Summary

The Data Study Group Principal Investigator (DSG PI) is an opportunity for postdocs to engage with industry, government and third sector; define research questions for industrial applications; and work with a multi-disciplinary team to explore novel data science solutions.

It is a great learning experience supported by The Alan Turing Institute with the possibility of getting involved in cutting-edge research with external partners and bringing in follow-on research to your home institution.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to the Data Study Group</td>
<td>1</td>
</tr>
<tr>
<td>What's in it for you?</td>
<td>2</td>
</tr>
<tr>
<td>How much time will this all take?</td>
<td>3</td>
</tr>
<tr>
<td>Process of a DSG challenge</td>
<td>4</td>
</tr>
<tr>
<td>The DSG PI</td>
<td>6</td>
</tr>
<tr>
<td>Responsibilities</td>
<td>6</td>
</tr>
<tr>
<td>What makes a good challenge?</td>
<td>7</td>
</tr>
<tr>
<td>Data</td>
<td>8</td>
</tr>
<tr>
<td>Ethics and project sensitivities</td>
<td>9</td>
</tr>
<tr>
<td>Reproducible research</td>
<td>9</td>
</tr>
<tr>
<td>Findability, accessibility, interoperability and reusability (FAIR) research</td>
<td>10</td>
</tr>
<tr>
<td>Research environments and data security</td>
<td>11</td>
</tr>
<tr>
<td>Tier assessment process</td>
<td>11</td>
</tr>
<tr>
<td>Your first meeting with the CO</td>
<td>12</td>
</tr>
<tr>
<td>DSG PI deliverables</td>
<td>13</td>
</tr>
<tr>
<td>The final report</td>
<td>14</td>
</tr>
<tr>
<td>Code</td>
<td>15</td>
</tr>
<tr>
<td>Roles and responsibilities</td>
<td>16</td>
</tr>
<tr>
<td>Preparing for the DSG</td>
<td>16</td>
</tr>
<tr>
<td>During the DSG</td>
<td>17</td>
</tr>
<tr>
<td>DSG follow-on</td>
<td>18</td>
</tr>
<tr>
<td>Training</td>
<td>19</td>
</tr>
<tr>
<td>The DSG week event schedule</td>
<td>20</td>
</tr>
<tr>
<td>In-person</td>
<td>20</td>
</tr>
<tr>
<td>Online</td>
<td>20</td>
</tr>
<tr>
<td>Participation in the event or how to work with your facilitator and group</td>
<td>21</td>
</tr>
<tr>
<td>DSG PI not attending the full event</td>
<td>21</td>
</tr>
<tr>
<td>DSG PI attending the full event</td>
<td>21</td>
</tr>
<tr>
<td>FAQs</td>
<td>22</td>
</tr>
<tr>
<td>Annex 1 – What makes a bad challenge and how to avoid it</td>
<td>24</td>
</tr>
<tr>
<td>Annex 2 – DSG challenge selection criteria</td>
<td>26</td>
</tr>
<tr>
<td>Annex 3 – DSG project readiness criteria</td>
<td>33</td>
</tr>
<tr>
<td>Annex 4 – The Turing data safe haven tiers and how to assess the tier of the projects</td>
<td>45</td>
</tr>
<tr>
<td>Annex 5 – Code of conduct</td>
<td>45</td>
</tr>
</tbody>
</table>
Introduction to the Data Study Group

Data Study Groups (DSGs) are intensive five-day collaborative hackathons hosted by The Alan Turing Institute. They bring together organisations from industry, government, and the third sector with talented multi-disciplinary researchers. Organisations act as DSG ‘Challenge Owners’ (COs) who put forth real-world problems to be tackled by small groups of gifted and carefully selected researchers.

During the DSG, researchers brainstorm and prototype data science solutions and investigate future directions. They present their work at the end of the DSG and produce a report that will be returned to the CO and subsequently published on the Turing website.

The DSG is first and foremost a training activity, not just for the participants, but for the DSG PIs, and the COs themselves. A close second is the research: creating a multi-disciplinary environment that promotes the cross-pollination of ideas from different data science and AI domains in a safe and serendipitous environment, which in turn we hope will lead to new research avenues that will benefit all involved.

The DSG is quite a large undertaking, not just for the participants, but also for the organisations that want to get involved. From the beginning of the engagement with us to the event itself, we estimate it takes about 3–6 months of preparation. COs must be fully engaged when shaping their challenges and take primary responsibility for providing adequate data.

To support the process, and as part of the Institute’s mission to train the next generation of data scientists, we recruit talented post-doctoral researchers to work with COs, manage the design and scope of the research questions that will feature during the event, and develop their own abilities in overseeing applied research projects.

Although the DSG is a big commitment, we offer constant support to DSG PIs at every stage of the DSG process. The Institute will manage the administration so that you can focus solely on matters that directly relate to the science and research. We offer training to better prepare you for project management, ethics, and engaging with COs. And you will be first author on the grey literature publication that will be produced at the end of the engagement.

What’s in it for you?

As a DSG PI you will have the opportunity to work with industry, government or third sector on real data science problems. Some challenges are also academic in nature, but may not have benefitted from the advances and expertise you would bring in data science and AI.

The DSG PI experience can benefit you in a variety of ways:

- Acquire further project management skills.
- Fine-tune your ability to convert commercial ideas into research projects, and vice versa.
- Fine-tune your ability to generate multiple research projects from loosely defined questions.
- Learn about research areas of interest to you that may be relevant for the challenge at hand, but that you have not yet had the chance to explore.

*Being a DSG PI was a great experience, because it enabled me to partake in one of the most fun activities as a data scientist – serving as a translator between people of different disciplines and backgrounds.

Additionally, this experience has sparked future collaborations with Dstl.

I have two suggestions for future DSG PIs. Firstly, I think that it is very important to meet with the CO one-on-one as early as possible before the DSG in order to help ‘translate’ their questions into the language of data science and help manage expectations.

Secondly, it is helpful to make sure that the data is in a format that would enable DSG participants to jump into data analysis straight away, instead of wasting precious time parsing strange data formats or reformatting tables.

Being a DSG PI is very rewarding, especially seeing all of the innovative ideas that DSG participants come up with!*

Daphne Ezer
Data Study Group PI
University of Warwick

---

It is important to note that participation in the DSG is not consulting. While you will work with the CO to design a challenge that meets their needs, it must be done with the consideration that the final report will be published, and be of benefit to other researchers or organisations. The resulting challenge posed for the DSG is a framework for the participants who attend to explore the dataset in any way they see fit. It is this freedom, coupled with no expectation to return positive results, that attracts them to get involved, and engenders possible novel solutions or directions for further research in the domain.

*During the COVID-19 pandemic DSGs moved online.
This remote edition is extended, to run over three weeks.
How much time will this all take?

Ideally, from the CO signing the DSG terms and conditions to receiving the final report, it will take around 9 months. The DSG PI engagement will usually last about 6-7 months. However, this can vary depending on the complexity of the challenge and data as well as drive from the CO. It can be shorter where the challenge and data are in a good state, and longer if further data preparations and/or collection are required.

The majority of the work will take place in the months preceding the event itself in the challenge scoping and data validation stage.

Process of a DSG challenge
- Challenge scoping and data validation: 3-4 months (up to the DSG week itself), approximately 1-4 hours per week, dependent on the complexity of the challenge and data. This includes sensitivity and ethics checks.
  - 60 hours/8-10 days 100% full-time equivalent.
  - The intensity may vary between weeks – some may have more hours, some fewer.
  - If there is less time for the project to be completed (e.g. 3 months) this will be compressed into those 3 months.
- GO NO GO: if the challenge is not ready then we reserve the right to delay the challenge and move it to the next available event. GO NO GO will take place a few weeks before participant recruitment.
- Participant recruitment: 3-4 weeks, opens 2 months before the commencement of the main event.
- DSG week:
  - In-person – 5 days, usually 9-12 hour days, except the last day.
  - DSG PI should join the first day for presentations and to meet the group, and also be present for the final presentations.
  - Tuesday, Wednesday and Thursday: at a minimum the DSG PI should attend several hours per day to answer academic questions about the challenge and to monitor the reports.
  - Ideally a DSG PI should join the full event for weeks 2 and 3, and participate with the group as an equal member.
- Online – 3 weeks (1 week part-time, 2 weeks full-time business hours).
  - During the first week the DSG PI should join the two CO Q&A sessions (around 30 mins) for their challenge to support – at lunchtime and early evening. The DSG PI is more than welcome to join some of the other participant activities during week 1.
  - DSG PI at minimum should join the first day of week 2 to meet the group and be present for the mid-event (end of week 2) and final presentations.
  - Between Monday week 2 and Friday week 3, at minimum the DSG PI should be present to answer academic questions about the challenge and to monitor the reports several hours per day.
  - Ideally a DSG PI should join the full event for weeks 2 and 3, and participate with the group as an equal member.

**Report finalisation:** the DSG PI will have up to two months to complete. The report will then be peer-reviewed and the DSG PI should be prepared to make corrections, amendments and redactions (one month for peer reviews) – again, the time required is based on the complexity of the challenge and the state of the report at the end of the event, but on average this should take about two full-time work days.

Before the report is published, COs will have the opportunity to review the final report to ensure nothing confidential has remained in the document.

### The DSG PI

The DSG PI leads on the academic design of a challenge, working closely with the CO. The DSG PI is the academic lead; the CO the problem and context giver. The DSG PI will scope the problem, taking it from an industrial/commercial problem and turning it into an academic challenge that can be presented to participants for the DSG and tackled in four days (nine days if event is online).

During the event itself, the DSG PI should provide academic input and suggestions to the group about the challenge. They should not direct, but support the group in what it is investigating. The DSG PI will be supported by a facilitator (taken from the group cohort) who will manage the day-to-day group coordination during the DSG event. The DSG PI will also need to review the contents of the report, ensuring that the narrative is coherent and well-organised, relevant to the DSG question, and scientifically rigorous (e.g. with assumptions and shortcomings clearly stated, and achievements not over-stated). The project will conclude with a published report (on the Turing website), co-authored by the DSG group and finalised by the DSG PI.

Great care is given to ensure that the work is not consultancy – the CO is not able to dictate how the academic challenge is shaped. The DSG challenge should present itself as a multi-method, multi-approach framework that allows participants the freedom to explore the CO data in any which way they choose, whilst answering questions that will give some insight to the CO problem.

The challenge needs to be prepared so that it is:
- doable in the time allowed for the event;
- an open, multiple approach challenge;
- publishable, with as little redaction as possible;
- ideally, leads on to further research.

### Responsibilities

A high-level overview of the key responsibilities is as follows, with detail and further guidance in later sections.

- Turn a CO challenge into a DSG challenge, with clearly formulated scientific/academic questions that are realistic for a 5-day event (or 8-9 days if online) while still offering ample learning opportunity for participants.
- Feed back and advise on the data to be used for the challenge (you will not be responsible for acquiring, cleaning or preparing the data): is it feasible to address the scientific questions posed with the data provided?
- Complete an ethics submission on the project.
- Complete a project sensitivity assessment.
- Oversee and proactively offer guidance to ensure the scientific integrity of the challenge during the DSG event itself.
- During the event, feedback on the content of the report that is written by the DSG participants.
- Finalise the report in preparation for publication.
What makes a good challenge?

A good challenge uses the strengths of the DSG scheme in providing participants and the CO with an enjoyable and informative experience, as well as creating ample opportunities for impactful follow-up.

- To ensure this, the DSG PI and CO must work together to shape the challenge into something that is suited to the five-day setting (or 8-9 days for online) of the DSG. Concretely, challenges should:
  - Be realistic to explore within one day of brainstorming and three/seven days of data science work.
  - Be realistic to address with the data provided.
  - Not have undue risks from data sharing, e.g. ethics, technical restrictions, privacy constraints, or data quality.
  - Be well-specified enough to give participants a good start with low-hanging fruit, leading into more exploratory or less well-defined questions that may be more difficult.
  - Focus on analytics and AI, rather than on rote tasks such as data munging, data curation, or data scraping.
  - Be appealing to participants, with real-world impact, the potential to turn into a long-term project, or the right level of intellectual data-scientific challenge.
  - Be likely to lead into impactful medium-term or long-term projects with Turing partners and participants, that can be kick-started by a DSG proof-of-concept or exploration.

The optimal trajectory of a challenge is as follows:

- Participants with the right skills read the long description of the challenge and watch the challenge presentations, then self-assign to the challenge they are interested in and feel they can contribute towards.
- During the week, the challenge team of CO representatives, DSG PI, and participants produces proof-of-concept solution for the low-hanging fruit challenges and brainstorm a series of approaches for the wider context.
- A report (and occasionally code) is produced which will be shared with the CO, as well as published on the Turing website.

In the context of a wider research engagement:

- In the short-term, a follow-on project group forms around the suggested directions from the DSG. Longer-term and larger-scale project planning is informed by this seed research.
- Results of the follow-on project get published in major scientific venues, and/or lead to disruptive innovations which in turn inform further collaborative research projects, embedded in a long-term partnership network.
- All challenges should have an element of social good — in that they benefit not just the CO but also wider society, be that the academic community, the industry sector as a whole or society as a whole. This should be emphasised where possible in the framing of the challenge and the report.
- Annex 1 has some more points on making a good challenge (or mistakes to avoid).
- Refer to the challenge selection criteria in annex 2 for more details on why challenges get accepted into a DSG.
- Refer to the project readiness criteria in annex 3 for an overview of the targets you want to achieve over the preparation stage.
- Another great resource for the design of a DSG project is The Turing Way: A guide for reproducible research. While a challenge might not be directly reproducible due to confidential data, methods developed should be constructed with reproducibility in mind, for similar datasets or problems.

Data

In parallel to defining the challenge, the DSG PI will help evaluate the appropriateness of the data being provided.

For instance:

- How sensitive is the project and what sensitivity tier does the data belong to?
- See next sections on ethics and project sensitivities, and research environments and data security.
- Is the data readily available?
- Is the data suitable for the challenge at hand?
- If not, what additional data is needed (if any)? What additional data could be collected or incorporated to enrich the challenge?

Information on the security tiers and how to assess them can be found in annex 4.

You are not responsible for the cleaning or preparation of the datasets. If it is a quick job you are comfortable with, by all means do help, but this is not your responsibility. Advise the CO what they need to do and what aspects of their datasets are of most value, and/or missing and required to make the challenge successful.
Ethics and project sensitivities

All research undertaken by the Institute is required to undergo an ethics review. The DSG PI will collate the required information and submit it, but will need information from the CO. Ethical considerations when shaping the challenge are:

- Is the data sensitive, e.g. personally sensitive and/or subject to GDPR?
- If there is personal data, what is the process by which informed consent has been given?
- What are the risks of the project posing potential harm to individuals or society?
- What measures of mitigation have been undertaken with respect to these risks?

While the DSG PI submits the (Turing internal) approval request form, much of the information will require input that only the CO can give. Incorrect or inaccurate information about the challenge and data can delay the project and in extreme cases risk the viability of the whole engagement. It is the CO’s responsibility to have the correct authorisation to use the data and to know the intricacies of the data they are providing. Try to be as specific as possible, as generalising can lead to requests for more information.

The CO should also prepare to assess the ‘sensitivity tier’ and security arrangements for the hosting of their data and challenge. The DSG PI will also independently complete the same project sensitivity assessment so that the organisers have complete confidence that we are hosting the challenge in the most appropriate environment.

For further information on the specific sensitivity tiers, what they mean, as well as the flow diagram that derives the tiers please see annex 4.

Reproducible research

Not all DSG challenges will be able to be reproducible. Often data is confidential and cannot be published with the report. However, the DSG PI should consider what aspects of the problem and challenge may be relevant to outside readers.

We recommend designing a challenge in such a way that minimises the need to reference confidential material in the outputs – be that the core datasets or specific patents, or confidential processes or IP. By considering confidentiality from the very beginning, final publication of the outputs will be less burdensome to assess, and will result in a cleaner report with reusable findings without the need for redactions.

Where some confidential information is integral to the final outputs, planning from the beginning how to represent such information will help preserve the integrity of the report narrative and reusability. Where redactions and obfuscations are needed at the end of the entire process, this can lead to patchy reports that are difficult to read due to entire sections needing to be removed, which also renders them unusable by the wider community.

Considerations can be things such as how to anonymise or generalise expected results that preserve the narrative of the findings. Alternatively, the explanation can focus on the methodological approaches and how the results support those methods. A more detailed view on Turing’s open research aims can be found on The Turing Way.

Read more about data sharing and open data.

Findability, accessibility, interoperability and reusability (FAIR) research

The FAIR guiding principles for scientific data management and stewardship were developed as guidelines to improve the findability, accessibility, interoperability and reusability of digital assets; all of which support research reproducibility. The FAIR principles play an important role in making your data available to others for reuse.

Similar to reproducible research, not every DSG challenge will be able to follow the FAIR guidelines. However, one should try to follow as many guidelines as feasibly possible.

To help DSG PIs with planning their challenge and for resources related to ethical, reproducible and FAIR research, it is strongly encouraged that DSG PIs make use of The Turing Way. You can sort by the DSG tag to show those chapters most relevant for DSG. It is an invaluable learning resource for all up-and-coming data science leaders. It is wise to bookmark this page and to check it regularly as it is being constantly updated.

The Turing Way is an open source community-driven guide to reproducible, ethical, inclusive and collaborative data science.

Its goal is to provide all the information that data scientists in academia, industry, government and the third sector need at the start of their projects, to ensure that the outputs are easy to reproduce and reuse.

The book started as a guide for reproducibility, covering version control, testing, and continuous integration. However, technical skills are just one aspect of making data science research accessible for all.
Research environments and data security

Each DSG challenge will have its own Secure Research Environment (SRE), used for the storage and analysis of data. Each environment is separated by an Azure subscription and is its own entity, completely isolated from that of other challenges.

The type of SRE provided will depend on the sensitivity tier assigned to it, which is done as part of the scoping of the project. The SRE can include the following security measures depending on data classification:

- No public internet.
- Two-factor authentication.
- Cut and paste restrictions.
- Access to the environment only available from specific sites.

In the higher tier SREs where there is no internet access, full package mirrors of popular tools can be found within the environment. However, you should consider the types of packages that will be required for your challenge and check with the DSG team to see if those are available as early as possible.

Tier assessment process

The CO, DSG PI and referee must arrive at the same conclusion before a tier can be successfully assigned.

DSG challenge data classification process:

1. The CO assigns their project a classification tier.
2. The data is moved into a SRE at the Turing.
3. The DSG PI then reviews and assigns a tier.
4. If the classification tier is 2 or above, a referee separately reviews and assigns a tier.
5. If all parties agree to the classification tier, it can be officially assigned. If not, a discussion is had to reach a consensus.
6. Ahead of DSG week, the SRE is then set up for each challenge and the necessary accommodations made surrounding permissions and restrictions depending on the tier.

After the DSG, prior to publishing any reports or code from a tier 1 or above project, a final declassification review is conducted to ensure no sensitive information is published. Here the process is reversed, where the CO, DSG PI (and referee) review the report (and other outputs), identify items that are deemed sensitive and then discuss how to safely represent said items for publication – delete, anonymise, generalise or obfuscate. The CO will also have the opportunity to review the final report and other outputs prior to publication, for minor redactions relating to confidentiality. Unless explicitly stated, data is never published.

Your first meeting with the CO

The primary goal of this first meeting is for both yourself and the CO to sketch out an overview of what the challenge will look like for the DSG.

The original challenge proposal from the CO’s initial submission, along with notes from the challenge assessment, will be shared with you prior to this first meeting. However, it would be good to have the CO reiterate to you what their organisation hopes to achieve from participating in the DSG. You should bring some ideas about how you want to frame their proposal as a challenge that is both an exciting research question as well as suitable for the timescales of a DSG (see the chapter ‘What makes a good challenge’, as well as the challenge selection criteria and project readiness criteria (in annexes) to help frame your ideas and guide the conversation).

The CO should bring a detailed data dictionary to the meeting or send one to the DSG PI beforehand. If they still don’t, remind them to send one to you, and if they still don’t speak to the DSG team to help push (the CO should be aware to provide this information to you as early as possible).

Ideally, the CO should appoint a single point of contact for the DSG PI who is able to commit a few hours a week and should have a fairly technical understanding of the core problem and datasets of the challenge. If the lead from their organisation is not a data scientist, they have been advised to bring someone who is able to discuss at length the technical requirements, ideally someone who has already worked with the data themselves. If this does not happen, ask for one, and if this still does not happen, please alert the DSG team to push on your behalf.

Subsequent meetings will be to iterate on and further refine the challenge description and prepare the dataset(s) for the DSG event, as well as prepare the required documentation listed in the deliverables section.

The DSG team will organise your first meeting together, but subsequent meetings should be self-organised between the CO and the DSG PI. The DSG team will check in with both over the preparation phase of the DSG.

Discussions on business functions, such as contracts, data access processes and IT provisions, are not the DSG PI responsibility and should be referred to the DSG team.

Clarification of roles – DSG PI vs the CO

The CO brings the business problem and the domain knowledge. The DSG PI, a data science researcher, will convert this into an academic challenge for the DSG. Expect to compromise. The CO will have final say on the domain focus and direction of the overall challenge. The DSG PI will have final say on the structure of the challenge and the framing.
DSG PI deliverables

The list below gives an overview of DSG PI actions and requirements throughout the DSG process. Deadlines will be provided for the specific event.

Pre-event:
- Agree to the terms and conditions of participating in the DSG.
- Attend the training.
- Work with the chosen CO to produce the challenge title, a short description and long description.
- Review data for appropriateness and relevance.
- Submit an ethics approval form.
- Assess the ‘sensitivity tier’ of the data.
- Support the CO in the preparation of their sales pitch presentation.

During the event:
- Attend the sales pitch presentation.
- Support the CO in any Q&A that follows.
- Attend the end-of-event presentation.
- Ideally try to attend each day of the event, even if not full-time, to support the group with academic questions they may have (the CO ideally will do the same from the domain knowledge aspect).
- Monitor the report as it is being written to ensure the right information is being recorded, ready for you to finalise the report after the DSG.

Post-event:
- Engage in the declassification of outputs where necessary.
- Review and finalise the report.
- Engage with the Turing and CO for potential follow-on research (optional).*

Post-event:
- Engage in the declassification of outputs where necessary.
- Review and finalise the report.
- Engage with the Turing and CO for potential follow-on research (optional).*

The final report

During the DSG week, participants will produce reports explaining their work and main findings. Throughout the week, the DSG PI will review the in-progress report and provide feedback to improve it, and make sure that they have all the information they need to thread the final narrative. Does what is being recorded make sense? Is there all the necessary detail? It should be presumed that the DSG PI will not be able to ask the participants of the group for clarifications once the event is over.

In terms of contribution, participants should write the majority of the report (around 85-90%).

Read a detailed description of what the report should be like, which is also shared with participants.

Following the DSG week, the DSG PI will have around two months to tidy and prepare the report for publication. The DSG PI’s task will be to collate any missing information, complete exposition, finalise the formatting and typesetting, and redact faulty or irreparably incomplete content – all without conducting any further analysis which would be out of scope. Ensure that:
- It is scientifically accurate.
- Claims are backed up.
- Clearly annotated and labelled (with axes marked) and are able to be resized in terms of resolution.
- Spelling and grammar have been checked.

After these iterations, a final peer review is carried out by the DSG team (usually a Turing researcher or academic staff member). The CO will receive the final document and be asked to feedback any required changes that focus on confidentiality only. These changes should be minimal and only absolutely necessary. The DSG PI will need to make any further amendments.

*Follow-on research may not necessarily be conducted via the Turing.

The challenge and report should be answering the core questions posed by the CO, about how the results/methodologies can benefit wider society – be that a particular domain, data science and AI research in general or a clear public benefit that could be developed through further research.

COs will not usually be given access to the report during the event week (unless they are active participants in the data exploration, and contributing to the content of the report on par with the other participants). Depending on the classification tier, we can usually give read-only access after the event. However, COs should refrain from directing the report finalisation (e.g. requesting edits to paint their organisation in good light, or limiting potential negative results).
Code

Code is also given to the CO and is also sometimes published through the Turing. However, code is not revised and improved in the same way that a report is.

The only action on the part of the DSG PI is to declassify code if it contains any sensitivities as identified by the declassification process. Participants should be discouraged from including sensitive information within the code.

The challenge should be structured so as not to require anything that might be sensitive within the code where possible.

Should the CO want clean code, support can be provided by the research software engineers (at the expense of the company and subject to availability).

Roles and responsibilities

Preparing for the DSG

<table>
<thead>
<tr>
<th>Role</th>
<th>PI</th>
<th>CO</th>
<th>DSG team</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writes short description of project</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Advises on data suitability / readiness</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prepares data</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Writes long description of project</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Completes data sensitivity assessment for data ingress</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Completes ethics approval application</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Provides information for ethics approval</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Advises on framework of challenge</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transfers data to the Turing</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Provides guidance and documentation on challenge curation process</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Records / presents challenge to participants</td>
<td></td>
<td></td>
<td>(We can help record)</td>
</tr>
<tr>
<td>Reviews all content before publication</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Organises participant recruitment</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organises DSG event</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Attends PI training sessions</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
During the DSG

<table>
<thead>
<tr>
<th>Role</th>
<th>Participant</th>
<th>Facilitator</th>
<th>PI</th>
<th>CO</th>
<th>DSG team</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chooses which project they want to work on</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joins challenge Q&amp;A sessions</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Organises and delivers training sessions</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Works on challenge during the event</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Presents findings at presentation sessions</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contributes to authorship of the report during the event</td>
<td>X</td>
<td>X</td>
<td>(X)*</td>
<td>(X)*</td>
<td></td>
</tr>
<tr>
<td>Advises on prioritisation of needs</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advises on content and quality of report</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Answers questions on challenge / data / domain specificities</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Answers questions on the event / event management</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administrative support</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal support</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group organiser</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decides route / way of exploring challenge</td>
<td>X</td>
<td>(X)*</td>
<td>(X)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acts as communication channel between groups and DSG team</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logs attendance</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensures all work is documented and saved</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*In the case that they are actively participating in the exploration of the data and coding prototypes during the event week.

† In the case of the group dithering on identifying potential directions for exploration.

DSG follow-on

<table>
<thead>
<tr>
<th>Role</th>
<th>PI</th>
<th>CO</th>
<th>DSG team</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completes data and outputs sensitivity assessment for data egress</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finalises report</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organises report review</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amends report based on reviewers' comments</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reviews and approves final report</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obtains participant consent for report publication</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publishes DSG final report</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organises follow-on work / projects</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Training

DSG PIs will be invited to attend a number of training sessions to help prepare for the role.

Ethics

To not only help with completion of the Turing ethical approval process, but to hopefully instil a broader perception of how ethical considerations are integral to data science research. Some of the questions you may have answers to at the end of this session are:

- Could any harms arise from the research – both as subjects or to the people involved in conducting this research (e.g. viewing violent content could harm researchers)?
- Could conducting or promoting this research create unintended negative outcomes, such as environmental damage, new power imbalances, or the misuse of technology?
- How do you plan to ascertain and acknowledge the limitations of your research, if any (e.g. does the data sample you use allow for generalisability of your research findings)?

What benefits could your research contribute that would balance or outweigh any potentially negative impacts that could arise?

Project design and management

An introduction to project design and project management approaches within the context of DSG.
- Team structure
- Control
- Timelines
- User stories
- Risk

Communication with non-academics (TBD)

COs are not always from academia and usually have very different motivations for the results than the researchers. This session will look at effective ways to communicate so that both parties are able to achieve the results they want – more details to follow.

The DSG week event schedule

In-person

Under usual circumstances, the DSG is a week-long event and is divided into three sections:
- The Monday morning sales pitches.
- The work.
- The Friday group presentations.

On Monday morning, participants will hear from all COs – up to six per event – about their challenges and why the participants should get involved. Following lunch, participants choose their groups and the work begins.

Online

The Turing has delivered several online DSGs during the COVID-19 pandemic with great success.

For remote events, the challenge preparation process and outputs remain the same. The event format is modified in the below ways to mitigate for remote collaboration.

The event is extended to run over three weeks and split into two stages:

Week 1 ‘Precursor stage’: The first week of the event is spent preparing participants for remote collaboration, including webinars, social activities and Q&A sessions with the COs. DSG PIs are invited to join these sessions but it is not a requirement.

Weeks 2 – 3 ‘Event stage’: Weeks 2 – 3 are spent working on the challenge. There will be an informal group-wide check-in at the end of week 2 where groups can tell the other groups what progress they have made and what barriers they have encountered (and other participants are encouraged to posit potential ways to overcome those barriers). The main presentations will take place on the Thursday of week 3. Friday of week 3 is for final report completion.

Usually on Monday evening, we will take all participants (and DSG PIs and COs if they want to join) to an offsite social event.

Tuesday to Thursday, the participants work hard in running experiments, brainstorming avenues of research and writing up what they have done and how well it worked, for the report and final presentation.

Friday, the individual groups present back to the whole DSG, including all the COs. This will be a brief overview of what happened and what was (or wasn’t) discovered.
Participation in the event, or how to work with your facilitator and group

All participants, including the CO (and their guests), the DSG PIs and organisers, must abide by the DSG code of conduct (see annex 5). The code of conduct is in place for the entirety of the engagement in DSG.

DSG PIs are encouraged to embed within the teams as much as possible. As the challenge experts, DSG PIs are required to join their teams for the first and last days of group work to help guide participants. At minimum, DSG PIs should be contactable and responsive a couple of times a day so that participants can run their ideas and questions by them.

Each group has a facilitator (or pair of facilitators) who are chosen from the participant pool before the main event. The PI should meet with the facilitators before the main event starts to discuss approaches and possible issues (DSG team will set up).

While groups should be self-organising, the facilitator will act as the lubricant to help get groups moving as quickly as possible. In some cases, they will need to lead the group; in other cases, where a natural leader emerges within the group, they will support them. Facilitators are the eyes and ears of the DSG team during the week and have daily catch-ups with the DSG team to ensure the smooth running of the event and to create the best possible experience for everyone.

It is anticipated that there will be members of the CO attending the full event in some capacity.

**DSG PI not attending the full event**

Be prepared to answer queries throughout the event, even if you are not going to be attending the day-to-day process. Block several hours daily in your calendar for the duration of the main event to be able to answer emails or messages, and to review the report.

Support the facilitators with technical issues relating to the challenge, the study area or the different directions from which participants are approaching the problem.

---

**DSG PI attending the full event**

Same as not attending the full event, but presents several additional opportunities.

Get involved in the exploration of some of the approaches, or support some of the more novel approaches proposed. The participants may have a very wide range of skills, so take advantage of approaches and skill sets that may not be directly in line with your own. Educate, assure and support the weaker members of the team. Network. If there is a lack of emerging leadership within the group then lead from behind.

The challenge posed for the DSG is a framework for the participants who attended to explore the dataset in any way they see fit. It is this freedom, coupled with no expectation to return positive results, that attracts them to get involved, and engenders possible novel solutions or directions for further research in the domain.

Do not send the participants in specific directions of research. Promote diversity of ideas and explorations. There have been situations where the more unusual ideas have borne fruit and even been presented at conferences to the wider academic community.

The DSG is an opportunity for all participants – including the COs, the facilitators and the DSG PI – to learn from one another. It takes everyone’s involvement in the preparation of the challenges, and in supporting the events, to help the events reach their full potential. Creating this positive environment will benefit the science, the results and everyone’s experiences of the DSG.

### FAQs

**When is the DSG?**

DSG is held around three times a year at the Turing offices in London (or online).

**How many challenges feature in an event?**

A DSG event usually hosts around 5 – 6 challenges.

**How is the DSG week structured?**

**In-person**

Day 1: Challenges are presented by COs in the morning, participants self-select which challenge they want to participate in after lunch, then begin to brainstorm.

Days 2 – 4: Brainstorming, modelling and problem solving.

Day 5: Progress and recommended routes forward are presented.

**Online**

Week 1: Part-time commitment of workshops and Q&A sessions with COs to prepare participants for remote collaboration and to inform challenge selection.

Weeks 2 – 3: Groups form and commence brainstorming, modelling and problem solving, presenting findings at the end of the event.

**Is the DSG a cheap consultant?**

No. Whilst the CO dictates the scope of the challenge, it is up to the DSG PI to shape the challenge into an interesting research question. The resultant question should provide a framework for the participants to explore during the week. Participants are offered the freedom to investigate the data as they wish and use the challenge description as a guide. What is more, the report findings and code will be published on the website, available for all to share and learn from.

**What is intellectual property?**

Any intellectual property (IP) arising from the DSG will be owned by the Turing. All of the CO’s background IP and that of our researchers remains with the inventor. For 3.5 days of work we do not expect any meaningful IP. Remember, all results will be published, and any code developed and published will be made available under permissive open-source licence.

**Do I need to attend the whole DSG week?**

It is preferable, but not a requirement. If you are unable to attend the whole week, we do ask you try to attend for the first and last days (group forming and presentations). Each group has a facilitator with whom you will be connected before the DSG event. You should liaise closely with the facilitator if you are not going to attend the whole event.

**What does the facilitator do?**

The facilitator is chosen from the pool of applicants. Their job is to manage the group, specifically people dynamics.

They will not be heavily involved in the data analysis but will have had a preview of the dataset before the event week to help support the other participants with familiarising themselves with the data.

They are also NOT responsible for writing the report. They will guide and encourage colleagues to complete report sections and ensure that the report is sufficiently complete by the end of the week, in preparation for finalisation.

**How much time am I expected to put in?**

See ‘Timeline of a DSG challenge’ on page 4.
What should be included in the report?

Visit our reports page for example reports to see what should be included.

Can the report be used for my university Research Excellence Framework?

The report will be a DOI-identified artefact with the PI as an author. It can be treated like any other co-authored open access paper and mention the DSG role / effect of DSG report in an impact case study (as long as it is merely mentioned as part of the wider research and activities you are undertaking at your university).

Are there opportunities for follow-on collaborations?

As well as training the next generation of data scientists, the DSG serves to kick-start research collaborations between the Turing and its partner universities, and industry, government and third sector.

It is our aim to continue as many DSG projects as possible in some form of follow-on. This could be, for example, writing a paper to extend the research; working with the CO to investigate the DSG results more closely; or brainstorming, e.g. conversion into follow-on work.

The participants have no idea what to do or where to start.

Who is this opportunity for?

Ideally the PI for a DSG project should be someone who is not yet a PI but wants to get some experience in leading a project, specifically working with a third party who will provide the overall scope of the investigation. They should be open to experimenting with a multitude of approaches. The DSG PI does not need to be affiliated with the Turing or the Turing's partner universities.

How do I get involved?

Expressions of interest should be submitted here.

Members of the list will be sent upcoming DSG PI opportunities and are invited to apply. The Turing will also post the challenge calls in the weekly bulletin.

When you apply, you will need to send your CV as well as a cover letter.

The cover letter (up to 2 pages) should demonstrate your ability to suggest multiple potential methodological approaches to the challenge being applied for, as well as demonstrate:

- Experience in applied data science.
- Willingness for multi-disciplinary collaborative work.
- Enthusiasm for working with industry, government and third sector to take their business problems into data science research projects.

Annex 1 – What makes a bad challenge and how to avoid it

Common failure points that create a ‘bad’ challenge arise from the negation of ‘good practice’. These pitfalls should be avoided:

- The challenge scope is mistakenly set up as a one-year research project rather than a three-day exploration.
- The data provided is unrelated to, or insufficient to answer, the challenge questions – the CO expects ‘magical’.
- Data gets stuck due to last-minute problems, or is shared in a bad shape.
- Questions are phrased too vaguely, or are without exception extremely ambitious.

A good rule of thumb is to have one or two simple off-the-shelf approaches in mind that can solve the ‘low-hanging fruit’ question in an afternoon (assuming perfect data quality). For precision, consider whether the questions are phrased in a way that easily allows the identification of a technical solution (as opposed to non-solutions).

For data quality, the DSG PI should iterate with the CO, and perform cursory exploration of common data quality issues, but not clean the data themselves.

- If the data is insufficient, it might be possible to weaken the questions, select an area which the data can support, select reasonable proxies, or recommend collecting different data and returning to the DSG at a later stage.
- This can be mitigated by ensuring the data is shared at least two months in advance of the DSG, and before publication of the challenge for participant recruitment. Note that (assuming due diligence on the side of the DSG team), having the data ready and in good quality is a 100% guarantee for it being ready for the DSG.
- Again, this is subject to scientific judgment in the scoping phase, by the DSG PI.
- A good rule of thumb is to have one or two simple off-the-shelf approaches in mind that can solve the ‘low-hanging fruit’ question in an afternoon (assuming perfect data quality). For precision, consider whether the questions are phrased in a way that easily allows the identification of a technical solution (as opposed to non-solutions).
- For data quality, the DSG PI should iterate with the CO, and perform cursory exploration of common data quality issues, but not clean the data themselves.
- A minimum requirement of description is a full data dictionary, with descriptions of all variables, and sampling conditions in every table, as well as a precise description of how tables relate, if there are multiple.
Here it helps to measure likely outcomes against common qualifiers of impact (e.g. knowledge transfer success, business disruption, academic excellence), as well as considering the participants’ common motivators for getting involved (e.g. decent technical challenge level, impactful applications, research potential, follow-on opportunities).

- While planning ahead is in general a useful idea, thinking concretely beforehand about immediate follow-on tends to greatly amplify positive outcomes from the DSG.

- It is recommended that the CO and DSG PI consider, based on educated guesses, how open avenues and momentum can be converted into productive follow-on work. Quite often, the DSG week leads to interesting work directions that give rise to proof of concept, but couldn’t be fully explored, or only in an ad hoc manner of provisional quality.

The Turing can provide DSG PIs with administrative support to organise multiple follow-on activities, including follow-on workshops, spring/summer projects, project organisation etc., subject to an organisational partnership with the CO.

In addition, DSG PIs may want to keep in mind the following ‘special’ considerations which may or may not apply to a given challenge:

- A CO may not be ready, organisationally, to specify a DSG challenge yet. This happens often if an organisation has just started to adopt data science infrastructure. In that case, extensive scoping and co-development of infrastructure may be more pertinent than running a DSG challenge, which can be done at a later stage when organisational maturity has increased. Alert the DSG team if you find yourself in this situation.

- A related, smaller version of this problem is a lack of data scientific competence when data storage infrastructure is present. In this case, organisational infrastructure, or at least cleanliness of the intended data batch, may have to be progressed together with the CO, outside the DSG, since participants are not data cleaners.

- Some COs exclusively look for brand association with the Turing, and may be enthusiastic to run the challenge just in order to be able to say “we are doing AI now”. This situation is not always easy to spot, but poses the danger of reputational damage since it is usually the same COs who are completely uninterested in valid outcomes or scientific rigour, as opposed to making grandiose marketing claims. Warning signs are vague claims or goals, statements of the kind “we don’t really mind what you do”, buzzword mashing, primary focus on PR and marketing, etc. From our experience, progressing such challenges should be avoided. Alert the DSG team if you find yourself in this situation.

Annex 2 – DSG challenge selection criteria

The DSG is a unique offering which may not suit the needs of every organisation. In order to assist COs, this document is a broad overview of the many considerations the DSG team factor in when selecting challenges. This is broken down into three categories, the challenge, data and impact.

If a challenge is in the red stage for any of the following sections, your challenge is likely unsuitable for the DSGs. We advise meeting yellow stage requirements at minimum before submitting your proposal, with a clear plan for achieving blue and green criteria.

During scoping, the DSG PI should strive to get the challenge to the point where it is in the blue or green criteria zones.

Challenge

The overall challenge is assessed by looking at project quality and organisation readiness.

Project quality

The problem should be challenging yet progress needs to be possible within the context of a DSG. It must be well-specified enough to give participants a good start with low-hanging fruit, leading into more exploratory or less well-defined questions that may be more difficult. A good rule of thumb is to have one or two simple off-the-shelf approaches in mind that might yield results.

Undefined and vague challenges are not suitable for a DSG, as the time frame of the event is not enough to complete meaningful research in general areas.

The challenge scope is instead set up for compact exploration. However, a long-term view should also be considered as to what kind of follow-on work may come out of the DSG challenge.

The focus should be on analytics and AI, rather than on rote tasks such as data munging, data curation, or data scraping.

The challenge must be appealing to participants, whether that’s due to real-world impact, the potential for a long-term project, uniqueness, or the ‘right’ level of data scientific challenge.

Organisational readiness

No project will succeed without a fully invested project partner. We need COs to provide insight into the problem and to drive the project.

It should be evident that organisations are committed to the challenge topic, demonstrating previous consideration and light-touch investigation in the area.

The DSG team is looking for organisations that will be able to invest time into the process. COs must be heavily involved in preparing the challenge (alongside the DSG PI), presenting the challenge to participants as well as embedding in the group during the event itself. The more hands-on the organisation can be, the more it will take home from the activity. This level of engagement usually requires an individual within the partner organisation to dedicate a couple of hours a week in the months running up to the event, and a near full-time commitment during the event. We recommend having one technical contact with primary responsibility.
<table>
<thead>
<tr>
<th>Challenge assessment matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Challenge difficulty</strong></td>
</tr>
<tr>
<td><strong>Challenge definition</strong></td>
</tr>
<tr>
<td><strong>Challenge scope</strong></td>
</tr>
<tr>
<td><strong>Challenge focus</strong></td>
</tr>
<tr>
<td><strong>Resource commitment</strong></td>
</tr>
<tr>
<td><strong>Demonstrable importance to organisation</strong></td>
</tr>
</tbody>
</table>
Data

Data is assessed by looking at data readiness and data quality.

Data readiness

Organisations must either provide their own datasets or have the right to use ones that they suggest bringing in.

Data should already be collated or there should be a clear path to collating the necessary data. If the data does not yet exist and there is no clear and concrete plan in place to collect and collate the data then the challenge is extremely risky. The challenge should be rejected and the CO will be asked to resubmit their challenge once they have data or a plan to obtain the necessary data.

Organisations must be able to share a subset of the data at the time of initial application, or a detailed description and/or data dictionary. By the time of the event, the full dataset must be ready for analysis. If this is not achievable, the CO is not ready to engage with the DSG offering.

Participants should spend the week on interesting analysis rather than data cleaning.

Data is assessed by looking at data readiness and data quality.

If the data contains personal or sensitive information, this must be GDPR compliant and organisations should be prepared to anonymise or generalise the dataset.

Data appropriateness

Data is unrelated to the challenge question

Data is somewhat related to the challenge question, but requires more work to be fit for purpose

Data is appropriate for the challenge question, but can be notably enhanced with data that the company could provide

Data is highly appropriate and relevant to the challenge question

Data quantity

Data consists of just a few subjects or points, single or few tables

Multiple tables with one central table, only one-to-many relations

Unstructured data / data lakes

Unstructured data / data lakes curated to contain only necessary information so that participants are not overwhelmed by information

Data sensitivity

Data contains highly sensitive information, data subjects have not consented to data being used

Data known to be sensitive, anonymity measures planned

Data has been anonymised, but with low level of confidence

Data has been anonymised to high degree of certainty (if containing sensitive information), data subjects / owners have consented to data being used

Data completeness and reliability

Many missing observations and unreliable data points

Some missing observations and unreliable data points, further investigation needed to assess whether okay for a DSG

Some missing observations and unreliable data points, but likely okay for a DSG

Very few missing observations (if any) and overall reliable data

Data complexity

The data proposed must be related to the challenge and expected to be sufficient to answer the challenge questions. The dataset(s) must be rich enough to provide ample avenues for exploration yet avoid overwhelming or unnecessary complexity.

Data assessment matrix

<table>
<thead>
<tr>
<th>Data readiness</th>
<th>Data appropriateness</th>
<th>Data quantity</th>
<th>Data sensitivity</th>
<th>Data completeness and reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data has not yet been collected and there is no methodology in place for doing so</td>
<td>Data is unrelated to the challenge question</td>
<td>Data consists of just a few subjects or points, single or few tables</td>
<td>Data contains highly sensitive information, data subjects have not consented to data being used</td>
<td>Many missing observations and unreliable data points</td>
</tr>
<tr>
<td>Some data has been collected, with firm methodology in place for obtaining more</td>
<td>Data is somewhat related to the challenge question, but requires more work to be fit for purpose</td>
<td>Multiple tables with one central table, only one-to-many relations</td>
<td>Data known to be sensitive, anonymity measures planned</td>
<td>Some missing observations and unreliable data points, further investigation needed to assess whether okay for a DSG</td>
</tr>
<tr>
<td>The data is collected and in workable condition, with some merging and unifying needed to be event-ready</td>
<td>Data is appropriate for the challenge question, but can be notably enhanced with data that the company could provide</td>
<td>Unstructured data / data lakes</td>
<td>Data has been anonymised, but with low level of confidence</td>
<td>Some missing observations and unreliable data points, but likely okay for a DSG</td>
</tr>
<tr>
<td>Data is collated, clean, with no missing gaps. The data is completed with full data dictionary and anonymised if necessary</td>
<td>Data is highly appropriate and relevant to the challenge question</td>
<td>Unstructured data / data lakes curated to contain only necessary information so that participants are not overwhelmed by information</td>
<td>Data has been anonymised to high degree of certainty (if containing sensitive information), data subjects / owners have consented to data being used</td>
<td>Very few missing observations (if any) and overall reliable data</td>
</tr>
</tbody>
</table>
Impact

Impact is assessed by looking at the challenge impact scope, learning potential and longevity. Moving towards the green stage is desirable but not a hard prerequisite for participation in the DSG in the way that the challenge and data criteria are. This does, however, increase the likelihood of a challenge being accepted into the DSG.

Impact scope

Challenges should allow for impact in terms of generating tangible positive impact to participant learning during the event, to the challenge organisations directly and to the wider data science community.

Participant learning potential – diversity of potential approaches

No potential for learning for participants by applying approaches

Potential for participants to learn how to use a limited number of approaches

Challenge lends itself to a variety of approaches, but mostly within the same discipline

Challenge lends itself to a variety of approaches and interdisciplinary perspectives

Participant learning potential – novelty of potential approaches

No learning potential; many precedents for applying the suggested approaches to the same or related data

Potential for participants to learn how to use off-the-shelf method in a new use case

Potential for participants to learn about a new methodology in a novel situation

Potential for learning about diverse novel methods

CO impact

No potential impact on production or no infrastructure in place to implement or maintain a solution

Potential insights about potential future research lines; some ability to implement and maintain a solution

Potential extensions to current business lines with a plan in place for how this would be integrated and maintained

Potential to start a business line on itself or to significantly change an existing one with a team in place who will own the solution

Research impact

No potential new insight. Off-the-shelf solutions only

Potential to use off-the-shelf solutions for new data

Potential to mix off-the-shelf solutions with custom approaches that could lead to further scientific outputs in the field

Potential to generate entirely new methods that would advance the field

Follow-on potential

Project is not suitable for continuation of research

Potential for follow-up on a specific branch of the project

Potential for project to initiate longer avenues of research

Project purposely designed as the first step of a long-term partnership

Longevity

Projects must also seek to generate tangible positive impact to both challenge organisations directly (by investigating an operational need where findings can be applicable / translatable to other contexts) and to the wider data science community.

Interdisciplinarity and potential for learning

Working across disciplines forms part of the learning opportunity. The project should have an interesting and unique angle so that the data science ecosystem is enriched by the project.

No potential for learning for participants by applying approaches

Potential for participants to learn how to use a limited number of approaches

Challenge lends itself to a variety of approaches, but mostly within the same discipline

Challenge lends itself to a variety of approaches and interdisciplinary perspectives

Potential for participants to learn about a new methodology in a novel situation

Potential for learning about diverse novel methods

Potential to mix off-the-shelf solutions with custom approaches that could lead to further scientific outputs in the field

Potential to generate entirely new methods that would advance the field

Potential to start a business line on itself or to significantly change an existing one with a team in place who will own the solution

Potential to start a business line on itself or to significantly change an existing one with a team in place who will own the solution

Potential to generate entirely new methods that would advance the field
Annex 3 – DSG project readiness criteria

DSG projects are complex: there are many considerations to factor in when progressing a challenge from the initial proposal. The DSG science team evaluates the project readiness using the criteria listed below.

When the challenge is accepted, the DSG team uses these in conjunction with the challenge’s selection criteria.

The DSG PI and CO should use these criteria as a guide to achieving a great and fruitful DSG challenge.

---

**Challenge**

Challenge readiness is assessed by looking at the project quality and organisational readiness.

---

**Project quality**

<table>
<thead>
<tr>
<th>Ethics approval*</th>
<th>Clear antisocial use of scientific work likely</th>
<th>Not submitted, but okay at initial ethical sense check</th>
<th>Has minor ethical issues which can be mitigated for with minimal amendments to the project</th>
<th>Approved by ethics board</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of challenge questions</td>
<td>Question is not empirical, and cannot be made empirical, e.g. ethical, philosophical, esoteric</td>
<td>Vague and unspecific question(s)</td>
<td>Can formulate empirical proxy that stakeholder is happy with</td>
<td>Question is specific, actionable, and open-ended enough to likely lead to high quality research</td>
</tr>
<tr>
<td>Data fit for scientific purpose</td>
<td>Data and question are related, key quantifiers missing</td>
<td>Data appropriate in principle, but specific dataset is plausibly of too low quality (sample size, SNR, etc.)</td>
<td>Data is robust enough to meet basic challenge goals, but may not provide scope for further exploration</td>
<td>Data can (in principle and plausibly) answer challenge question(s) fully</td>
</tr>
<tr>
<td>Data format and storage</td>
<td>Data not in file format(s) supported by R / Python (n.b. not xls/x)</td>
<td>Data distributed in one or more files of comparable structure/data structure may be non-standard or esoteric</td>
<td>Data in minimal necessary number of formats and linkable by unique ID / enumerator</td>
<td>Data in intuitive structure, well documented, optionally with API or other structured access mechanism</td>
</tr>
</tbody>
</table>

*Must hit green for project go.
<table>
<thead>
<tr>
<th>Organisational readiness</th>
<th>Internal stakeholder availability</th>
<th>Stakeholder data accessibility</th>
<th>Documentation</th>
<th>Communication loop</th>
<th>Readiness of data for sharing with the Turing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data storage located, no data sharing agreement in place</td>
<td>Data owner agrees in principle to sharing of data</td>
<td>Data owner has concrete strategy for data extraction</td>
<td>Data cleared for use in project by legal team (including GDPR)*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stakeholder does not have access to dataset</td>
<td>Stakeholder has access to relevant databases or file systems</td>
<td>All data extracted to disk / online bucket</td>
<td>Data is in analysis-ready format (e.g. delimited text file, feather, json, hdf5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Headings and data types for all data fields present</td>
<td>Descriptions of data fields (e.g. high-level explanation, formulae and/or relationships to other fields)</td>
<td>Description of data collection process and enumeration of known data quality issues</td>
<td>Full formal data dictionary including the rest of this row</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Regular point of contact without data expertise</td>
<td>Expert point of contact for addressing mutual questions, queries, requests</td>
<td>Regular meetings with expert point of contact who is familiar with the dataset, data infrastructure, and procedures</td>
<td>Regular contact and high availability of data creator or equivalent expert</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sharing process not agreed</td>
<td>Sharing agreement approved but not yet signed</td>
<td>Sharing process agreed and data sample shared</td>
<td>Full dataset on site for use by data scientists*</td>
<td></td>
</tr>
</tbody>
</table>

*Must hit green for project go.*
Data

Data is assessed by looking at data readiness and data complexity.

## Data readiness

Considering the datasets themselves, the fewer items below that the datasets feature, the better. These are not a barrier to a DSG challenge, and can be challenge-specific, but the more of these items a dataset features, the slower and more difficult the research will be, making it less desirable for a DSG challenge.

<table>
<thead>
<tr>
<th>Data recording (raw data and acquisition)</th>
<th>Generic data entry errors including typos</th>
<th>Inconsistent data meanings between tables or over time</th>
<th>Changepoints in data collection mechanism</th>
<th>Missing data for some entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data encoding (machine readable 'formatting')</td>
<td>Missing unique identifiers (e.g. to link tables)</td>
<td>Inconsistencies of entity/value encoding (e.g. IBM, I.B.M., ibm etc.)</td>
<td>Missing value code clash (NA vs -1 etc.) or not known</td>
<td>Inconsistent file formats or encoding between files/batches</td>
</tr>
<tr>
<td>Data extraction</td>
<td>Non-representative extract, including deletion bias (e.g. censored samples)</td>
<td>Problematic manipulation before sharing (e.g. aggregates including missing values)</td>
<td>Suboptimal data imputation or 'completion' performed before sharing</td>
<td>Inappropriate merging/joining before sharing</td>
</tr>
</tbody>
</table>
### Data complexity

The table below is a guide for finding a balance with data complexity. This is something that will develop as the challenge is scoped, but good to consider from the outset. Going extreme in either direction is less desirable for a DSG challenge, but at the same time, this is very challenge-specific.

Make the challenge too easy and it will not test and develop participants’ skills. Too complex and this reduces the number of participants that could get involved, thus reducing serendipitous possibilities that the DSG can create.

<table>
<thead>
<tr>
<th>Easier</th>
<th>More complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data volume (figures are ballpark estimates)</td>
<td>Amenable to in-memory operations</td>
</tr>
<tr>
<td>Data velocity (in real scenario)</td>
<td>Static, e.g. concluded study</td>
</tr>
<tr>
<td>Relational complexity of data model</td>
<td>Single table</td>
</tr>
<tr>
<td>Sample dependence</td>
<td>Independent samples: observations are independent instances, e.g. patient B is independent of patient A</td>
</tr>
<tr>
<td>Effective sample size (take into account identical replicates and sample dependence)</td>
<td>ca. 50–100,000 – standard range of effective sample size for most methods</td>
</tr>
<tr>
<td>Data types</td>
<td>Excel table types: only numerical, ordinal, categorical/strings (‘primitive types’), date/time; no missing values</td>
</tr>
<tr>
<td>Modelling complexity (given question and likely approach)</td>
<td>Supervised learning in tabular setting</td>
</tr>
<tr>
<td>Data fit for scientific purpose</td>
<td>Data and question related, key quantifiers missing</td>
</tr>
<tr>
<td>Data generalisability</td>
<td>Data is too specific or biased to generalise</td>
</tr>
</tbody>
</table>
### Impact

Impact is assessed by looking at impact scope and longevity.

### Impact scope

<table>
<thead>
<tr>
<th>Impact to CO</th>
<th>No potential impact on production</th>
<th>Potential insights about potential future research lines</th>
<th>Potential extensions to current business lines</th>
<th>Potential to start a business line on itself or to significantly change an existing business model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impact to research field</td>
<td>No potential new insight and off-the-shelf solutions only</td>
<td>Potential to use off-the-shelf solutions with new data</td>
<td>Potential to mix off-the-shelf solutions with custom approaches that could lead to further scientific outputs in the field</td>
<td>Potential to generate entirely new methods that would advance the field</td>
</tr>
<tr>
<td>Impact to the industry</td>
<td>No changes to the industry</td>
<td>Potential new insights about how to proceed in a specific case</td>
<td>Potential to generate new methods that can complement existing state-of-the-art methods or increase their performance</td>
<td>Potential to deeply disrupt the industry and create a new standardised state-of-the-art</td>
</tr>
<tr>
<td>Impact to society</td>
<td>No positive impact to society expected</td>
<td>Some positive impact could be argued</td>
<td>Expected to bring positive impact to society down the line, challenge serves as a stepping stone piece of work</td>
<td>Potential to bring positive impact to society as a direct result of the research undertaken</td>
</tr>
<tr>
<td>Impact to participant learning</td>
<td>No potential for learning for participants</td>
<td>Potential for participants to learn how to use off-the-shelf methods in a new use case</td>
<td>Potential for participants to learn about a new methodology in a novel situation</td>
<td>Potential for learning about diverse new methods in a variety of situations</td>
</tr>
<tr>
<td>Project multidisciplinarity</td>
<td>Related to a single discipline</td>
<td>Possibility to approach the problem from at least two independent fronts</td>
<td>Potential to approach the problem from at least two different fronts with a possibility of crossover</td>
<td>Potential to approach the problem from a variety of fields, providing new holistic approaches</td>
</tr>
</tbody>
</table>

Moving towards the green stage is desirable but not a hard prerequisite for participation in the DSG and is not situation critical in the same way as the challenge and data criteria.
<table>
<thead>
<tr>
<th>Longevity</th>
<th>Project continuity with partner</th>
<th>Project continuity within the Institute</th>
<th>Partner commitment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Partner does not have a team with sufficient technical expertise to implement and maintain the solution</td>
<td>Partner has a team in place but team does not have required expertise</td>
<td>Partner is not prepared to spend time before, during, after the event</td>
</tr>
<tr>
<td></td>
<td>Partner has a team in place but team does not have required expertise</td>
<td>Partner has a team and a plan of how the expertise will be built for the solution to be maintained and used</td>
<td>Partner is committed to minimum effort – weekly calls, two in-person meetings and provision of data</td>
</tr>
<tr>
<td></td>
<td>Partner has a team and a plan of how the expertise will be built for the solution to be maintained and used</td>
<td>Partner has a team in place with necessary expertise who will own the solution</td>
<td>Champion(s) at partner available to ensure resources are made available during and after fellowship for continuation</td>
</tr>
<tr>
<td></td>
<td>Partner has a team in place with necessary expertise who will own the solution</td>
<td>Project purposely designed as a first step of a long-term partnership</td>
<td>Additionally, provision of further funding</td>
</tr>
<tr>
<td></td>
<td>Project is not suitable for continuation within the Institute/host</td>
<td>Potential for project to initiate longer partnership with the Institute</td>
<td></td>
</tr>
</tbody>
</table>
Annex 4 – The Turing data safe haven
tiers and flow diagram

For information on our bespoke secure environment, the Turing safe haven:

Visit the data safe haven project page.

One-page overview: Poster with overview of our data classification approach, security measures, data management and technical architecture. This is the best one-page high-level overview of our systems and process.

Overview presentation: Slides from our standard presentation about the data safe haven, giving a more in-depth overview.

Overview video: An extended version of our overview presentation that also demonstrates our data classification web application and how to use the environment as a researcher.

Design choices: Our preprint ‘Design choices for productive, secure, data-intensive research at scale in the cloud’, outlining our policies, processes and design decisions for the data safe haven.

<table>
<thead>
<tr>
<th>Tier 4</th>
<th>Very sensitive personal, commercial or government data</th>
<th>Access only from known dedicated secure rooms and stricter package whitelist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier 3</td>
<td>Personal data with weak or no pseudonymisation, or more sensitive commercial or government data</td>
<td>Access only from known physical spaces and access only via managed devices, and whitelisted packages</td>
</tr>
<tr>
<td>Tier 2</td>
<td>Most commercially sensitive data and strongly pseudonymised personal data</td>
<td>Access only from known networks, remote desktop only, no outbound internet, no copy/paste, full package mirrors</td>
</tr>
<tr>
<td>Tier 1</td>
<td>Data with very low consequences for disclosure</td>
<td>No safe haven required, outbound internet okay, access from internet okay, still requires good standard security practices</td>
</tr>
<tr>
<td>Tier 0</td>
<td>Open data</td>
<td>No safe haven required, outbound internet okay, access from internet okay, still requires good standard security packages</td>
</tr>
</tbody>
</table>
How to assess the tier of the projects

Start

Will the research generate (including by linking, sorting or combining) any personal data?

No

Will any project input be personal data?

Yes

Is that personal data made accessible to the general public with no restrictions on use?

No

Is that personal data pseudonymised?

Yes

Tier 0

No

Tier 1

Will releasing any of the datasets or models impact on the competitive advantage of the research team?

Yes

Will you be working with confidentiality/confidential information, or private third-party intellectual property?

Yes

Do you have absolute confidence that it is or will be possible to identify individuals from the data, either at the point of entry or as a result of any analysis that may be carried out?

No

Yes

Do you have absolute confidence that it is or will be possible to identify individuals from the data, either at the point of entry or as a result of any analysis that may be carried out?

Yes

No

Do you have high confidence that the consequences of unauthorised disclosure of this data will be low?

Yes

No

Do you have high confidence that the consequences of unauthorised disclosure of this data will be low?

Yes

No

Tier 2

Will you also be working with confidentiality/confidential information or third-party intellectual property?

No

Yes

Do you have high confidence that the consequences of unauthorised disclosure of this data will be low?

Yes

No

No

No

No

Tier 3

No

Yes

Do you have high confidence that the consequences of unauthorised disclosure of this data will be low?

Yes

No

Did the attacker include sophisticated, well-researched and determined attacks, such as those that may be conducted by state actors?

No

Yes

Tier 4 - reconsider project

No

Yes
Annex 5 – Code of conduct

A condition for all our participants in joining the event is that they adhere to the DSG code of conduct. This extends to the DSG PI plus all attending representatives of the CO organisation.

- **Be respectful to others.** Do not engage in homophobic, racist, transphobic, ageist, ableist, sexist, or otherwise exclusionary behaviour.

- **Use welcoming and inclusive language.** Exclusionary comments or jokes, threats or violent language are not acceptable. Do not address others in an angry, intimidating, or demeaning manner. Be considerate of the ways the words you choose may impact others. Be patient and respectful of the fact that English is a second (or third or fourth!) language for some participants.

- **Do not harass people.** Harassment includes unwanted physical contact, sexual attention, or repeated social contact. Know that consent is explicit, conscious and continuous – not implied. If you are unsure whether your behaviour towards another person is welcome, ask them. If someone tells you to stop, do so.

- **Respect the privacy and safety of others.** Do not take photographs of others without their permission. Note that posting (or threatening to post) personally identifying information of others without their consent (‘doxing’) is a form of harassment.

- **Be considerate of others’ participation.** Everyone should have an opportunity to be heard. In update sessions, please keep comments succinct so as to allow maximum engagement by all participants. Do not interrupt others on the basis of disagreement; hold such comments until they have finished speaking.

- **Don’t be a bystander.** If you see something inappropriate happening, speak up. If you don’t feel comfortable intervening but feel someone should, please feel free to ask a member of the code of conduct response team for support.

- **As a general rule, please be intentional in your actions and humble in your mistakes.**