LEEDS Institute for Data Analytics

The Alan Turing Institute

LIDA Turing Data Study Group

Challenge Call Guide 2021

12-23 July 2021

Contents

Introduction

What are Data Study Groups?

Data Study Groups (DSGs) are 'collaborative hackathons' bringing together organisations from academia, industry, government and the third sector with multi-disciplinary researchers. Organisations act as '**Challenge Owners**' who provide real-world problems to be tackled by small groups of highly talented, carefully selected researchers.

Each DSG features 4 to 6 challenges to be worked on by researchers (**Participants**). Participants select which challenge to work on and work in groups of 8 to 15 during the event. Groups brainstorm avenues of exploration, run experiments on the data and document their activities. They present their work on the final day of the DSG and produce a report that will be returned to the Challenge Owner and published on the Turing website. Participants are early career researchers (ECRs) from various disciplines in academia or industry (the majority are PhD level). By bringing together diverse expertise the DSG offers an opportunity for personal development and the chance to spark the all-important moments of serendipitous collaboration.

For each challenge, a **Principal Investigator (PI)**, usually a more senior individual whose area of expertise aligns with the challenge, is recruited. PIs support Challenge Owners throughout the process with the preparation of the challenge framework, data and final report.

The final report is a key DSG outcome, as it provides a tangible output for both Challenge Owners and Participants. Participants produce reports detailing their work and main findings during the event. PIs then review this document before it is published online. Some examples of past DSG reports can be found here.

What can Data Study Groups achieve?

Above all, the purpose of the DSG is to offer **skills development** to the Participants who will experience a powerful mix of curated data, a real-world research question, domain expertise and academic leadership, brought together within the Turing, the national institute for data science and AI. This attracts talented individuals from around the globe, and the number and quality of applicants is typically very high.

Proof of concept and exploration. The DSG is well-suited for concept testing and rapid exploration. A multi-disciplinary team of ECRs with quick, agile thinking are well-positioned for generating many potential avenues for investigation.

Kickstarting larger projects. Many Challenge Owners who have participated in the past continue their engagement with the Turing and Participants, through collaboration on research articles or conference abstracts, follow-up investigation of results and recommendations or the recruitment of Participants.

Knowledge transfer. Challenge Owners are encouraged to be as involved as possible throughout the process. Challenge Owners work with PIs during the challenge curation and are embedded in the groups during the event. All insights, reasonings and methods are shared with Challenge Owners. DSG is a great learning opportunity for those directly involved (i.e. Participants, PIs and Challenge Owners) and acts as a catalyst to stimulate wider learning at the Challenge Owner organisation.

Research challenges

We are open to considering challenges in any area. Staff in Leeds Institute of Data Analytics have specific expertise in areas related to health and medicine, including biomolecular science, urban analytics, population and consumer data. We also have a good track record of working with large and small companies on problems of commercial importance. We are particularly interested in challenges or direct relevance to the local area and economy, in the Leeds region and across the North of England.

Funding

Participation is free of charge, but proposing organisations need to commit staff time and resources to challenge preparation and participation in the DSG.

Who should apply as Challenge Owner?

We welcome proposals from a wide variety of organisations from business, charities, research institutes, the NHS, public health bodies or industry The Challenge Owner should have support from senior stakeholders within their organisation.

During the challenge curation, the Challenge Owner will be required to nominate a member of their staff who has a deep understanding of the proposed problem and the associated data. Ideally, the nominee would have already considered a few ways to address the problem. It would be beneficial, albeit not essential, if the nominee has coding and/or data science experience. Nominees with basic coding and/or data science experience will benefit most from the DSG.

During the event we reserve up to three Participant slots per group for the Challenge Owner. They should be the same individuals throughout the event. It is vital to note that the Challenge Owner Participant is not there to direct or lead the group, but to take part as a Participant and a domain expert.

Ways of working

We expect Challenge Owners to be involved in every step of the DSG process. They should be open and transparent about their reason for participation, data and time commitments. The DSG is a collaborative endeavour rather than a consultancy-type engagement. The challenge organisation sets the challenge but does not direct the conduct of the research. However DSGs can often lead to more focused project activity, and we welcome the interest of organisations seeking broader long-term engagement. Challenge Owners must agree to have the final report (detailing the challenge, the process and the methodology) published at the end of the DSG. A challenge needs to be carefully curated so that it is:

- A framework for exploration and identifies benefits and beneficiaries beyond the Challenge Owner;
- A question with data-driven solution(s);
- Achievable in 3.5 days;
- Challenging, non-trivial and interesting for Participants;
- Clearly scoped but is not a discipline in its own right e.g. protein folding, stock prediction;
- With potential for follow-up work.

An academic counterpart

The PI will be an experienced data scientist with expertise relating to the challenge. S/he will work with the Challenge Owner on the research question to be investigated during the DSG, and help make the challenge suitable for the DSG. The PI is also integral in overseeing the suitability and quality of the data the Challenge Owner provides (e.g. data sensitivity tiers and how that affects the Participant experience). Challenge Owners are expected to be able to meet with the PI as often as needed during the challenge curation stage: depending on the complexity of the challenge, we estimate that 4-5 meetings, as well as some time between meetings to prepare key documentation and data for the challenge, will be required. This curation process yields documentation that contains:

- A title for the challenge
- A short challenge description: A one-paragraph description of the challenge, to be published with the call for Participants.
- A long challenge description: A two-page description of the challenge, to be included in the information pack for Participants and for the Turing's internal ethics approval process.
- A "sales-pitch" presentation: Each Challenge Owner will give a 15-minute presentation (including 5 minutes for questions) of their challenge on the first day of the event. If a virtual DSG is held due to COVID-19, this would be a pre-recorded video, with separate Q&A sessions led by the Challenge Owner.

Where appropriate the PI will complete a University of Leeds ethics check. The Challenge Owner is expected to support the PI in completing all necessary forms in advance of ethical approval from the appropriate University Committee being sought.

Data

The PI will be able to advise on the relevance and quality of the data in preparation for the challenge. It is, however, the responsibility of the Challenge Owner to provide the data and ensure it is ready for the DSG. This will include the necessary cleaning of the data; if Challenge Owners are unable to do so, the challenge will not be progressed. Please see Selection criteria for more detail on data requirements.

Prior to data transfer, the Challenge Owner will take part in a project sensitivity exercise, the outcomes of which will inform the most appropriate computing environment for the challenge.

A signed data sharing agreement will be required for each challenge.

The event

Challenge Owners are expected to attend the event. Particularly important are the first-day pitch presentation and the final group presentation at the end of the DSG. Challenge Owners should also be accessible a few hours per day throughout the DSG to answer Participant questions. We, however, strongly encourage Challenge Owners to embed with their challenge group throughout the week, and be actively involved to provide Participants with the necessary domain knowledge and to learn from them. This will upskill the Challenge Owner Participants and expedite the knowledge transfer back to the Challenge Owner organisation.

Groups consist of 8-15 data scientists (PhD level and above). Each group has a facilitator (invited before the event from the applicant pool) to support the organisation of the group; the PI provides academic support and report feedback.

Post event

After the DSG, the PI will collate reports written by the Participants. The final report will be shared with the Challenge Owner to identify any commercial sensitivities. Data will NOT be published, unless the Challenge Owner gives permission.

Follow on opportunities

DSGs offer an excellent opportunity to launch an initial exploration of the proposed problem. We particularly welcome challenges with potential for further development. Experience from previous Turing DSGs has shown that DSG outputs are greatly enhanced if there are dedicated resources and/or personnel who are committed to build on the work within the organisation. We are particularly keen to work with organisations that are data-rich and demonstrate a collaborative approach that enables the development of a meaningful research agenda. This mode of working has helped deliver crucial breakthroughs first conceived at a DSG.

Selection criteria

The Data Study Group is a unique offering which may not suit the needs of every organisation. If an organisation or a project meets the criteria below, it has the potential to make a great DSG challenge.

Quality

The problem should be challenging, yet progress needs to be possible within a week. It must be well specified to allow Participants a good start with tangible outcomes at the end, leading to more exploratory or less well-defined questions that can be further developed in follow-up projects. A good rule of thumb is to already have one or two off-the-shelf approaches in mind that might yield results. Undefined or vague challenges will be rejected due to the time constraint.

Challenges must focus on analytics and AI, not rote tasks such as data munging, curation or scraping. It must be appealing to Participants by being unique, having real-world impact, offering potential in terms of long-term projects, and being the "right" level of data scientific challenge, i.e. not trivial but also not so hard that nothing can be achieved.

Interdisciplinarity and learning potential

Niche projects that are disciplines in their own right do not typically make a good challenge, as they do not tend to lend themselves to this interdisciplinary approach. For example, protein folding is a discipline in its own right, so there would need to be a concrete new angle to consider this for a DSG challenge. Projects should have an interesting and unique angle to make them appealing to Participants' curiosity, but also have potentially generalisable outputs so that similar data science challenges within the sector might benefit from the same research.

Potential for impact and follow on

Projects should seek to generate positive impacts for the Challenge Owner directly and for the wider data science community by investigating where findings can be applied/translated to other contexts. There should also be a clear path from DSG outputs to a larger, more in-depth project.

Data readiness – appropriate, relevant data

Challenge Owners must provide the data and have the right to use it. It is the responsibility of the Challenge Owner to ensure they have the necessary rights and permissions to use and share the data they are proposing to use.

The data provided should be sufficient to address the challenge. The dataset(s) must be rich enough to provide ample avenues for exploration, but not overwhelmingly/unnecessarily complex. Challenge Owners must be able to share a subset of the data at the time of the challenge application. By the start of the event the full data must be cleaned (if applicable) and ready for analysis.

Data can come from publicly available sources, Challenge Owners are, however, responsible for providing assurance that they have the right to use the data as part of the challenge.

Data can be sensitive, but the Challenge Owner should strive to keep any sensitivities to the bare minimum. If the data contains personal/sensitive information, it must be GDPR-compliant, and the Challenge Owner should be prepared to anonymise. It should also be noted that the more sensitive the data, the more restrictive the computing environment and the analysis will be, which can substantially slow research progress during the event.

A motivated, committed and capable Challenge Owner

No project will succeed without a fully invested partner. We need committed partners to provide insight and help drive the challenge. It should be evident that Challenge Owners are committed to the challenge topic and have demonstrated previous consideration and light-touch investigation in the area.

We are looking for organisations who will be able to invest time in the DSG process. Challenge Owners must be heavily involved in preparing the challenge (alongside a PI), presenting the challenge to Participants, as well as embedding in the group during the event. The more hands-on a Challenge Owner can be, the more they will take home from the DSG. This level of engagement usually requires an individual within the Challenge Owner organisation to dedicate a few hours a week in the months running up to the event and a near full-time commitment during the event.

The estimated time depends on the complexity of the challenge and data, but it should be expected that after the initial few conversations between the Challenge Owner and the PI, discussion will ramp down as the Challenge framework is solidified at the beginning, and slightly ramp up during the Participant recruitment call when the Challenge descriptions and presentations need to be finalised.

We recommend having one contact with primary responsibility.

Application process and deadlines

Applications are now open. Please submit your Expression of Interest via the online form on the LIDA website here. Alternatively please use the hard copy at Annex 2 below and email it to Ros McDonnell (r.a.mcdonnell@leeds.ac.uk). If your application is shortlisted we will contact you to discuss your ideas with a group of academics. At this stage organisations should be prepared to discuss the proposed challenge and associated data in depth, as well as the objectives and how the outcome may contribute to research in a wider context. This is also the chance to discuss the challenge framework and definition with the experts, so we would advise that at least one member of the organisation with deep technical knowledge attends this discussion.

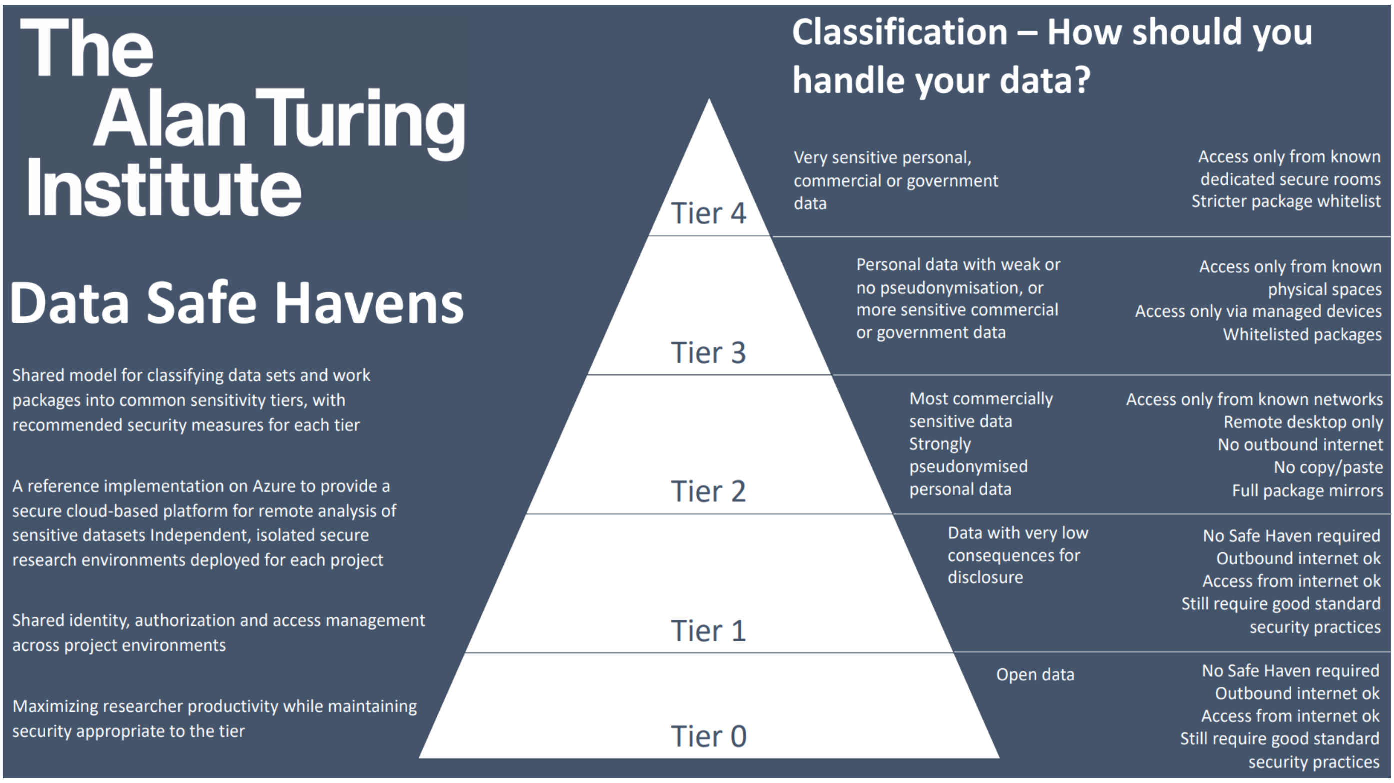
Contacts

DSG Science Lead, David Westhead d.r.westhead@leeds.ac.uk
Turing University Liaison Manager, Rosaleen McDonnell r.a.mcdonnell@leeds.ac.uk

Covid Mitigation

Due to the pandemic and ongoing uncertainty the DSG will be held online over the two weeks 12 to 23 July 2021.
The University is unable to host challenges with high-sensitivity requirements in an online setting. We are thus unable to accept challenges that involve non-anonymised personal data, are likely to reveal identifiable personal data, or highly commercial sensitive data. For more information on the sensitivity levels, please refer to Annex 1: Project Sensitivity tiers. Please get in contact if you have any questions.

Annex 1: Project Sensitivity tiers



The Alan Turing Institute

Data Safe Havens

Shared model for classifying data sets and work packages into common sensitivity tiers, with recommended security measures for each tier

A reference implementation on Azure to provide a secure cloud-based platform for remote analysis of sensitive datasets Independent, isolated secure research environments deployed for each project

Shared identity, authorization and access management across project environments

Maximizing researcher productivity while maintaining security appropriate to the tier

Classification – How should you handle your data?

Tier 4	Very sensitive personal, commercial or government data	Access only from known dedicated secure rooms Stricter package whitelist
Tier 3	Personal data with weak or no pseudonymisation, or more sensitive commercial or government data	Access only from known physical spaces Access only via managed devices Whitelisted packages
Tier 2	Most commercially sensitive data Strongly pseudonymised personal data	Access only from known networks Remote desktop only No outbound internet No copy/paste Full package mirrors
Tier 1	Data with very low consequences for disclosure	No Safe Haven required Outbound internet ok Access from internet ok Still require good standard security practices
Tier 0	Open data	No Safe Haven required Outbound internet ok Access from internet ok Still require good standard security practices

Note: Due to Covid mitigation, we cannot accept Challenges likely to be tier 3 and above.
‘Safe-haven’ data facilities can be provided for sensitive data sets (level of sensitivity subject to confirmation).

Annex 2: DSG Challenge Expression of Interest Form

1. Challenge proposer (organization name, address, website)

2. Contact (the main contact person within the organization for the DSG)

3. Contact details (address, phone, email)

4. Challenge proposal

Describe the challenge and include any key questions you hope the challenge will answer. (500 words)

5. Data set description

Describe the format, size and content of the data set. (200 words)

6. Data set sensitivity

Is your data set commercially sensitive? Does it contain personal data? If the data set contains personal data have the participants consented to data use for this purpose? If the data set contains personal data could it be anonymised for DSG use? (100 words)

7. Data sharing and confidentiality agreements

A data sharing agreement will be required for DSG participation. Please indicate your likely main requirements in this respect.