<br>**Decide NPL's role** |
| **Milestones** | | Repository of datasets; model development library | | Additional tools to repository | | Validated repository | |
| **Resource requirements** (people, equipment; prototyping, etc.) | Consultation; Lab equipment; IOT sensors; validated data; Data analysts; Software engineers; Academia | | HPC | | Crowd sourcing | | |
| **Funding mechanisms** | Industry for domain-specific; Government for public good. | | Grants | | Self-sufficient system[community] – cost model? | | |
| **Other enablers** | People sharing data and expertise" including IBM, Microsoft etc.; Buy-in by major stakeholders. | | | | | | |
| **Risks** | Worrying about IPR | | Too generic to work | | | | |

### Project C - Next generation integration algorithms and methodologies for multiple data sources

## D. METHODS AND STATISTICS TO ESTIMATE THE UNCERTAINTY (AND DEVELOP APPLICATIONS) FOR SPATIAL-TEMPORAL MODELS

This project is closely linked to the "**Next Generation Integration Algorithms…"** project, in that it is looking at methods and statistics to estimate the uncertainty, and therefore develop applications associated with temporal and spatial modelling. To achieve this, a multiple-data-layer, multiple-data-source approach is required. A particular application of such a method would be to establish the environmental truth for a local area. Such a method would support decision-making in terms of future resilience in a number of different themes.

The development of such a method would enable the identification of the degree of confidence that could be obtained by combining different spatial and temporal layers with different resolutions. Initially, 'spatial' would be synonymous with 'geo-spatial'. Later the project will allow exploration of other factors, which might be spatially separated.

This project would address the challenges of extracting the maximum value from temporal and spatial data, especially where multiple types of data are combined. This will include compound uncertainty with multiple different datasets, different temporal parameters, and the associated different spatial parameters. It would exclude the confidence measure of the individual datasets, which should be addressed by a different project.

*Project:* ***Methods and Statistics to Estimate Uncertainty (and Develop Applications) for Spatial-temporal Models***

| | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>Identify degree of confidence of combination of different spatial and temporal layers with different temp & space resolutions; Independent & dependent sources | Scope<br>**What's IN:**<br>Propagation of uncertainty | **Desired future:**<br>Quantitative confidence estimate of a combined output;<br>Intelligent user - understanding of outcome |
| | Industry needs/ challenges it addresses:<br>Extract maximum knowledge and value - make decision | **What's OUT:**<br>Individual data sets confidence (given) | |

| | ***Short term (+1 year)*** *2018* | | ***Medium term (+3 years)*** *2020* | | ***Long term (+5 years)*** *2022* | | ***NEXT STEPS*** |
|---|---|---|---|---|---|---|---|
| **Required research/ technology development** | Multivariate time series analysis; Trends, seasonal & acyclic; Dynamical PCA factor analysis MGLMM regression modelling;<br>4D time & space; Identify other sources of uncertainty and combine this | TARGET<br>Data scientist;<br>Geospatial | Sensor technology structured; Software new AI/ML visualization predictive model | TARGET<br>Users involved;<br>Multi-spatial | | TARGET<br>Unstructured data | *Immediate next steps:*<br>**Identify data start point & customer requirements** |
| **Milestones** | | | | | | | |
| **Resource requirements** (people, equipment; prototyping, etc.) | Multi-disciplinary team;<br>ML & adequate computing;<br>Research | | Validation: User testing & validation;<br>Wider collaboration (share of resources) | | Commercial | | |
| **Funding mechanisms** | Government research fund | | Joint funding | | | | |
| **Other enablers** | | | | | | | |
| **Risks** | Data sustainability high;<br>Reliance on the model | | Updating capability & standards of data sources;<br>Interoperability of data sources | | | | |

### Project D - Methods and statistics to estimate uncertainty for spatial-temporal models

## E. APPLYING HPC, BIG DATA AND COGNITIVE SYSTEMS IN SCIENCE AND ENGINEERING

Within this project, relevant case studies and applications will relate to improving workflows and comparing experimental data with a series of tools that allow productivity improvements. When the data sources are heterogeneous, high performance computing in the form of simulation and data analytics tools or text analytic tools is important. It can assist cognitive advisers in the process, and generate insights on how they act as decisions makers.

Within the scope of this project is the building of algorithms and software, applications, workflows, knowledge portals and simulation tools that allow decision makers to optimise the decision-making process. Specific use cases could include a formulation workflow for pharmaceuticals or the use of graphene as a detector. These processes require the merging of various data sources, and two simulation tools could be developed – one for each process.

Ultimately, such a development could create much more productive knowledge/work for a researcher who is designing a new device, a new chemical, or new process.

*Project:* ***Applying HPC, Big Data and Cognitive Systems in Science & Engineering***

| | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>Framework that ingest data from heterogenous sources (simulations, sensors, instruments, etc.) aiming to increase productivity in development & optimisation of neurotechnologies;<br>The platform is an integrated set of tools for data analytics, presentation & interpretation | Scope<br>**What's IN:**<br>Scientific gateways optimisation techniques, machine learning, automated systems, knowledge portals; Algorithm development & optimisation including parallelisation, theoretical model development; Easy usage by non-experts | **Desired future:**<br>Deliver a system/appliance capable of support decisions for science based on simulation and real data; Demonstrate value for very specific use – cases agreed with experimentalist |
| | Industry needs/ challenges it addresses:<br>Speed up development; Increase productivity; Reduce trial & error; Understanding big volume of data | **What's OUT:**<br>Build new hardware infrastructure; Build new computing hardware | |

| | ***Short term (+1 year)*** *2018* | ***Medium term (+3 years)*** *2020* | ***Long term (+5 years)*** *2022* | ***NEXT STEPS*** |
|---|---|---|---|---|
| **Required research/ technology development** | Identify group of use-cases (min 3); Identify feasible methods to apply to data gathered from multiple sources; Examples: surface properties, graphene for gas sensors, targets for drug delivery, optimize power grid distribution | Select most successful proof-of-concept and engineering them into more robust product (increase TRL);<br>Integrate components, APIs, Optimize ease of use, 10-15 early adopters;<br>Iterate rate with early adopters to verify added functionality, productivity and ease of use | Framework, Application, Commercialize, Sustainable project, Increase adoption | *Immediate next steps:*<br>**Identify partners in academia, industries**<br>**Organize thematic workshop around various use cases with multiple stakeholders**<br>**Link to 'Machine augmented learning & knowledge extraction from (scientific) documents' project** |
| **Milestones** | Minimum Viable Product for each use-cases, evaluate impact in real scenarios | Prototype product or service working used by domain experts on a special computing platform | Black-box product (ISV) that can be used by anyone who can access input data sources (it can be sold as service or product to market) | |
| **Resource requirements** (people, equipment; prototyping, etc.) | Hardware Infrastructure (on sites), ML/DL frameworks, Storage ("Hot"/"Cold");<br>Build skilled and capable project team; Early adopters, internal; "Customer with a vision"/ Industrial & Scientific Advisory Board;<br>Skills: mathematical modelling, Software engineering, HPC, Domain expertise from end-users, experimentalist people | Increased engineering team and rebalance skillset based on use-case; Increase competing resources (simulation and validation) | Support team to promote and disseminate tools to partners and collaborators | |
| **Funding mechanisms** | Data fidelity centre; NPL strategic research funding; Hartree Centre (STFC), Industrial partners | Innovate UK, H2020 EPSRC | VC, Private funding | |
| **Other enablers** | | Japan – UK framework | | |
| **Risks** | | | | |

### Project E - Applying HPC, big data and cognitive systems in science and engineering

## F. STANDARDS FOR DATA SECURITY

This project seeks to develop standards for data security. The project vision is to facilitate data sharing using untrustworthy infrastructure. Data identity of data sources is important and this needs to be protected. The project scope includes data in transit or cached en route between devices during the sampling and the end point, but not data that rests at any of the bulk end points. This constrains the problem, as the two different data strands require two very different security solutions.

In order to achieve the required solutions within a reasonable timeframe, quantum resistance approaches maybe required. In the first year, some cases will need to be agreed with some generic standards to ensure a broad selection of suitable representative case studies. This will help identify appropriate industry partners that have a commercial incentive to collaborate in such a project. This project will not necessarily increase business revenues, but will reduce corporate risk.

*Project:* ***Standards for Data Security***

| | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>National standard for data security that is adopted by HMG; Data integrity, data privacy, data availability, authenticity/ID | Scope<br>**What's IN:**<br>What to standardise: Trace provenance, Process steps, Temporal issues, Tamper evident; Data in transit or cached | **Desired future:**<br><br>Standard adopted by … |
| | Industry needs/ challenges it addresses:<br>Facilitating data sharing using untrusted infrastructure | **What's OUT:**<br>No new security primitives, No new architecture; Data at rest <u>at end points</u> | |

| | ***Short term (+1 year)*** 2018 | | ***Medium term (+3 years)*** 2020 | | ***Long term (+5 years)*** 2022 | | ***NEXT STEPS*** |
|---|---|---|---|---|---|---|---|
| **Required research/ technology development** | Partner selection; Background research; Domain awareness; Select case studies; Threat modelling; Understand measurement architectures; Information sharing forum | TARGET<br>Overarching principles - link to cyber essentials | Concept standard; Use cases; Policy engagement; Create scenario specific test ranges; Revisit case studies and threat modelling | TARGET<br>Draft standard | Guide standard through process; Compliance testing – can this be driven from the standard?; Certification body? – BSI, - IETF, - Collaboration with NIST? | TARGET<br>Adopted/ Accepted/ Approved standard | *Immediate next steps:*<br><br>**Get on with it.**<br><br>**Business case for HMG funding.**<br><br>**Fund consortium.** |
| **Milestones** | | Overarching principles – link to cyber essentials | | Draft standard | | Adopted/ Accepted/ Approved Standard | |
| **Resource requirements** **(people, equipment; prototyping, etc.)** | £ (€?); Domain experts for case studies; Threat modellers; Workshops and White papers; BEIS (for academia) | | Range connectivity infrastructure (secure remote access); Labs for ranges; Demonstrators; Info sharing & training packages | | People | | |
| **Funding mechanisms** | HMG Grant; Active research programme in novel light weight quantum crypto | | KTP | | - Industry, - Grant funding, - Quango time | | |
| **Other enablers** | International alliances; Broad industry engagement; Industry group engagement in case studies | | | | | | |
| **Risks (& risk appraisal)** | Where is the commercial imperative?; SQUEP availability. | | Step change in technology (e.g. Quantum computing) **(high risk)** | | Competing standards (see policy engagement) **(low risk)**<br>Not adopted **(medium risk)** | | |

### Project F - Standards for data security

## G. MACHINE AUGMENTED LEARNING AND KNOWLEDGE EXTRACTION FROM [SCIENTIFIC] DOCUMENTS

The project vision is to be able to query structured and unstructured data sets regardless of the data sources and to receive information back.

The initial challenge with such a project is to extract the information in a computable format, which is difficult if the information provided is text. The second challenge is to take individual extracted text facts and assess which ones are true, and relevant to the question at hand. The challenge is to be able to take this type of data and provide answers that one might expect from an expert in a particular domain.

*Project:* ***Machine Augmented Learning & Knowledge Extraction from (Scientific) Documents*** *(Linked to 'Applying HPC, Big Data and Cognitive Systems in Science & Engineering')*

| | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>Develop the ability to ask any question of any data and get a quantitative answer | Scope<br>**What's IN:**<br>Scientific data (varying quality) from any source | **Desired future:**<br>Knowledge extraction from structured and unstructured data via NLP |
| | Industry needs/ challenges it addresses:<br>Machine learning for data processing and analytics; quantify data quality to assure high quality information and decision making; standards for archival, metadata and searching of data; management, use and learning from historical, legacy or available data; improved provenance of measurements, data and databases; research study and application of data science to data-driven materials design; high-speed algorithms for real-time analytics | **What's OUT:**<br>Non-digital data | |

| | ***Short term (+1 year)*** *2018* | ***Medium term (+3 years)*** *2020* | ***Long term (+5 years)*** *2022* | ***NEXT STEPS*** |
|---|---|---|---|---|
| **Required research/ technology development** | Metadata cataloguing for structured data;<br>Data extraction organization & annotation on unstructured data | Data and metadata integration & indexing;<br>NLP to understand data sources needed & computations required | Specific (quantitative) answer to natural language questions including both structured and unstructured data sources | *Immediate next steps:*<br><br>**Set up Working Group**<br><br>**Feasibility study (MLN)** |
| **Milestones** | Metadata standards (e.g. EFO) agreed; Partner with the likes of Elsevier of text processing; NLP of lab notebooks | Pilot Markov logic network approach; Pilot NLP with leading academic/ industry | Pilot integration | |
| **Resource requirements** **(people, equipment; prototyping, etc.)** | Initial PoL using 10-20 FTEs | Strategic Government Initiative | | |
| **Funding mechanisms** | HIS, Elsevier, ATI<br>NPL | H2020 | Google, Microsoft, IBM | |
| **Other enablers** | Computer Science Team UoFC | Computer Science Team UoFC | | |
| **Risks** | | | | |

### Project G – Machine augmented learning and knowledge extraction from [scientific] documents

## H. CURATION AND ANNOTATION OF VERY LARGE DATASETS

The curation and annotation of very large datasets is a broad problem that is normally addressed with machine learning technologies that require a sufficiently large or representative dataset. To make this problem manageable and real, medical data has been specifically discussed here.

Many universities and commercial entities have access to medical data, but frequently they are unable to share it due to anonymity requirements and possibly ethical regulations. However, derivations of this datasets can often be shared, so this roadmap examines how to generate a centralised dataset in order to share data whilst complying with legislation and ethics regulations.

In the short-term, it will be important to identify existing resources, legislation barriers, and infrastructure, and build a team. In the longer-term it would be important to identify different datasets and push them into a central resource so that people can share data.

To be sustainable in the long-term, a public-private funding model maybe required in the short-term with potentially free access for academic use and a fee for commercial entities in the longer-term. For academics this will provide the advantage of increasing citations, and for industry, the opportunity to access more data.

*Project:* ***Curation and Annotation of Very Large Datasets***

| | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>Facilitating broad use of data whilst maintaining anonymity and complying to ethics regulations<br><br>Industry needs/ challenges it addresses:<br>Machine learning for data processing and analytics; Medicine, climate, finance, security, humanisation/ personalisation of consumer products | Scope<br>**What's IN:**<br>Electronic data; Medical data; Understanding of legislation; Centralisation of data; Engaging broad research communities (Government/Academia/Industry)<br><br>**What's OUT:**<br>Non-UK (initially); Paper records | **Desired future:**<br>Framework utilised by all; A defined user group; Case study examples; Machine learning test set for common applications; A sustainable funding model (academics free use, commercial free to use) |

| | ***Short term (+1 year)*** *2018* | ***Medium term (+3 years)*** *2020* | ***Long term (+5 years)*** *2022* | ***NEXT STEPS*** |
|---|---|---|---|---|
| **Required research/ technology development** | Maintaining data provenance; Identify existing resources (datasets); Standardisation of formats, translation of formats; Specifying infrastructure; Identify legislative barriers; Build engaged community & define goals and objectives<br><br>TARGET<br>Completed planning and built distributed leadership team & identified resource requirements | Research community to create metadata which can be shared, e.g. segmentation, models etc.; Build infrastructure requirements; Research community online presence; Workshops to share & develop; Work on case studies<br><br>TARGET<br>Basis of working infrastructure, set of case study projects, a first set of centralised data & developed engaged team | Established an operational framework (designed to be extensible); Completed & published case studies; Open source framework launch with clear community guidelines (free/fee academics commercially)<br><br>TARGET<br>Citation of dataset increasing, Active researchers increasing | *Immediate next steps:*<br><br>**Identify core team (champion)**<br><br>**Identify critical barriers for data sharing;**<br><br>**Identify community – try to find commonalities**<br><br>**Identify data resources; Identify funding resource & vision for long-term sustainability (and thereafter, investigate legislative issues around data sharing).** |
| **Milestones** | Identify community<br>Identify & resources & barriers<br>Plan case studies | First release of collated resources<br>First workshop & build online presence<br>Start working on case studies | Completed:<br>• Case studies<br>• Operational framework<br>• Self-sustaining dataset | |
| **Resource requirements** (people, equipment; prototyping, etc.) | 2STEs to drive process;<br>Centralised data store. | 2 FTE to drive process (same people as before)<br>3 FTE to build data infrastructure | Total 5 FTE (as before) | |
| **Funding mechanisms** | Public-private partnership (NHS/NMS/Academia/ Industry) | | Licence/subscription model | |
| **Other enablers** | Early identification of case studies; Enhancing reputation of dataset owner; Maximise the re-use of existing datasets | Funders OA dataset generation oblige maximum re-use of data; Opportunity for industry to promote capabilities; Aligning agendas from research community & funders | Self-sustaining model | |
| **Risks** (& risk appraisal) | Complexity of legislation **(high risk)**<br>Lack of willingness of dataset owners **(high risk)** | Community is too fragmented/competitive **(high risk)** | | |

### Project H - Curation and annotation of very large datasets

## I. INTEGRATED OPTIMISATION OF SUPPLY CHAIN

This roadmap explores the integrated optimisation of supply chain with "just in time supply" approaches where costs are reduced without incurring downtime. The routing of certain products to different areas is also within the scope of this project as it can be facilitated by better use of data.

It will be important to review existing supply chains, and work with experienced managers to understand what is currently working well, where the bottlenecks are, and where improvements are required. The project should demonstrate the value in appropriate generalisation, but also visibility of the real constraints, and options on how to balance flexibility versus standards.

In terms of actual implementation, a lot of effort is required in metadata research with links to other areas, such as algorithms for optimisations, etc. The short-term milestones would be to gather and understand the actual requirements, and if possible to create a sandboxed or idealised demonstration to show the likely impact of this approach to potential users. In the medium-term, it will be important to demonstrate improvements to existing supply chains, and in the long-term, to be able to demonstrate fully optimised supply chains.

*Project:* ***Integrated Optimisation of Supply Chain***

| | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>Integrated optimisation of the supply chain. Start with existing supply chains (find experiences managers (guinea pigs))<br><br>Industry needs/ challenges it addresses:<br>Decision making from multiple sources of information; management use and learning from historical, legacy or available data; reliable methods for combining data streams with different characteristics | Scope<br>What's IN:<br>Showing value in appropriate generalisation; Awareness of real constraints, e.g. bulk order discounts etc. (transport costs); Balance flexibility vs standards; Data interop (compatibility) (supply chain logistics) – "Just in time"; Useful upstream info to suppliers e.g. sensitivity to supplier delays<br><br>What's OUT:<br>Sensor/source specifics; Detailed implementation inclusiveness except case studies | Desired future:<br>Demonstration of optimisation (cost of spares/inputs & down-time through supplier mismanagement leading to £ savings); Optimised routing of grades of product to produce lines (yielding cost savings and increased reliability); Human element (buy-in from individual supply managers & overcome supplier resistance) |

| | *Short term (+1 year)* 2018 | *Medium term (+3 years)* 2020 | *Long term (+5 years)* 2022 | **NEXT STEPS** |
|---|---|---|---|---|
| **Required research/ technology development** | Stakeholder engagement (understand challenges, current State-of-the-Art (SOTA), e.g. data stored, mine experts (e.g. in automotive ("just in time" and FMCG), context (business drivers); Metadata & data integration theory; Training. | Algorithms for optimisation; Input to standards/guidance & accreditation; Prioritisation of sampling frequency/priority | Autonomous decision making/ managing human intervention; Predicting likely supply chain performance | *Immediate next steps:*<br>**Industrial survey – where are the bottlenecks?**<br><br>**Initial assessment of likely impact & priorities (stakeholder survey)**<br><br>**Context SOTA review, supply chain management theory SOTA existing methods)**<br><br>**Initial training resource (overlap with other projects?)** |
| **Milestones** | Understand RQs & business drivers; Ensure relevant (NMI activity) complies with good practice; Sand boxed/ idealised demo to indicate financial impact | "Module"/ blocks prototypes ready; Realise/demo benefits for existing supply chains; Demo for impact. Ranking of supply processes | Demonstrate flexibility & commonality; Enable totally new improvements with financial impact | |
| **Resource requirements** (people, equipment; prototyping, etc.) | Skills (maths, software, logistics supply, expertise (current SOTA), artificial intelligence, data science, computer science); "A little bit of a lot of people" (multi-disciplinary teams, stakeholder committee); Representative data? - Development environment. | Equivalent to discretionary funding for SMEs; Test/use case owners & their resources (2+); Artificial Intelligence experts | Staff to maintain new infrastructure | |
| **Funding mechanisms** | Funding: Re-target existing projects (e.g. Empir); New – Innovate UK H2020 | H2020 | | |
| **Other enablers** | Facilitation to find & secure use case dinners; e-Training on basics; NPL product verification programme – model data? | Inability to track industry development (resource problem?) | | |
| **Risks** | Reluctance to disclose current practices; Key data not recorded (and hard to change – regulation); Supplier reluctance resistance to change; Routine changes in suppliers – back to Square 1 for parts of use cases | Re-inventing the wheel & overspecialised/ proprietary outputs; Outpaced by international competition? | | |

**Project I - Integrated optimisation of supply chain**

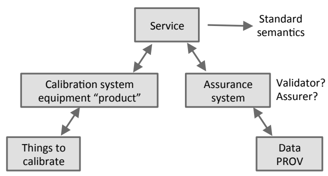
## J. DEVELOP TOOLS AND STANDARDS FOR SENSOR PRECISION AND CALIBRATION OVER INTERNET

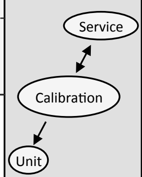
The goal of this project is to develop software as a service model where a trusted third-party would create a service that allows users to get sensors, run calibrators against these sensors, and report them back to the service. Any future user of these sensors could query the service to obtain the calibration data and their tolerances.

The first step would be to create a definition for this service, followed by defining all the parameters around it so that the quality of the calibration and/or tolerances could be established (maybe using a semantic model). The service implementation would be the first milestone.

This model could enable users to utilise other services also, for example any data source, which may or may not be based around the sensor data such as stocks, shares and commodities. It could enable the development of an assurance service, possibly integrating AI or similar algorithms. This would not change the overall service definition and over the long-term it could integrate services in security and provenance of data in a more unified service.

*Project:* ***Develop Tools and Standards for Sensor Precision and Calibration over Internet***

| | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>Data quality assurance over the internet<br><br>Industry needs/challenges it addresses:<br>Need to understand the pedigree & control of data; standardisation of sensor metadata, storage of sensor datasets, encryption of data to individual sensors and validation and governance of the data from sensor to analytics system; certification of trusted algorithms; quantifying data drift and its effect on data quality; trustworthy real-time data and information. | Scope<br>**What's IN:**<br>Realisation of SI units into factory floor;<br>Continue to calibrate sensors throughout 10 year life-span;<br>Profile of degradation over time<br><br>**What's OUT:**<br>Data security; Action based on output; Correcting erroneous data | **Desired future:**<br>Service & associated products to assure quality<br> |

| | *Short term (+1 year)* 2018 | *Medium term (+3 years)* 2020 | *Long term (+5 years)* 2022 | ***NEXT STEPS*** |
|---|---|---|---|---|
| **Required research/ technology development** | Service definition;<br>Standard/ semantics definition | IoT calibrator;<br>Products & services even AI;<br>Algorithmic QA | Extension to non-physical data;<br>Franchise the service? | *Immediate next steps:*<br> |
| **Milestones** | Service implementation | Example implementations | | |
| **Resource requirements** (people, equipment; prototyping, etc.) | Service definition specialists;<br>Web development | Product/instrument designers;<br>Partner companies FTEs | | |
| **Funding mechanisms** | Innovate UK? | Series A? | | |
| **Other enablers** | Requirements gathering;<br>Stakeholder engagement;<br>Business plan | Business launch/development | | |
| **Risks** | Credible user community;<br>Protection of 3rd party IP | Establishing trust | | |

### Project J - Develop tools and standards for sensor precision and calibration over internet

## K. IMPROVE DATA QUALITY THROUGH ADVANCES IN MEASUREMENT AND SIMULATION

This roadmap explores how to improve data quality through advanced simulation measurements, and how to develop simulation methods, software and an experimental system for supporting this domain. Domain examples are, for instance, robot intelligence or high speed of computers. A protocol of data simulation needs to be established. Subsequently, the simulation and measurement can be applied to the experimental system to improve the quality and quantity of data.

*Project:* ***Improved Data Quality through Advances in Measurement and Simulation***

| | | |
|---|---|---|
| **Project description/ scope** | Project summary description: Improve productivity by combining experiment & theory for breakthrough data science<br>Industry needs/ challenges it addresses: Quantify data quality to assure high quality information and decision making; research study and application of data science to data-driven materials design | Scope<br>**What's IN:** New (faster) simulation techniques (physical, data science, HPC); Matching measurement data like imaging & simulation parameters; NOUEL V&V & UQ for prediction; High throughput experiment with unknown uncertainty; Explore/define interface between raw data collection and its use (after processing)<br>**What's OUT:** Not about measuring physical parameters |
| **Desired future:** | Improved data quality with better outcomes | |

| | ***Short term (+1 year)*** 2018 | TARGET | ***Medium term (+3 years)*** 2020 | TARGET | ***Long term (+5 years)*** 2022 | TARGET | ***NEXT STEPS*** |
|---|---|---|---|---|---|---|---|
| **Required research/ technology development** | Use case development (materials); Relate to product dev process system engineering (V&V) process engineering; Specifying type of data to material scientists – what needs to be measured for simulation; Software systems; Big data (high throughput experiments, robotic, miniature/micro) | Working use cases | Making model (standard) system for computing simulation and experiment data | Establish protocol (general purpose) | Validated models. Virtual certification. Better prediction of the effects of uncertainty | | *Immediate next steps:*<br>**Turn this roadmap into a white paper to lead to funding** |
| **Milestones** | | Proof of concept (impact) | | Demonstrate impact | | | |
| **Resource requirements** (people, equipment; prototyping, etc.) | Industrial engagement from domain experts (post docs/ PhD);<br>HPC system (CPU and GPU);<br>Hardware (measurement, HPC) | | Software engineers | | Resource for implementation (user friendly) | | |
| **Funding mechanisms** | Exploratory funding → triage experiment and theory special joint | | Precompetitive Industry-Government collaboration | | Advanced development | | |
| **Other enablers** | Engagement/buy-in | | | | | | |
| **Risks** | Needs strong collaboration | | Use cases don't deliver | | Low impact | | |

### Project K - Improved data quality through advances in measurement and simulation

## L. DEVELOP RISK PREDICTION AND ANALYSIS MODELS USING MULTIPLE DATA SOURCES/TYPES

A successful outcome for this project would be guidelines on how to perform risk analysis in various scenarios, and a toolbox to assist the fusion of different sources of data.

In the short-term, the project should try to quantify and categorise risk, as well as identify the different data sources and types to generate real guidance for one sector, for example the communication sector. In the medium-term the communications data could be integrated with other systems such as autonomous vehicles, and the weather, or GPS data used therein. (This would then create a set of guidelines for risk analysis of autonomous vehicles.) In the long-term, a commercial product is envisaged, which will generate risk predictors, possibly with star ratings and NPL certified risk analysis on the data. Some research of how the analysis could be transferred between sectors, for example, from autonomous vehicles, to finance or energy infrastructure would also need to be included. Finally, public and/or industrial funding would be required for such a project.

| Project: ***Develop Risk Prediction and Analysis Models using Multiple Data Sources/Types*** | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>Developing risk prediction and analysis models using multiple data sources/types<br><br>Industry needs/ challenges it addresses:<br>Quantify data quality to assure high quality information and decision making; decision making from multiple sources of information; reliable methods for combining data streams with different characteristics (data type, uncertainty, etc.); drive toward probabilistic engineering; methods for propagating uncertainties through data curation methods | Scope<br>**What's IN:**<br>Cloud; Generic models; Publicly available data; Any data source (images, measurement data, admin data, calibration etc.); Data fusion; Standards on risk prediction.<br><br>**What's OUT:**<br>Collecting data; Sensors/development of hardware | **Desired future**:<br>Guidelines on how to do risk analysis in various scenarios;<br>Guide books and toolbox to fuse different formats of data |

| | ***Short term (+1 year)*** 2018 | | ***Medium term (+3 years)*** 2020 | | ***Long term (+5 years)*** 2022 | | ***NEXT STEPS*** |
|---|---|---|---|---|---|---|---|
| **Required research/ technology development** | Quantify risk, categorize risk; Identify data sources and types | TARGET<br>Guidance on risk for one sector e.g. communications protocol | Fuse communications and other sources for autonomous vehicles (e.g. weather, GPS);<br><br>From risk analysis to risk prediction | TARGET<br>Guideline on risk analysis of autonomous vehicle;<br>Apply models to different sector e.g. earth observation, infrastructure, finance | Generic models; Implementation in generic software | TARGET<br>Commercial product which generates a star rating e.g. NPL certified. | *Immediate next steps:*<br><br>**Find sectors interested.**<br><br>**Find people interested.**<br><br>**Find partners interested.** |
| **Milestones** | Company agrees to enter a partnership for medium term activities | 1st guideline on Risk Analysis for one sector, e.g. communications | | Demonstration to OEMs, DoT, Telecoms | | $ | |
| **Resource requirements** (people, equipment; prototyping, etc.) | Communications people (Electrical Engineering) Mathematicians, Cryptography expertise, Big computers, Machine learning; People from OEMs, Telecoms, ETSI; Engagement with Standard bodies | | Company to partner with; Autonomous car | | Social engagement; Cyber security | | |
| **Funding mechanisms** | EMPIR NPL & Horizon 2020 | | CCAV & partner company IP | | Partner companies from multiple sectors (IP) | | |
| **Other enablers** | Organizations/bodies already doing risk analyses; Government – start dialogue with society industry | | Dept. of Transport, OEMs, Telecoms, General public | | Legislation | | |
| **Risks** (& risk appraisal) | Lack of qualified individuals **(high risk)**<br>Funding **(critical risk)** | | Too sector specific **(low risk)**<br>Lack of partner company **(high risk)**<br>[Lack of] Industrial collaboration **(medium risk)** | | One size (model) fits all **(high risk)** | | |

**Project L - Develop risk prediction and analysis models using multiple data sources/types**

## M. NEXT-GENERATION ANALYTICS *(DEVELOPING NEXT-GENERATION TOOLSETS FOR DATA ANALYTICS)*

This roadmap explores how to develop the next generation of toolsets for data analytics. Data analytics is hard and difficult to implement effectively without data scientists, and there are not many data scientists available. The problem can be approached in two ways: training and employing more data scientists, or simplifying the job and trying to do more with less. One approach for the latter was to enable less specialised people or staff to create analytic solutions, and embed them into software solutions.

A key element on this approach would be around user interfaces, where drag and drop options could be provided for some types of analytic solutions. Starting small, it can capture the main industry and data scientist expertise in a plug-in model that could be integrated into a tool delivered to users, IT staff, or software developers, for example, in order to facilitate their job.

The key skills needed are data science and software development skills, in order to implement that knowledge in software.

| *Project:* **Next-Generation Analytics** | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>Interfaces for general users. Hide complexities; Replace "training" with "easier to use" tools; Integrating experiment, theory or simulation data; Help industrial scientist decide next experiment/study | Scope<br>**What's IN:**<br>New maths & algorithms; Decision trees; New (parallel/ distributed) hardware & software; Embed data analytics tools in existing software | **Desired future**:<br>Drag & drop data analytics "modules" into IT systems;<br>Gain data analytics capabilities required using existing people & skillsets;<br>Machine learning integrated in all data analytics software |
| | Industry needs/ challenges it addresses:<br>Skills shortage; Lower entry barrier to data analytics/ machine learning | **What's OUT:**<br>Development of underlying maths or techniques | |

| | ***Short term (+1 year)*** 2018 | ***Medium term (+3 years)*** 2020 | ***Long term (+5 years)*** 2022 | ***NEXT STEPS*** |
|---|---|---|---|---|
| **Required research/ technology development** | Identify missing skills for next steps;<br>Data import | Develop UK for IT people to drag/drop simple analytics solutions;<br>Plug-in abstraction | Commercial analytics solutions delivered without needing data scientists;<br>Outlier detection | *Immediate next steps:*<br>**Pick an exemplar;**<br>**Assemble partners ('supply', 'delivery', 'user')** |
| **Milestones** | | | | |
| **Resource requirements** (people, equipment; prototyping, etc.) | Data scientist;<br>DFC;<br>UI/tool developer;<br>Platforms | Domain expertise | More maths & algorithms | |
| **Funding mechanisms** | Innovate UK, EPSRC;<br>STFC, Hartree, Customers. | | | |
| **Other enablers** | | | | |
| **Risks** | Codifying data expertise is too hard<br>Insufficient funding | Lack of adoption<br>Unable to grow in scope | No ongoing development/ maintenance | |

### Project M - Next-generation analytics

## N. ETHICAL STANDARD(S) FOR TOTAL DATA LIFECYCLE *(DEVELOP STANDARDS (INCLUDING ETHICS AND PRE-HARVEST/RECONNAISSANCE PROCESSES) AND HOMOGENOUS TOOLS/TECHNIQUES FOR DATA COLLECTION (AND USE ON LARGE SCALE))*

The exploration of this project started quite broad, but was narrowed down to ethical standards for total data lifecycle. Various ethical issues were discussed on company acquisitions, autonomous vehicles, etc. The idea here is to have some standards for handling ethics. Before an international standard can be developed, an important step would be the setting up of a working group, and thereafter, running through some use cases.

Although an international standard would be good, it is complicated by different countries having different views on ethics. This is probably the biggest risk that this project might face as it would get complicated quite quickly. But this is an increasingly important issue, and the earlier it is addressed the more chances of success it would have. This project should be publicly funded initially with private support in its later phases.

*Project:* ***Ethical Standard(s) for Total Data Lifecycle***

| | | | |
|---|---|---|---|
| **Project description/ scope** | Project summary description:<br>To produce an 'ethics' standard framework and test it against ***x*** number use cases<br><br>Industry needs/ challenges it addresses:<br>Robotics & autonomous systems; ethical decisions not made by humans; Individuals (health and consumer privacy); Society (defence, security, health); Company (disclosure of company performance, trade secrets, IP, Confidentiality); standards for archival, metadata and searching of data; trustworthy real-time data and information; quality indicators of AI algorithm and the data it produces | Scope<br>**What's IN:**<br>Generalised framework; End goal is assurance; Must be independent of individual/ company/ sector/ company interests<br><br>**What's OUT:**<br>Technology agnostic | **Desired future** :<br>Companies to adopt to act in an ethical way; Drivers behaviour; Consumers understand ethical framework and use of their data |

| | ***Short term (+1 year)*** 2018 | ***Medium term (+3 years)*** 2020 | ***Long term (+5 years)*** 2022 | ***NEXT STEPS*** |
|---|---|---|---|---|
| **Required research/ technology development** | 1.What good practice exists?<br>2.Does it translate to other use cases?<br>3.Produce interim design guide<br>4.Apply across test scenarios (and refine)<br>5.Devise ethical framework/options for levels of moral consideration | International standard (kite mark) and adoption and inspection | Ongoing and repeat | *Immediate next steps*<br><br>**Find a champion/ leader**<br><br>**Government action to help initiate**<br><br>**House of Commons Enquiry to kick off?** |
| **Milestones** | Form working groups/committees/ governance; Publish and consult on interim guidance; Limited trial of beta rollout | Agree initial standard and publish; Widespread adoption and inspection | Revision | |
| **Resource requirements** (people, equipment; prototyping, etc.) | Devise framework: committee and networking and consultation time; £2m;<br>Test framework: data experiments, collection, design, programming, storage, & operational research;<br>Publish & route to market: education & training & assurance/inspection | Market should self-sustain.<br>Auditing and inspection.<br>Business to Business services. | | |
| **Funding mechanisms** | Public sector | Private sector | | |
| **Other enablers** | Political and regulatory; Analysis of value add/market advantage to encourage uptake; Consumer and public opinion and practice | | | |
| **Risks** | National security; H&S – individual/society;<br>Very difficult; Lack of consensus on what is 'ethical' or 'right' | Business case and added value difficult to make<br>May need a regulatory approach;<br>Public perception and/or lack of knowledge and apathy | | |

### Project N - Ethical standard(s) for total data lifecycle

## O. DETERMINE NEW MODELS OF DATA STORAGE, ACCESS, AND DISTRIBUTION THAT CAN ALLOW A NEW DISTRIBUTED ECONOMY IN MANUFACTURING TO THRIVE UNDER EXISTING RESTRICTIONS

This project tried to determine new models of data storage, access and distribution that may allow a new distributed economy to thrive under existing restrictions. This could enable the next industrial revolution, and has strong links from a data perspective with Project B (**Develop data (and Metadata) Provenance Standards…)** and Project E (**Applying HPC, Big Data and Cognitive Systems…)** projects, in terms of data provenance, quality, reliability and certainty.

The drivers for this project include the increasing need for mass customisation and small batch production, as well as high-value manufacturing. This results in a distributed IP creation or design, and production of items. In such a system, IP protection could be rendered worthless and so UK manufacturing needs to become more agile to out-think the competition.

Different technologies, such as additive layer manufacturing and others, will be required, but as this is a long-term project (25 to 35 years potentially) the technology options are uncertain. The risks could potentially be catastrophic for UK manufacturing, if UK manufacturing does not embrace a new phase – a new industrial revolution. Potential solutions could be the creation of a Data Fidelity Centre where a manufacturing supply chain will be created in a digital realm before transferring it into the real world.

Project: ***Determine New Models of Data Storage, Access and Distribution that can Allow New (More) Distributed Economy in Manufacturing (and raising productivity) to Thrive under Existing Restrictions – or Rewrite Legislation***

| | | | | |
|---|---|---|---|---|
| **Project description/ scope** | Project summary description: Distributed (resilience) & decentralised (strengthened) (supply chain); Stress testing of supply chains – confirm (or otherwise) resilience<br>Industry needs/ challenges it addresses: Mass customisation; Consumer products provider (Ultra responsive to 'Star Trek' canteen); Small batch high value (delivered at an acceptable price point!!) | Scope<br>**What's IN:** IoT (not the buzzword); Delinking production & IP generation; New market areas & applications; Cloud computing<br>**What's OUT:** New business models; Platforms (e.g. bespoke suit which fits); Software | **Desired future:** Innovation & productivity of UK manufacturing base; 2-sided business models | |
| **Required research/ technology development** | ***Short term (+1 year)*** 2018 | ***Medium term (+3 years)*** 2020 | ***Long term (+5 years)*** 2022 (& beyond) | ***NEXT STEPS*** |
| | Repeatability & reproducibility, e.g. ALM; Reputation/provenance/ fraud/ liability management | Trusted global catalogue of enabling information; Discoverability of customers/ suppliers; orders/ catalogues etc.; Additive manufacturing (1. Materials, 2. Machine/nearer "net shape" than current) | IP protection; End to end integrity | *Immediate next steps:*<br>**Lobby Government (understand societal, technical, employment risks and benefits)**<br>**Awareness raising (international competition)** |
| **Milestones** | Measurably more resilient supply chain; Distributed small-batch high value additive manufacture | | Digital marketplace coupled to distributed manufacturing | |
| **Resource requirements** (people, equipment; prototyping, etc.) | Additive manufacturers (Rolls-Royce etc.) | | | |
| **Funding mechanisms** | Direct government funding; Commercial funding | … Commercial | | |
| **Other enablers** | Incentive for radical business models; SME support; Stress testing of supply chains | | | |
| **Risks** | Near term specificity; Lack of government investment; Implementation of GDPR | Fraud; Industry decline/collapse; Resistance to change (or apathy); | Societal change; Quantum computing & issues for securing Data/IP/ Communications | |

**Project O - Determine new models of data storage, access and distribution**

## CONCLUSIONS

There were 89 participants from industry and academia that attended the UK Workshop on Data Metrology and Standards commissioned by the National Physical Laboratory, and delivered with the Universities of Cambridge and Huddersfield. The aim of the workshop was to engage UK industrial users of data to identify data measurement challenges and explore research project ideas to address them.

The following 15 projects were identified as priorities to respond to identified challenges. These projects were judged to be most important given the significant industry opportunity they potentially can open up, and also that they were reasonably achievable:

A. Develop standards (and optimisation models) for data quality (including accuracy, confidence and fidelity)
B. Develop data (and metadata) provenance standards and requirements
C. Next-generation integration algorithms and methodologies for multiple data sources
D. Methods and statistics to estimate the uncertainty (and develop applications for) spatial-temporal models (and best practice techniques/algorithms for analysis and modelling of sensor data (including data compression for storage of previously recorded sensor data)
E. Applying HPC, Big Data and Cognitive systems for decision support in chemistry, materials, life science and engineering discovery
F. Develop standards for data security
G. Machine augmented learning and knowledge extraction from scientific documents
H. Curation and annotation of very large datasets available for public and commercial usage
I. Integrated optimisation of supply chain
J. Develop tools and standards for sensor precision and calibration over internet
K. Improved data quality through advances in measurement and simulation capability
L. Develop risk prediction and analysis models using multiple data sources/types
M. Developing next-generation toolsets for data analytics
N. Develop standards (including ethics and pre-harvest/reconnaissance processes) and homogenous tools/techniques for data collection (and use on large scale)
O. Determine new models of data storage, access and distribution that can allow a new distributed economy to thrive under existing restrictions – or rewrite legislation

The workshop marked an important first step in evaluating the challenges and needs of UK industrial users of data, and the outputs of the project formulation will be used in developing NPL's data science research strategy. Engagement with NPL allows companies to leverage national data infrastructure, facilities and knowledge to maximise their investment in generating, understanding and using data far beyond the current level with private investment in big data facilities. To that end, NPL is committed to connecting with new partners and collaborators – for more information please contact datascience@npl.co.uk

## APPENDIX

### UN-SHORTLISTED INDUSTRY NEEDS/CHALLENGES AND PROJECTS

**Table 4 - Industry needs and challenges not voted for by workshop participants**

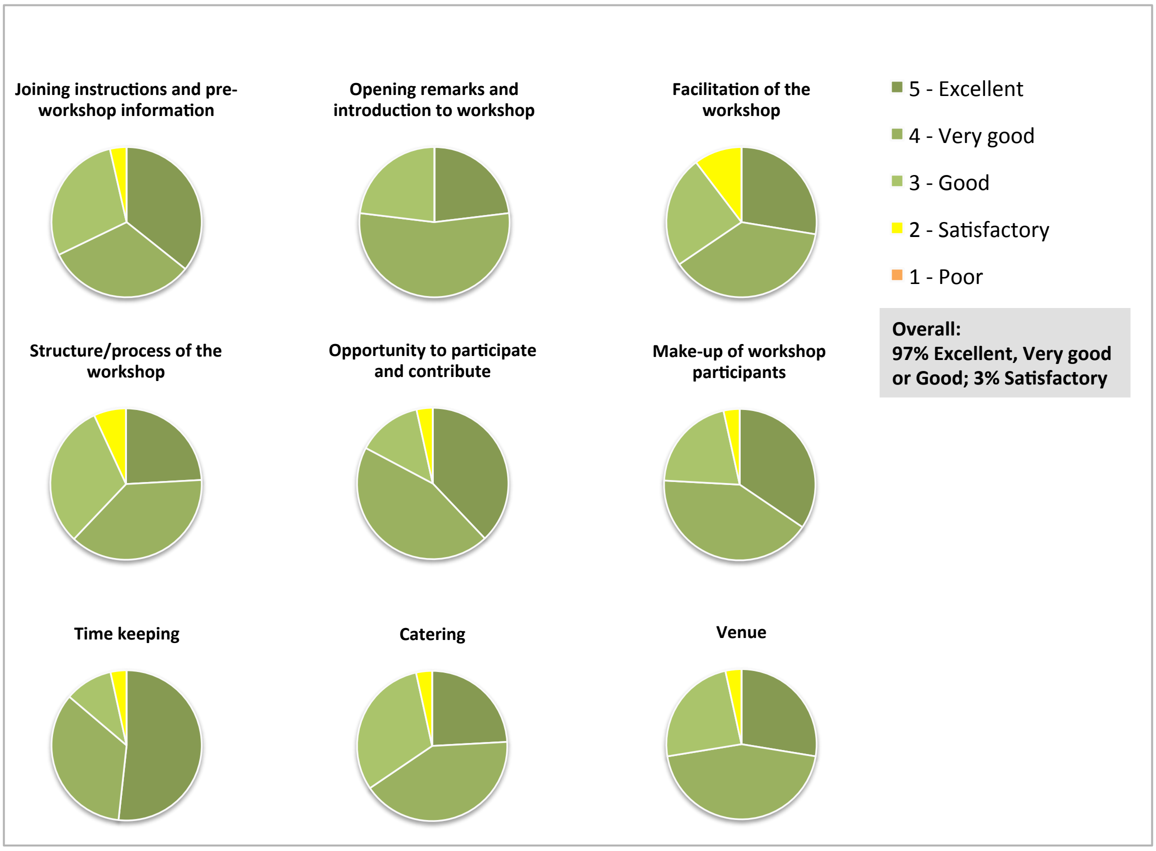
| Industry needs and challenges | Timescale | Votes |
|---|---|---|
| Continuous monitoring of pipeline flow rates using acoustics sensors and data | ST-LT | 0 |
| Monitoring of compressors on oil and gas platforms and liquefaction plants | ST-MT | 0 |
| Companies that do not evaluate and embrace new data technologies risk becoming uncompetitive | ST | 0 |
| Exascale computing application giving real time insight to complex problems with embedded smartness | LT | 0 |
| Recent ability to process unstructured data provides new capabilities that can be encoded in applications | ST | 0 |
| Using general purpose operating systems (e.g. Windows and Linux) | ST | 0 |
| Unregulated data as principle source of dynamic human information exchange vice info/data publishing/archiving | MT | 0 |
| Predictions using a suite of environmental variables and open data to reduce risks from environmental hazards | LT | 0 |
| Introduce digital trading schemes in real time | LT | 0 |
| Codification of best farming practice (and options lists) | ST | 0 |
| Augmenting human capabilities especially among least able | LT | 0 |
| Low-cost, high throughput measurement of multiple ‘omics data for biomedical research and suitable for cohort studies | ST-LT | 0 |
| Simulation of dynamic processes, estimation of interactions beyond ‘omics layers, and identification of key molecules, biomarkers, drug targets using trans-omics data | MT | 0 |
| Workforce including sales, geared towards traditional products | ST | 0 |
| Acceptance of new technologies making errors very low (when compared to human making errors at greater rates) | MT | 0 |
| Management of human judgement interventions for contextual (therefore) intelligent understanding - think people | ST-LT | 0 |
| How to comply with data import/export laws | ST | 0 |

## Table 5 - Proposed projects that did not make the shortlist

| Proposed projects (not shortlisted) | Timescale | Opp. votes | Feas. votes |
|---|---|---|---|
| Algorithms for real time data dependant modelling experiments and applications | MT | 5 | 0 |
| Analyse interface between data (algorithm, data format, etc.) by materials simulation and feasible data analysis methods | ST-LT | 3 | 0 |
| Constructing and subsequently implementing (with the ability to modify for cloud applications) a framework for measuring instruments based on virtualisation | ST-LT | 0 | 0 |
| Develop and promote standards/guidance material on data analysis for cyber-physical systems | ST-MT | 0 | 0 |
| Collaboration with institutes to establish accreditation curriculum in Data Science | ST | 0 | 0 |
| Create efficient algorithms which can process high frequency data in pipeline monitoring | ST-LT | 0 | 0 |
| Define and broaden stakeholder engagement and provide leadership in standardisation | ST-MT | 0 | 0 |
| Develop reliable and seamless multi-scale materials simulations by support of data science | MT | 0 | 0 |
| X-ray computed tomography as a metrology method for integration into adaptive machining | MT-LT | 0 | 0 |
| Activities akin to ExCape in other areas (e.g. microscopy) | ST | 0 | 0 |
| Adoption of new networking standards (e.g. White Rabbit equivalent) | LT | 0 | 0 |
| Case studies to deliver impact e.g. complex system uncertainty and integrity assessment to drive decision making | ST | 0 | 0 |
| Data filtering/classification to determine temporal 'tick' of data type. | LT | 0 | 0 |
| Develop degree accreditation scheme (like BCS) in collaboration with industry. | ST | 0 | 0 |
| Identity of Entities project in the context of BIM Level 3 | ST | 0 | 0 |
| Internal data analytics training and mentorship programme | ST | 0 | 0 |
| Outreach to policy-makers | ST | 0 | 0 |
| Standardisation in digital/online age verification | ST | 0 | 0 |
| Standardisation of metrics and descriptors used for dose data | ST | 0 | 0 |
| Measurement of emissivity within non-equilibrium processing | ST | 0 | 0 |
| Policing and high profile prosecutions of trolls and phishers | ST | 0 | 0 |
| Dissemination of expert systems (algorithms and information) | MT | 0 | 0 |
| Extend existing data hubs with Web of Data Capabilities and spatio-temporal support (e.g. GeoSPARQL) | ST | 0 | 0 |
| Development of algorithms for diagnosis of patients using trans-omics data | LT | 0 | 0 |

## WORKSHOP FEEDBACK

There were 29 workshop participants who provided feedback. Of these 96 per cent considered the workshop to be excellent, very good or good. The detailed feedback is shown below.



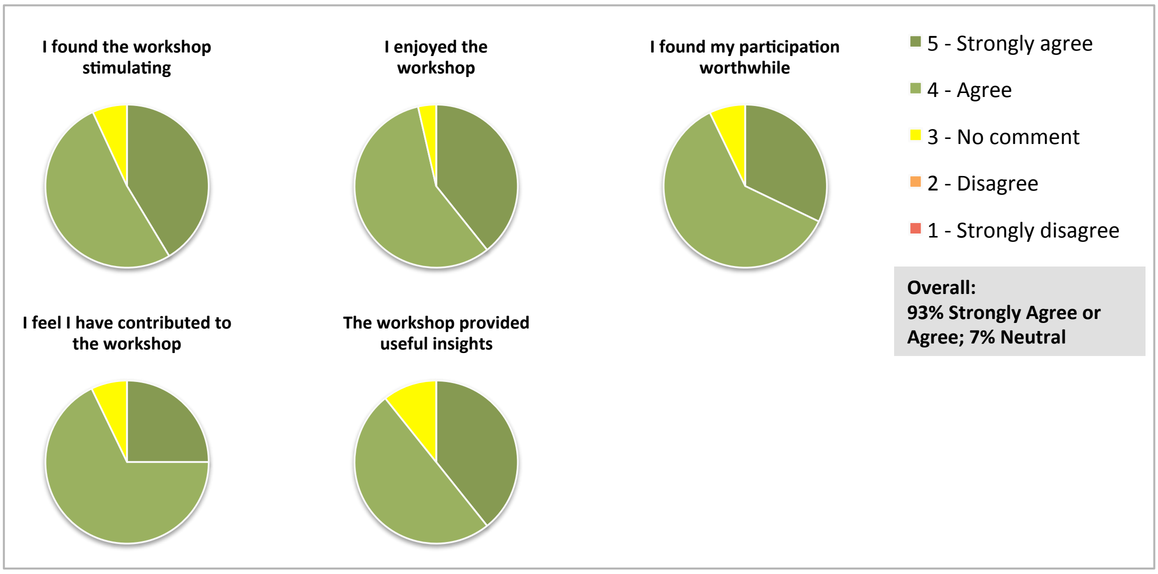




Figure 4 - Workshop participants' feedback

## WORKSHOP PARTICIPANTS

**Table 6 - List of workshop delegates and their respective organisations**

| Name | Organisation |
|---|---|
| Claire Donoghue | 3M |
| Suzanne Shea | 3M |
| Ralph Ecclestone | Access Cambridge |
| Hisao Nakamura | AIST |
| George Dibb | All-party Parliamentary Group on Data Analytics |
| Claus Bendtsen | AstraZeneca |
| Julian Hill | BAE Systems Applied Intelligence Ltd |
| Hugh Boyes | Bodvoc Ltd, Warwick Manufacturing Group |
| James Aston | BRE |
| James Gbadamosi | BRE |
| Richard Wishart | Delivery Management Ltd |
| Paul Galwas | Digital Catapult |
| Stuart Homann | Environment Agency |
| Monty Mountford | FREMO Ltd |
| Steve Morgan | Fujitsu |
| Steven Wilson | GCGP Enterprise Partnership (LEP) |
| Simon Thornber | GSK |
| Andy West | GSK |
| Michael Gleaves | Hartree Centre |
| Martyn Winn | Hartree Centre |
| Glenn Martyna | IBM |
| Nigel Rix | Innovate UK |
| Matt Sansam | Innovate UK |
| Mark Wharton | Iotic Labs |
| Kris Kobylinski | Jaguar Land Rover |
| Liqun Yang | KTN |
| Yasuhide Fukumoto | Kyushu University |
| Hiroyuki Sasaki | Kyushu University |
| Jim, Roche | Lenovo UK |
| Ruth Boumphrey | Lloyd’s Register Foundation |
| Mike Dewar | NAG |
| Robert Hanisch | NIST |
| John Bancroft | NPL |
| Elena Barton | NPL |
| Sunny Bhandari | NPL |
| Andy Blackmore | NPL |
| Lindsay Chapman | NPL |
| Stephane Chretien | NPL |
| Alistair Forbes | NPL |
| Nigel Fox | NPL |
| Ian Gilmore | NPL |

| | |
|---|---|
| Peter Harris | NPL |
| JT Janssen | NPL |
| Christopher Jones | NPL |
| Amir Kayani | NPL |
| Lisa Leonard | NPL |
| Valerie Livina | NPL |
| Ric Parker | NPL |
| Stephen Robinson | NPL |
| Ivan Rungger | NPL |
| Sophie Smith | NPL |
| Peter Thompson | NPL |
| Jenny Wooldridge | NPL |
| Rob Woollin | NPL |
| Vibin Vijay | OCF |
| David Yip | OCF |
| Daniel Peters | PTB |
| Henning Baldauf | QuoData |
| Ron Bates | Rolls Royce |
| Michael Cunningham | Rolls Royce |
| Pete Loftus | Rolls Royce |
| Elizabeth Quaglia | Royal Holloway University London |
| Mark Halling-Brown | Royal Surrey County Hospital |
| Mishal Patel | Royal Surrey County Hospital |
| Tim Park | Shell |
| Bryan Edwards | STFC |
| Amanda Lane | Unilever |
| Pete Davies | Uniper |
| Nathan Gould | Uniper |
| Paul Alexander | University of Cambridge |
| Yin Chang | University of Cambridge |
| Clare Dyer-Smith | University of Cambridge |
| Ayat Fekry | University of Cambridge |
| Alan O'Neill | University of Cambridge |
| Mark Reader | University of Cambridge |
| Filippo Spiga | University of Cambridge |
| Tien-Chun Wu | University of Cambridge |
| Grigoris Antoniou | University of Huddersfield |
| Andrew Ball | University of Huddersfield |
| James Devitt | University of Huddersfield |
| John Remedios | University of Leicester |
| Paolo Missier | University of Newcastle |
| Hongjie Ma | University of Portsmouth |
| Michael Grinfeld | University of Strathclyde |
| James Irvine | University of Strathclyde |

| | |
|---|---|
| Blair Johnston | University of Strathclyde |
| Jiazhu Pan | University of Strathclyde |
| Greig Paul | University of Strathclyde |
| Robert Elliott | University of Surrey / NPL |

### Table 7 - Participant groupings for exploring the fifteen priority projects

| Project | | Participants |
|---|---|---|
| A | Develop standards (and optimisation models) for data quality (incl. accuracy, confidence and fidelity) | Vibin Vijay, John Remedios, Robert Elliott, James Devitt, Valerie Livina |
| B | Develop data (and metadata) provenance standards and requirements | David Yip, Kris Kobylinski, Andrew Ball, Paolo Missier, Alistair Forbes |
| C | Next-generation integration algorithms and methodologies for multiple data sources | Mark Reader, James Gbadamosi, Amanda Lane, Blair Johnston, Peter Harris |
| D | Methods and statistics to estimate uncertainty (and develop applications) for spatial-temporal models | Monty Mountford, Henning Baldauf, Stuart Homann, Jiazhu Pan, Elena Barton |
| E | Applying HPC, Big Data and cognitive systems for decision support in chemistry, materials, life science and engineering discovery | Mike Dewar, Yasuhide Fukumoto, Michael Gleaves, Filippo Spiga, Ivan Rungger |
| F | Develop standards for data security | Elizabeth Quaglia, James Aston, Hugh Boyes, Julian Hill, Ric Parker |
| G | Machine augmented learning and knowledge extraction from scientific documents | Michael Cunningham, Claus Bendtsen, Simon Thornber, Andy West, Stephane Chretien |
| H | Curation and annotation of very large datasets available for public and commercial usage | Claire Donoghue, Nigel Fox |
| I | Integrated optimisation of supply chain | Liqun Yang, Tim Park, Grigoris Antoniou, James Irvine, Christopher Jones |
| J | Develop tools and standards for sensor precision and calibration over internet | Mark Wharton, Ron Bates, Robert Hanisch, Ian Gilmore |
| K | Improved data quality through advances in measurement and simulation capability | Hiroyuki Sasaki, Hisao Nakamura, Pete Loftus, JT Janssen |
| L | Develop risk prediction and analysis models using multiple data sources/types | Alan O'Neill, Sascha Eichstaedt, Suzanne Shea |
| M | Developing next-generation toolsets for data analytics | Steve Morgan, Martyn Winn, Michael Grinfeld, John Bancroft |
| N | Develop standards (including ethics and pre-harvest/reconnaissance processes) and homogenous tools/techniques for data collection (and use on large scale) | Ruth Boumphrey, Nathan Gould, Stephen Robinson |
| O | Determine new models of data storage, access and distribution that can allow new distributed economy to thrive under existing restrictions - or rewrite legislation | Daniel Peters, George Dibb, Greig Paul, Paul Galwas, Rob Woollin |

### Table 8 - Workshop facilitators

| Name | Organisation |
|---|---|
| Imoh Ilevbare | IfM Education and Consultancy Services Ltd.<br>Institute for Manufacturing<br>University of Cambridge |
| Nicky Athanassopoulou | |
| Michèle Routley | |
| Rob Munro | |

### Table 9 - Workshop steering group

| Name | Organisation |
|---|---|
| Jenny Wooldridge | NPL |
| Lindsay Chapman | NPL |
| Lisa Leonard | NPL |
| Alistair Forbes | NPL |
| Ian Gilmore | NPL |
| JT Janssen | NPL |
| Sundeep Bhandari | NPL |

**National Physical Laboratory**
Hampton Road
Teddington
Middlesex
United Kingdom
TW11 0LW
www.npl.co.uk/contact

**IfM Education and Consultancy Services Limited**
Institute for Manufacturing
Department of Engineering
17 Charles Babbage Road
Cambridge
CB3 0FS
http://www.ifm.eng.cam.ac.uk/services/overview/

IfM ECS works with companies of all sizes to help them create and capture value, and with national and regional governments to support and help grow their industrial sectors. It does this by transferring the new ideas and approaches developed by researchers at the IfM through a programme of education and consultancy services. IfM ECS is owned by the University of Cambridge. Its profits are gifted to the University to fund future research activities.