Research Challenges in Data Science

(a SICSA Data Science Event)

23rd November 2017

Authors:
Özgür Akgün, Youssef Al Hariri, Laila Alabidi, Oana Andrei, Evgenij Belikov, Elvis Bernard, Sheriffo Ceesay, Mike Chantler, Jessica Chen-Burger, Laura Cram, Vincenzo Crescimanna, Monica Dearden, Ruth Falconer, Muyao Fan, Azimeh Gharavi, Dimitra Gkatzia, Alasdair Gray, Tristan Henderson, Matthew Higgs, Robin Hill, Alistair Lawson, Hans-Wolfgang Loidl, Michael Lones, Mila Maidarti, Tom Martin, Thomas Methven, Yashar Moshfeghi, Nikos Ntarmos, Stefano Padilla, Jeff Pan, Cyril Pernet, Andrea Rosales, Usman Sanusi, Ian Simpson, Vanessa Zervogianni, Yajie Zhu

sicsa* The Scottish Informatics & Computer Science Alliance

Contents

Introduction

Data Science is increasingly being accepted as one of the crucial technologies for the wellbeing and prosperity of nations. As Data Science can be the source and enabler of large-scale social and commercial change, finding the research challenges in Data Science is now a critical aspect for most scientific research and businesses. Data Science is also the subject of considerable investment by the Scottish Government, the United Kingdom, and the European Union.

Three years ago SICSA[1] held the inaugural meeting of their Data Science theme. It surveyed the research challenges at that time (July 2014) and found nine overarching challenges (https://www.well-sorted.org/explore/SICSADataScience). Much has changed since then, however, and in response to informal conversations with organisations such as the Data Lab[2], Scottish Enterprise, and EPSRC[3], we saw the need to refresh the Scottish Research Challenges in Data Science from the view of researchers.

This document aims to provide a concise overview of what are the major Data Science research challenges over the next five years and their possible impact. We did so by collecting feedback from researchers specialised in Data Science from around Scotland and discussing the various challenges during a one-day event in Edinburgh. This document will be split into three sections:

- A section detailing the pre-meeting activity which was done via Well Sorted[4]
- A section detailing the morning activity, where attendees generated synthesised research challenges for each coloured group created in the Well Sorted process.
- A section detailing the afternoon activity, where groups suggested the possible impact of research in each coloured group.

[1] The Scottish Informatics and Computer Science Alliance. (http://www.sicsa.ac.uk/)
[2] The Data Las is one of the eight innovation centres funded by the Scottish Funding Council. (https://www.thedatalab.com/)
[3] The Engineering and Physical Sciences Research Council. (https://www.epsrc.ac.uk/)
[4] Well Sorted is a free (for academic use) tool which allows for remote, collaborative idea suggestion and grouping. Each attendee submits ideas, and then individually sorts them into groups. Well Sorted then creates an 'average' group which accounts for everyone's opinions equally. See www.well-sorted.org for more details.

Meeting Agenda

10:00	Registration & coffee
10:20	Event opening and introduction – Mike Chantler, Strategic Futures Laboratory
10:35	Overview of recent activities – Gillian Docherty, The Data Lab
11:05	Breakout session – Identifying and expanding data science challenges
11:50	Ten Minute Break
12:00	Government data sources and challenges – Peter Winstanley, The Scottish Government
12:30	Speed feedback from challenges
13:00	Lunch break
14:00	Breakout session – Potential impact and sectors of identified challenges
14:45	Speed feedback from impact and sectors
15:15	Mixer of challenges and impact
15:50	Next steps, Summary, and SICSA Opportunities
16:00	Close

Pre-Meeting Well Sorted Activities

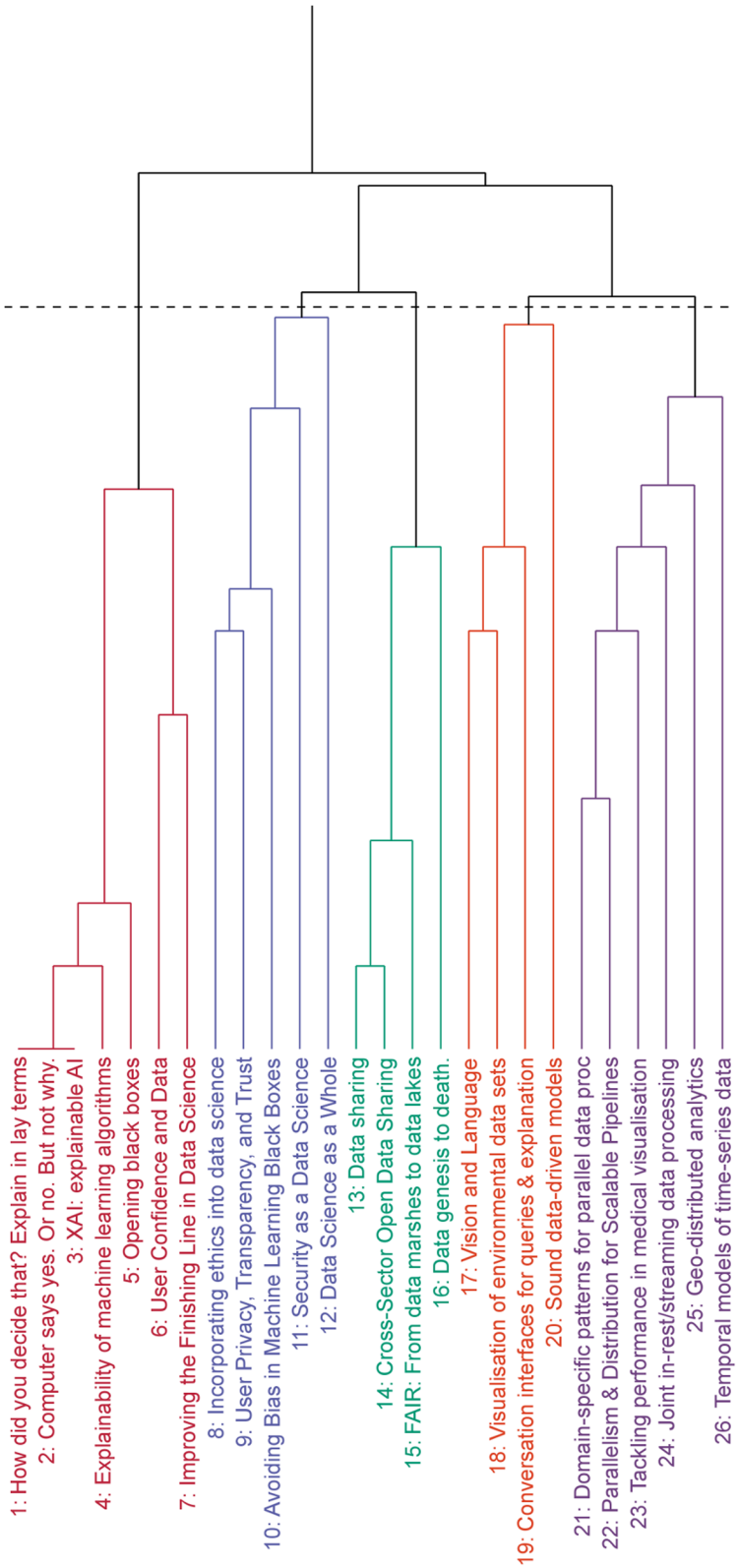
In this section, we'll present the output from the Well Sorted activity which we performed in a month running up to the meeting itself. This consisted of the following steps:

1. Attendees were asked to submit two (2) ideas which answered the following question: "Please describe two research challenges that you would passionately argue are important for Scottish data science?"
   These ideas consisted of a less than 50 character title, and a less than 255 character description.
2. Once all the ideas had been submitted, attendees were asked to individually sort them into groups however they saw fit.
3. Once all attendees had the chance to sort the ideas, Well Sorted created an 'average' grouping which will be shown in this section. As meeting organisers all we did was choose the number of groups, a number we set to five (5).
4. Attendees were asked to select a coloured group that they wished to discuss during the morning and afternoon breakout sessions. Attendees' names will show which groups they attended.

In the following subsections, we present the output of the Well Sorted visualisations, as they were shown to attendees before and during the meeting.

Dendrogram

This tree shows each submitted idea and its similarity to the others. The lower two ideas 'join' the more people grouped those two ideas together. For example, if two ideas join at the bottom, every person grouped those two together.

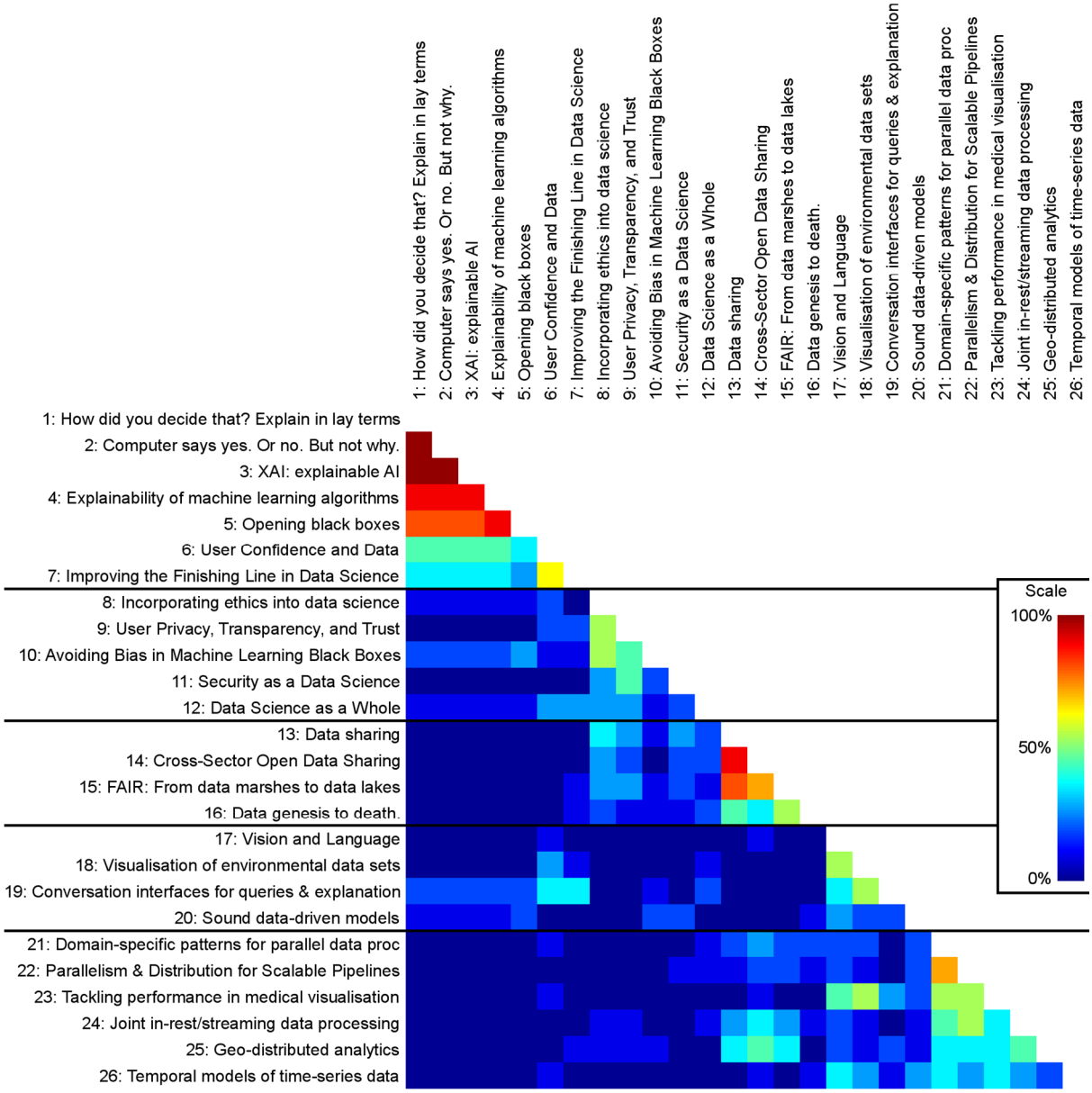


Tree Map

This presents an 'average' grouping. It is calculated by 'cutting' the Dendrogram at the dashed line so that any items which join lower than that line are placed in the same group. In addition, rectangles which share a side of the same length are more similar to each other than their peers.

Heat Map

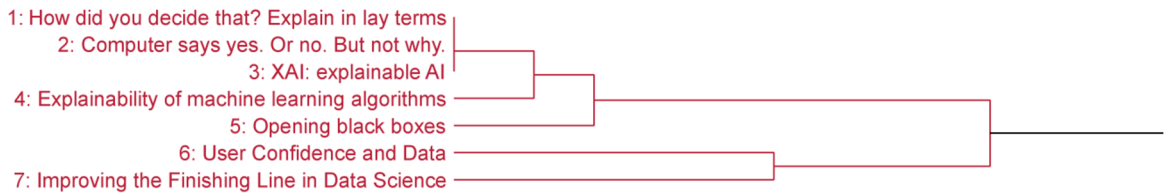
This shows a similarity matrix where each idea is coloured at the intersection with another idea, showing how similar the two are. This is useful to see how well formed a group is. The more red there is in a group (shown by the black lines), the more similar the ideas inside it were judged to be.



Raw Group Data

This shows every submitted idea and its longer description. They are shown in the same order as the Dendrogram (so similar ideas are close to each other) and split into the coloured groups used in the Tree Map. In addition, each idea has been given a unique number so they are easier to find.

Red Group

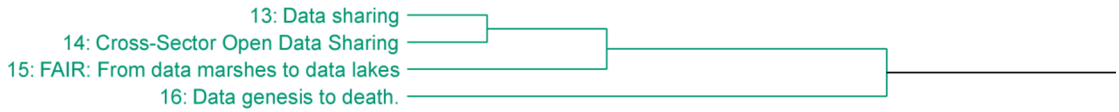


RED	1	How did you decide that? Explain in lay terms	Whether it is the result of a machine or deep learning algorithm, or a query answer over multiple sources, we need effective ways to explain in easily comprehensible terms the effect of our data science techniques
	2	Computer says yes. Or no. But not why.	Trust extends from transparency, comfort and acceptance. Data science requires more than just faith. Decision processes and output need to make sense; ultimately offering justified choices, adaptive planning and degrees of belief/accuracy (XAI).
	3	XAI: explainable AI	Ability of data mining and visualisation systems to explain decisions, layouts and focuses to various classes of user (researcher, developer, stakeholder, naive user).
	4	Explainability of machine learning algorithms	Deep learning algorithms employ thousands of layers to model data. These layers are "black boxes" for the researchers. Finding a way to understand and explain the derived models would be beneficial.
	5	Opening black boxes	Complex data mining or data mapping processes often rely on black box process. It could be beneficial for users to get more details about algorithmic decisions.
	6	User Confidence and Data	It should be important to consider the 'naive' (or non-technical) user perspective and focus on understanding the mechanisms involved in user confidence. Especially when users are given results (e.g. data visualisations) as is.
	7	Improving the Finishing Line in Data Science	We have grasp crunching data early in the process, we mastered analysing in the middle, but often overlook the decision step. The challenge is to improve this crucial last decision-making step allowing us to reinforce confidence within our stakeholders.

Blue Group

BLUE	8	Incorporating ethics into data science	There is much concern about the societal and legal effects of data science, for instance when used in employment, policing or media (e.g. Brexit). We need to understand if, and how, to embed ethical considerations at all stages of the data science process
	9	User Privacy, Transparency, and Trust	While we all want to take advantage of the opportunities of big data and new methods of data processing, how do we do this in such a way that we avoid damaging user privacy, or the tacit agreement and trust between data provider and processor.
	10	Avoiding Bias in Machine Learning Black Boxes	Recently there have been many examples of machine learning deciding on rather unfortunate things, like the racist Microsoft Chatbot, or Google's homophobic Sentiment Analyzer. How do we design these systems while avoiding encoding our own bias into them?
	11	Security as a Data Science	Scotland is small enough that Government, Universities, industry and medical, policing & security forces can come together with real problems, data share and find solutions. Data mining and AI techniques can aid malware detection.
	12	Data Science as a Whole	Data Science comprises many stages from defining a problem, collecting and analysing data, to high-level decision making. The challenge is to improve on Data Science as a whole to enhance reliability, confidence and trust across the entire process.

Green Group

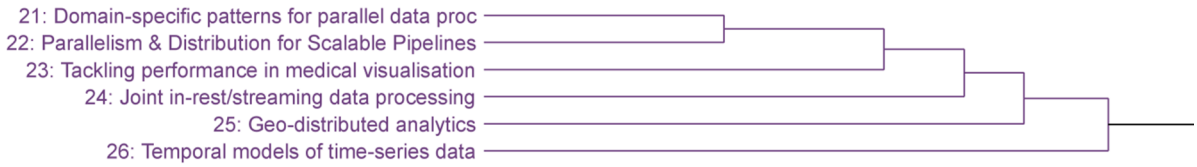


GREEN	13	Data sharing	Improved data sharing would help many data science projects, but we need frameworks for responsible, accountable, legal and ethical data sharing that provide appropriate incentives to both data producers and data users.
	14	Cross-Sector Open Data Sharing	Scottish Enterprise estimate the potential benefits for economy from cross-sector data sharing at £15-20 billion in the next years. To facilitate sharing, a general data-sharing API, similar to the Open Banking API, should be defined and standardized.
	15	FAIR: From data marshes to data lakes	Getting all your data in one place solves the Accessibility problem but you still need to Find the relevant data and get it to Interoperate in order to enable Reuse.
	16	Data genesis to death.	Meaningful datasets are dynamic and not isolated or frozen in time. Interpretation and contextualisation requires a complete understanding of the lifecycle from conception, through growth, maturity and finally redundancy (graceful degradation if lucky).

Orange Group

ORANGE	17	Vision and Language	Being able to bridge the gap between vision algorithms (object recognition, features recognition, etc.) and language generation (provide descriptions of images).
	18	Visualisation of environmental data sets	The growth of VR could be harnessed to visualise geographical, marine and environmental data to enable not just environmentalists but archaeologists, farmers and school children to see the world around them.
	19	Conversation interfaces for queries & explanation	Conversational interfaces for queries, focusing, and explanation of visualisation and data mining tools.
	20	Sound data-driven models	Create sound data-driven models amenable to rigorous analysis and verification techniques

Purple Group



PURPLE	21	Domain-specific patterns for parallel data proc	To efficiently leverage the power of parallel processing on big data sets across application domains, a high-level programming approach is needed, that is easily accessible to non-specialist users. We argue for a cloud-based, domain specific patterns.
	22	Parallelism & Distribution for Scalable Pipelines	To address the 4 Vs of Big Data in a scalable way, novel computer architectures and programming languages should be co-designed and implemented where parallelism, distribution and fault-tolerance are first-class considerations, not added as afterthoughts.
	23	Tackling performance in medical visualisation	The latest generation of 7T MRI scanners, deployed in Glasgow this spring, increases the data volume per scan by a factor of 27, through increased resolution. Parallel processing is needed to tackle the performance challenges in medical visualisation.
	24	Joint in-rest/streaming data processing	How can we build a system that can process queries over both in-rest and streaming data, with complex interdependencies? How can we detect feedback loops in our data/control flows and harness them to improve the performance and quality of our results?
	25	Geo-distributed analytics	How can we support analytic queries on data partitioned/distributed across different datacentres, when transferring all of the data to a central location is not an option (e.g., due to bandwidth/storage/administrative constraints).
	26	Temporal models of time-series data	Analysis of temporal models of time-seried data generated by some interactive system.

Morning Session: Research Challenges

On the morning of the event we asked attendees to create 'breakout' groups, by joining up with everyone else who had selected the same colour as them at registration. For this first morning session, attendees were given 45 minutes to synthesise the ideas which fell within their colour's group and come up with three (3) research challenges. Afterwards, one member of the group acted as a raconteur and presented these challenges in a 60 second quick-fire session.

In this section, we present the challenges which were presented by each group, along with any additional comments they provided in the week after the meeting.

Red Research Challenges

XAI (eXplainable Artificial Intelligence)

Mike Chantler, Jeff Pan, Andrea Rosales, Robin Hill, Usman Sanusi, Michael Lones, Laila Alabidi, Özgür Akgün, *Oana Andrei*

Model Simplifications For Explainability

- Going from models which are good for computational efficiency to models which are good for human understanding
- Incorporation of human expertise and theory and domain knowledge
- Improve stakeholders trust and confidence
- Explaining to 3rd parties about decisions with consequences
- Explanation of CAV behaviour

Development of Socio-Technical Pipeline for Explainability

- Developing shared understanding between humans and machines
- Explanation of the combined human machine decision process

Formal Specifications for Explainable Validation

- Capturing the properties you expect from the outcome before even touching the data
- Ensuring compliance with mandatory and optional requirements
- Building in cost functions for consequences

Additional Notes

Formal Specifications for Explainable Validation

- Explainable counterexamples of a system requirement failing to hold
- Ensuring the validity of learned models of system behaviour from execution traces

Blue Research Challenges

Data Science the Big Picture

Stefano Padilla, Tristan Henderson, Alistair Lawson

Ethical Impact Assessment for Data Science?

- How can we access the ethical impact of a project?
- How can we embed the culture of ethics into every aspect of data science?

Auditing Data Science

- How can we make data science projects accountable to different stakeholders? Buyers, end users, regulators, developers will all have different requirements.
- Challenges in measurement, provenance, explanation, trust, privacy…

Multidisciplinary Data Science

- Treating data science as a whole needs input from computer science, STS, Management, Law, Politics and more.
- How can we connect these disciplines and focus them? Can we learn from software engineers?
- Data science != computer science

No additional notes were provided for this group.

Green Research Challenges

Data as a Service

Ian Simpson, Matthew Higgs, Alasdair Gray

Inter-operability

- Establish cross-sector standards
- DCAT, schema.org, "core vocabularies", w3c specification
- Creating meta-models to describe data
- Federatable registries for reconciliation/validation/enforcement
- Discovery

Privacy by Design

- Enable "fictionless" data access controlled by default based on data description/access control
- Transparency and metrics/analytics for data use/provenance
- Trust system? Transactional control. Blockchain? Smart contracts
- Control of anonymization/"levels" of obfuscation of data
- Legal

Facilitating Adoption

- Smaller numbers of common tools for "data processing/analysis"
- Incentivise both good practice and contribution – reward
- Discovery improvement
- Training, awareness

No additional notes were provided for this group.

Orange Research Challenges

Visual and Aural Interfaces

Ruth Falconer

Developing Novel and Intuitive Interactions with Data in an AR/VR/XR Context

- Ways to interrogate and query data in AR/VR/XR
- Ways to present, visualise, and navigate data in XR/AR/VR ➔ 3D? 4D?

Communicating Data Quality Fed into Modelling and Resulting Measure of Uncertainty

- Techniques for conveying data quality in aural and visual interfaces for decision making

Measure of User Experience in Visual and Aural Interfaces

- What are the benefits of VR/AR/XR?
- How does it compare with traditional media?
- In what context does it work best?
- Information overload?

No additional notes were provided for this group.

Purple Research Challenges

From Data Processing to Model Exploration

Hans-Wolfgang Loidl, Nikos Ntarmos, Evgenij Belikov, Oana Andrei, Jessica Chen-Burger

Joint Static & Dynamic/Streaming Data Processing

- Hybrid systems for processing static and dynamic/streaming data.
- Tackle data processing challenges (and communication) challenges ➔ parallel processing.

Infrastructure for Clouds/Swarms of Data Centres

- Scaling communication, processing, knowledge extraction, on a global scale

Formal Models for Big Data

- Formal languages and interfaces for model exploration, capturing temporal aspects of data
- Formal languages as a Lingua Franca, and define resource interfaces

Additional Notes

Joint Static & Dynamic/Streaming Data Processing

Additional challenges are to mine static and dynamic data and to make sense of them to assist decision making in either static or real-time applications.

Infrastructure for Clouds/Swarms of Data Centres

The execution environment needs to provide generic high-level abstractions for end-users, whilst transparently adapting the execution to the underlying hierarchical/heterogeneous hardware (e.g. including Accelerators).

Such large-scale infrastructure will also need to provide fault tolerance to recover from intermittent network failures, bit-errors in messages etc, as well as making dynamic trade-offs wrt CAP theorem.

Formal Models for Big Data

This challenge includes the use of type systems and static analysis in programming languages that to provide composable high-level domain-specific abstractions and allow prediction of resource consumption and scaling properties of different pipelines built from such abstractions.

Another established technique applicable to analysing large systems is assume-guarantee reasoning: it is based on a "divide-and-conquer" approach to verify a large system while relying on guarantees provided by the system components when assumptions about each component's environment are satisfied. The assume-guarantee reasoning can be automated by learning component interfaces or assumptions.

Learning temporal models from stream data using statistical or machine learning algorithms allows the use of formal modelling and analysis techniques (e.g., Markovian models, Dynamic Bayesian Networks, temporal logics) to provide richer understanding about the generating processes.

Afternoon Session: Possible Impact Ideas

In the afternoon, we asked attendees to recreate the 'breakout' groups that they created in the morning. Several attendees decided they'd rather join a different colour for the afternoon session, and as such some groups were merged. Attendees were again given 45 minutes, but this time they were asked to reflect on the possible impact that the coloured group might have. As before, afterwards, one member of the group acted as a raconteur and presented these challenges in a 60 second quick-fire session.

In this section, we present the possible impact ideas which were presented by each group, along with any additional comments they provided in the week after the meeting.

Red Impact

XAI (eXplainable Artificial Intelligence)

Robin Hill, Mike Chantler, Özgür Akgün, Michael Lones, Usman Sanusi, Oana Andrei

Model Simplifications For Explainability

- Medicine – e.g. diagnosis, generating disease understanding, informing best practice
- Finance – profiling, Explanations for GDPR, Basis of insurance/loan decisions
- Fraud detection

Development of Socio-Technical Pipeline for Explainability

- Legal processes – e.g. aids to legal argument research
- Resource Scheduling
  - Travel, resource, people, logistics
  - Re-planning, re-scheduling, adaptive planning
  - Plan failure
- Policy
  - Economics
  - Social
  - Energy
  - Environmental

Formal Specifications for Explainable Validation

- Autonomous systems
  - CAVs,
  - Robots in hazardous environment,
  - Assisted cars,
  - Automated trading systems
- Assurance + self-certification

Additional Notes

Formal Specifications for Explainable Validation

Other examples of autonomous systems: UAVs, nursing care robots

Blue & Green Impact

Data Ethics Kitemark
Stefano Padilla, Jeff Pan, Andrea Rosales, Laila Alabidi, Alasdair Gray, Tristan Henderson, Alistair Lawson
Willingness to Share Data
- Improving Data quality - Improving update on services - Reducing Bias - Transparency on whole - Comprehension and Comprehensibility ➔ Ethics Data
Avoid Misuse
- More trustworthy Data Science services - Faculty Policy making
Fair Reuse
- More start-ups : Innovative services - Health care : Improvements

No additional notes were provided for this group.

Orange Impact

Visual and Aural Interfaces

Ruth Falconer, Hans-Wolfgang Loidl, Jessica Chen-Burger, Nikos Ntarmos, Evgenij Belikov

Developing Novel and Intuitive Interactions with Data in an AR/VR/XR Context

- Modelling the invisible in sectors such as Medicine (e.g. explorative surgery), Construction (e.g. visualising hidden structures such as pipes or electricity), Marine Science (e.g. visualising sea floor), or Archaeology.
- Enhancing Decision Making with more immersion in the data
- Improving interactions with self-driving cars
- Other impacts in tourism, game industry, therapy.

Suggested Sectors: Medicine, Tourism, Archaeology, Marine Science, Construction and Game Industry

Communicating Data Quality Fed into Modelling and Resulting Measure of Uncertainty

As the orange group was merged with the purple group, the impacts were all merged under challenge a), with agreement from everybody in the group.

Measure of user experience in visual and aural interfaces

As the orange group was merged with the purple group, the impacts were all merged under challenge a), with agreement from everybody in the group.

No additional notes were provided for this group.

Purple Research Challenges

From Data Processing to Model Exploration

Ruth Falconer, Hans-Wolfgang Loidl, Jessica Chen-Burger, Nikos Ntarmos, Evgenij Belikov

Joint Static & Dynamic/Streaming Data Processing

- Crowdsourcing image data for any real time use
- Feedback aspect of this: streams affecting your static model
- Use Cases:
  - Disaster management first response
  - Games real-time large scale
  - Recommender systems
  - Backtracking data provenance
  - Geographical and weather data combined for disaster response, city planning or tourism

Suggested Sectors: Health and Social Services, FINTECH, GDPR, Tourism, PSD2

Infrastructure for Clouds/Swarms of Data Centres

- Regional data centres cloud computing can be a) Close to where it is processed, or b) Located in the desired jurisdiction
- Ability to query access several data repositories
- Elastic data storage across data centres
- Performance benefit of structuring data into "core" and "halo"
- Consolidation of resources across organisations, Data redistribution, e.g. Tourism

Formal Models for Big Data

- A unified standardised way of querying data:
  - e.g. time series data: apply temporal logic
  - e.g. Would be nice to be able to know pre-query how long will it take and how much data will get back
- Expressive and rich modelling languages for event data
- Provable computation
- Quality assurance of data

Additional Notes

Joint Static & Dynamic/Streaming Data Processing

Addressing this challenge can have impact on any use case where we have a high volume/high

velocity stream of (dynamic) data that needs to be processed in tandem with a large amount of in-rest (static) data. Examples can be found across most sectors and application domains; for example:

- Insights gained through processing of static data can inform the selection of appropriate (AI, processing, forwarding, etc.) models for related streaming (dynamic) data and vice versa, leading to systems that converge faster and produce higher quality/accuracy clusterings/classifications/predictions/etc.
- Joint processing of (static) geographical and historical weather data, alongside fast paced (dynamic) weather readings or other (social media, smart city, etc.) streams, can be used for smart city management (smart routing), disaster management, tourism (route recommendations), etc.
- Joint processing of (static) historical customer data, alongside (dynamic) streams of current transactions, can be used to improve several fintech solutions (ranging from fraud detection to automated stock trades); this use case will become increasingly more relevant when the GDPR/PSD2 directives will come into full force.
- Joint processing of (static) historical player data, alongside dynamic (near-real-time) user input streams, can be used to improve the quality and responsiveness of AI in online gaming.

Infrastructure for Clouds/Swarms of Data Centres

Addressing this challenge is of paramount importance to address the current bottlenecks in designing, resourcing, maintaining and operating large scale data storage/processing infrastructures. Most large corporations (including search giants and content distributors) are abandoning the huge, monolithic, "centralised" datacentres of yore in favour of a swarm of smaller, more cost/energy-efficient so called "regional" data centres spread around the globe. The rise of the Internet of Things also points in the direction of a model more akin to what is now called the "fog computing", where devices connect to local aggregators, thus forming small virtual "regional datacentres", which are then interconnected through the wide area network. At the same time, several organisations (in the UK and abroad) operate a number of geo-distributed datacentres, while there are also several state/RC-funded datacentres around the UK serving different data needs and users. Being able to access and query data across all of these locations, without the need to transfer the data to a central location, will solve a number of showstopper issues, related to resource allocation and consolidation, data provenance, privacy, licensing, etc., and will allow for a more efficient and effective use of the currently available data and resources. The use cases listed in the original pro-forma give concrete examples of the above.

Formal Models for Big Data

Research and development efforts in the field of big data storage and processing have provided us with a slew of frameworks, systems, techniques, etc., to address our big data needs. Yet the vast majority of these tools have evolved in an ad-hoc manner, growing organically from the characteristics of the specific use cases they were built to address. What is really lacking is a formal, theoretical approach to provide us with accurate, provable models of algorithmic correctness, operational costs, and computations trusted to finish and produce the expected/desired results. Furthermore, formal models could provide a "lingua franca" that would allow researchers and practitioners from across research fields and sectors to reason in and communicate; the need for the latter is painfully apparent whenever people with different scientific/industrial backgrounds are called to design a big data solution from scratch, often needing several iterations just to agree on a common terminology, let alone to describe the required functionality and associated solutions. Furthermore, creating ‘small’ models from Big Data that can be transferred instead of the data and allow truthful reconstruction at the destination would make efficient use of available bandwidth and reduce communication cost.

Wrap Up

We hope that this document will engender discussion and encourage collaboration within this exciting, and rapidly expanding research area. To facilitate this, we intend to distribute this document as widely as possible to interested parties such as EPSRC, ESRC, and the Scottish Government, and we invite you, the reader, to do similar.

To finish, we would like to thank all of the people who submitted ideas, performed sorts, or attended the meeting in person. Without their support, this document would not have been possible.