Data Study Group High Level Scoper Pack

Introduction to DSG

Data Study Groups (DSGs) are an intensive five day 'collaborative hackathons' hosted by the Turing. 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. They present their work on the final day of the DSG and produce a report that will be returned to the CO and subsequently published on the Turing website.

The Role of the Initial High-Level Scoper

Prior to signing any agreements with a CO, an evaluation of the proposed challenge must take place with a data science expert. This is a 1-hour scoping meeting to evaluate whether a proposed challenge:

- aligns with the Turing's mission
- can be potentially reduced into a 3.5-day DSG activity
- has the relevant data to support the investigation (including flagging up potential sensitive data sets)
- has potential follow on possibilities

There will be 3 outputs from the meeting:

- a recommendation to proceed or reject a challenge
- a supporting half-page description of the challenge, as the scoper understands it, with any considerations, concerns or opportunities in support of the recommendation.
- If the recommendation is positive, a 2-sentence high level description of the challenge, with the preferred background of a possible DSG PI

The outputs will be returned to the DSG team and the relevant BDM to feedback to the CO. In the positive instance, the 2-sentence description will be used to advertise for the DSG PI. If you are interested in being the DSG PI yourself, please let us know.

The rest of this document goes into more detail the timeline, the required outputs and guidance on what to look for when assessing a DSG challenge.

Timeline

The potential CO's journey begins when they send in a high-level brief of their challenge, using the enquiry form on the DSG website. If the proposal initially looks sensible, a meeting is setup up with one of the business development managers (BDM). The BDM will discuss in more detail what is involved in participating in a DSG. We need to be sure the CO has the means and drive, as well as the finances, to participate in a DSG. Contracts will be discussed, and expectations set.

Following this BDM meeting, the CO will be given the CO pack which contains everything they will need to know and do to participate in the DSG, and be asked to complete the Challenge Proposal form - a more detailed version of the enquiry form, which they will need to return prior to the high level scope meeting.

On return of the Challenge Proposal form, we will arrange for the potential CO to attend a high-level scope meeting to explore the scientific feasibility of their proposed challenge. The scoper will receive the completed Challenge Proposal form before this meeting.

Scoper Outputs

Scopers must produce 3 outputs. Please find an exemplar high level scope review at Annex 1. Both the review and 2-liner description should be handed back to the DSG team within 1 week of the meeting.

1. Proceed or Reject recommendation
    - Rejecting a challenge doesn't mean it is not suitable for the Turing. Just maybe not a DSG. There are many forms of collaboration with the Turing, with the DSG being only one avenue.
2. The half-page High Level Scope Review
    - The half-page review doesn't need to be a work of art, it simply needs to summarise what was discussed during the meeting/call and document your judgements.
    - The detailed scoping of a challenge consists of four main considerations and must be present in the half-page review:
        - Does it align with Turing priorities? Which ones?
        - Can this initial challenge description be optimised into a well-defined scientific question?
            - The question is *can it*? – it is not part of this role or meeting to actually optimise the scientific question
        - Does the CO have relevant data for this challenge? Is this data sensitive at all (commercial or GDPR)?
        - Is there potential for a follow up projects? Can solving this particular challenge solve other adjacent problems?
3. The 2-line description
    - Only if the recommendation is positive
    - Neatly summarise the challenge at a high level. These will be publicised in the Turing weekly bulletin to recruit a permanent DSG PI. As the description is intended to attract competent data scientists please feel free to use common technical data scientific terms. The most important thing is that the challenge is framed in the best light possible and stimulates scientific curiosity to engender applications of interest.
    - If you are interested in picking up the DSG PI role, please let us know when submitting the recommendation

Things to be aware of during the discussion

In order to fully assess the challenge, it is advised that discussions touch on the challenge itself, the data that will be used and an iteration of what a DSG requires from its challenge owners.

The Challenge

It will be your role to assess whether the challenge you are presented with could have potential as a DSG challenge. A good challenge leverages the strengths of the DSG scheme, in providing participants and challenge owners with an enjoyable and informative experience, as well as creating ample opportunities for impactful follow-up.

To ensure this, the challenge must work well in a 1-week setting, it is expected that challenges will need to pass through many rounds of refining either scaling up or down so don't worry about this too much at this stage - This will be a job for the DSG PI and challenge owner team in the weeks that follow. It is more valuable for you to focus on the challenge question itself.

Concretely, challenges should:

- Be realistic to explore within 1 day of brainstorming and 3 days of data science work.
- Be realistic to address with the data provided.
- Not be at risk due to issues with 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, by real world impact, potential long-term project, or the "right" level of 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.

Discussing the Dataset

It is important for you to discuss the data that the CO plans to use for the DSG, as this will form the basis of the challenge. It may be helpful to ask preliminary questions such as; How was it collected? What do you think its sensitivity might be? What is the state of the data?

Data in poor condition or with a very high sensitivity level will not be suitable for a DSG and the challenge will not be able to progress. Equally too little or too much data can also be problematic, the latter presenting less of a problem as scaling down can be distinguished later with the PI.

DSG Requirements

The DSG represents a big commitment for CO's, it takes up more time/resources than some might assume. It may be worth reiterating the following deliverables to the CO to ensure they understand the process. However, please note that expansion on these points were covered in the BDM meeting and will be revisited by the DSG team so don't focus too much on this section.

All CO's will be expected to;

- Complete the necessary contracts and documentation to participate in the DSG
- Work with the PI (usually at least 4/5 calls or meetings) to refine and prepare the challenge
- Assign a sensitivity tier to their data
- Prepare presentation slides for the DSG

- Ensure a representative is present for DSG week
- Review the final report

Example high level scopes

1. Good review – not a good challenge from CO

Scientific problem:
Predicting function of proteins based on amino-acid sequence, with a view to inform integrated or active learning strategies.
To be used in an integrated exploration/active learning cycle of drug discovery that they claim to have set up as their business model.

Dataset:
They proposed to use a publicly available dataset based on yeast organisms.
Despite multiple queries, they were unable to confirm the exact dataset in the scoping session.
The rep was also unable to confirm whether they have any data in-house or would be ready to share such (more interesting) data sets.

Challenge:
It sounded like supervised prediction from (amino-acid) strings, with a strong transfer learning aspect – but difficult to confirm through data.

Impact:
They claimed they were not interested in scientific research, but primarily in building their "internal knowledge base" through DSG outcomes. They were surprised when I told them DSG are usually about follow-on, with a view towards enabling collaborative research or knowledge transfer.

Summary:
We could run the challenge if the dataset exists, but it might be somewhat boring (and impact-free for Turing).

Confidential comments:

This was a strange scoping session; the rep was very evasive and tried to dazzle with business speak.
In summary, it looked like:
(a) the rep, supposedly a data scientist, had only superficial understanding of the problem itself; in particular it also looked like there is currently no "internal knowledge base" that deserves the name. I.e., the company might have identified a relevant business problem, but have not managed in 5 years to develop the technical means to address it.
(b) in consequence, they may be trying to abuse DSG participants, or Turing, to plug the gaping (technical) hole in their business model.

I have the feeling that it hence may be problematic to proceed on the basis as currently proposed. The technical knowledge they are looking for is essentially what would eventually constitute the entire value of their venture. In my opinion, if we proceed with them, it should be as part of a fairer arrangement – e.g., they should fund a 250k project or so.

2.

Type of data:

- (Unbalanced) panel data: Set of "processes" (that can be grouped by "function") from supply chain, and for each process a set of measurements over time
- Some labels for the different "stages" of some of the processes

Question:

- Automatic identification of the different stages in a process

Desired skill:

- (Supervised) Panel data segmentation

Prior work:

- Apparently someone in-house has obtained preliminary results on this using Random Forests