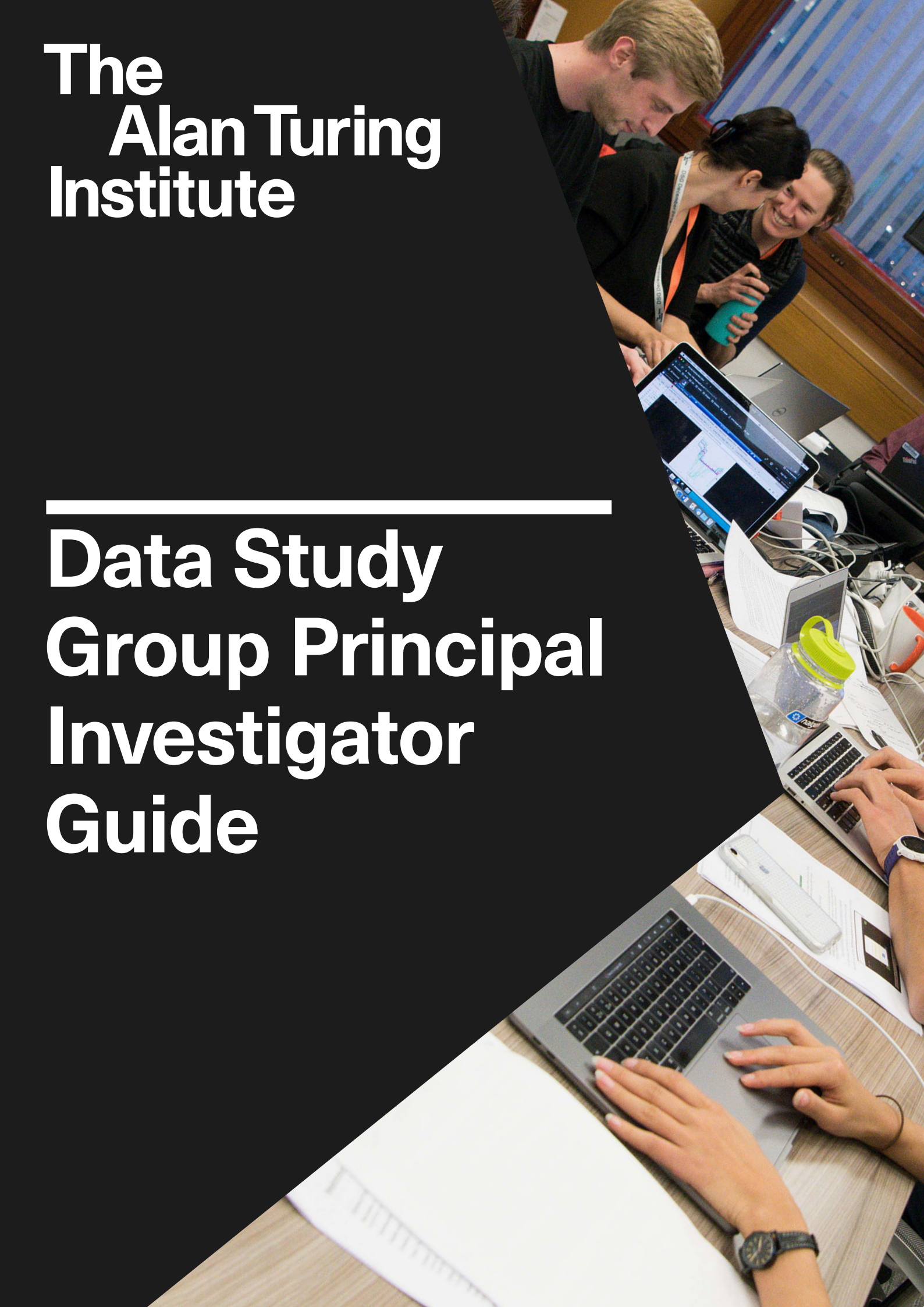
The Alan Turing Institute

Data Study Group Principal Investigator Guide



Summary

An opportunity for post-docs and early career researchers 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.

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.

Intro to Data Study Groups

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 talented and carefully selected researchers.

During the DSG week, researchers brainstorm and engineer data science solutions. 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.

As part of the Institute’s mission to train the next generation of data scientists, we are looking for talented early career researchers who want to act as a Principal Investigator (PI) on a DSG challenge (which, as further mentioned below, have the potential for research follow up).

The Institute will manage the administration so that you can focus solely on matters that directly relate to the science and research.

Lifecycle of a challenge

Organisations can register their interest through the DSG website.

After an initial qualification from our business development team, prospective COs are invited to a briefing, where they will be given an in-depth overview of what is involved in a DSG. During the briefing, COs also speak to one of the DSG executive team's 'Initial Scopers' about their proposed challenges. These preliminary steps allow an initial challenge viability check and ensure that the organisations fully understand what a DSG entails.

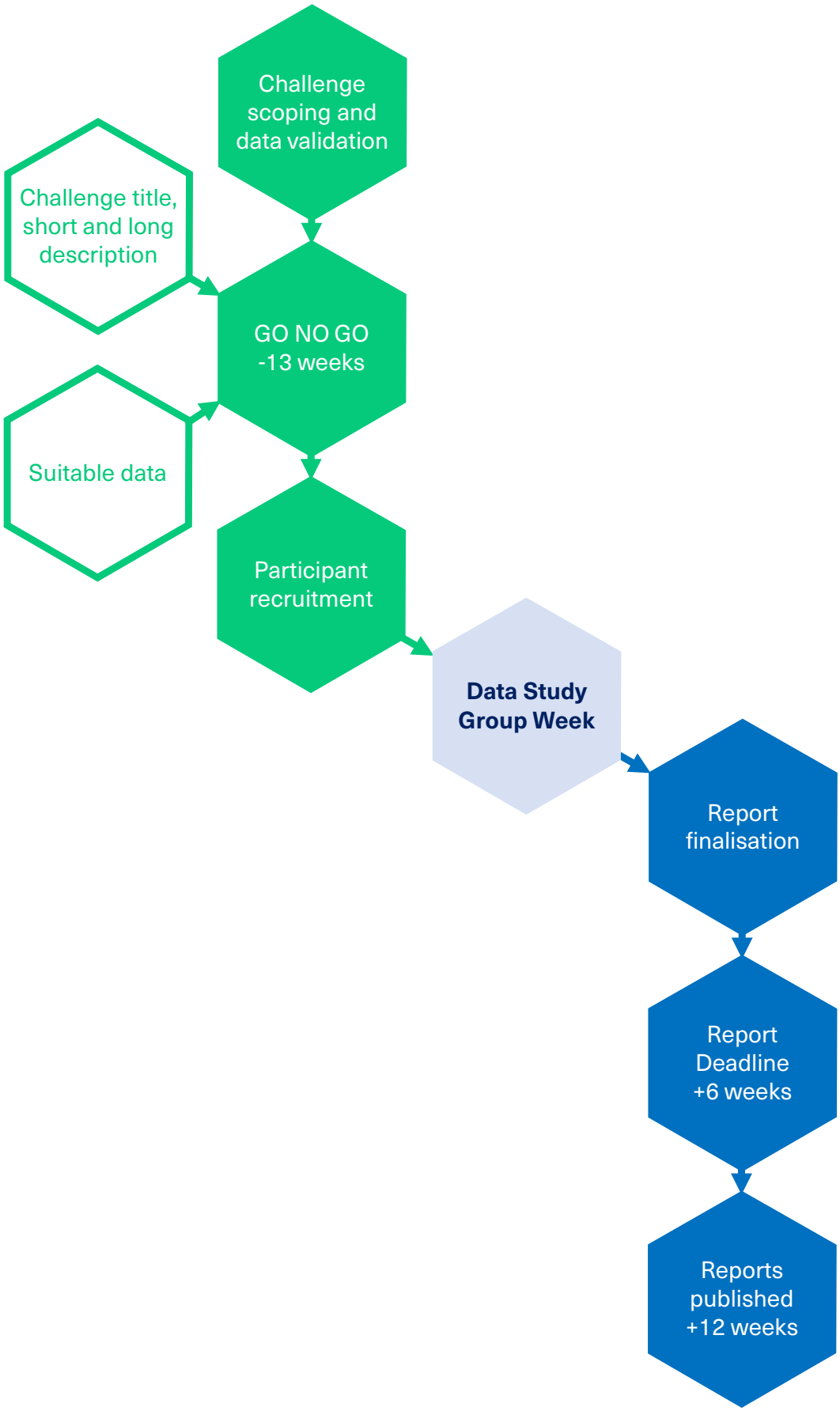
The DSG is quite a large undertaking, not just for the DSG team, but also for the organisations that want to get involved. From the beginning of their engagement with us to the event itself, we estimate about 3-6 months of preparation. COs need to be fully engaged when shaping their challenge, as well as take primary responsibility for providing adequate data. However, they will need help and support on both fronts.

Provided the CO's challenge is aligned with the Institute's aims and can likely be framed within a five-day DSG, the Initial Scoper hands over the detailed scoping of the challenge to the DSG PI. The detailed scoping of a challenge consists of two main considerations:

1. Turning the initial challenge description into a well-defined scientific question.
2. Optimising the outcome of the DSG itself (e.g. report, tangible impact) as well as the potential for a follow up project (which the PI can lead if he/she wishes to).

Business functions, such as contracts, data access, IT provisions are taken care of by the DSG executive team and wider Turing support providers.

Timeline of a DSG challenge



DSG Principal Investigator

The PI will be an early career researcher or post doc. The PI will liaise with the CO only on matters that relate to the science and research questions that will be investigated during the DSG.

They will help turn a preliminary (and often loosely stated) challenge description into a challenge description suitable for a DSG, with potential for follow up work.

They will also need to consider the type of data being provided (e.g. data sensitivity tiers and how that will affect the participant experience during the DSG).

The role of the PI will last on average 6 months per DSG challenge. It is possible to have multiple PIs per challenge, provided all PIs involved agree to this. This is a voluntary role.

Responsibilities

Challenge definition

In collaboration with the CO, the PI will help shape the challenge into a research question that can be tackled during 3.5 days of intensive work (Monday am is the CO pitches, and presentations take place on Friday morning. Friday afternoon after lunch is the official close).

The PI and CO will collaboratively write the following documents, which will help steer the challenge through to the DSG:

- A challenge title, with a 2-sentence description and a short list of relevant participant skill sets
  - This will be used in the recruitment of participants for the DSG
  - It needs to be signed off by both the CO and the Turing prior to publication
- A detailed 2-page description of the challenge (roughly 2 pages, includes detail on for e.g., data at hand, question(s) tackled, approaches suggested, output expectations)
  - This will be included in the delegate pack for the event, which is sent to the participants about 2 or 3 weeks prior to the event taking place.
  - It needs to be signed off by both the CO and the Turing prior to publication
- Both are required about 14 weeks before the DSG event for ethics review

More information on what makes a good challenge found in Appendix 1 at the end of this document.

Data

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

For instance: What security tiers does the data belong to? (Please refer to the paper 'Design choices for productive, secure, data-intensive research at scale in the cloud'.) Is the data readily available? Is the data suitable for the challenge at hand? If not, what additional data is needed (if any)? Information on the security tiers and the tiers flow diagram are in the appendix of this document.

The CO will provide access to the data before they transfer the main tranche to the Turing. The PI will need to assess whether the data is challenge-ready before public challenge announcement (which usually occurs 7 weeks before the DSG) and if not, advise the CO on what needs to be done in order to achieve data readiness in time for the DSG.

Additional support can be provided by the research software engineers (at the expense of the company and subject to availability) or the data scientists on the DSG executive team.

Ethics

The PI will need to submit an Ethics Advisory assessment before the DSG to ensure that any potential ethical issues have been thought of and mitigated. These can take an hour to complete on average and need to be completed before the Go No Go date.

The ethics advisory group (EAG) can be found on Turing complete, the Institute intranet. If you do not have access to the intranet, please contact us for the latest version of the EAG introduction document.

Key questions to be answered regarding the ethical implications of this research:

- Consent
- Privacy and Security
- Other harms

Please see pages 5 – 8 of the 'Introducing the Turing Ethics Advisory Group' document. The Ethics Advisory form is a web form accessible through the Turing intranet. If you are not a Turing affiliated fellow you will not have access to this.

Non-Turing affiliated researchers should review page 18 of the 'Introducing the Turing Ethics Advisory group' and send your answers to the DSG project manager.

Deliverables

Reports

Reports are a key outcome of the DSG as they provide tangible output for organisations, are publishable by the Turing, and form the basis for any follow up research work.

During the DSG event week, participants will produce reports explaining their work and main findings. On the Wednesday or Thursday of the event week, the PI should review the in-progress report and provide feedback to improve it.

Following the DSG event week, there will be a period of report revisions to ensure the document is publishable for the Turing website. This is usually conducted by one of the DSG participants who is hired on a short-term basis to complete the report, or by the PI themselves. The task will be to collate 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.

If one of the participants in completing the report, the PI should provide guidance and feedback to the report finaliser to help optimise the quality of the final report. The minimum engagement is giving detailed feedback in weeks 2 and 4 post-DSG, but closer collaboration is encouraged, especially if the intention is to take the project forward post event.

After these iterations, a final editing and review cycle is carried out by DSG editors and the Turing communications team.

If the challenge was tier 1 or above, a final declassification review is also conducted to ensure no sensitive information is published. The CO will also have the opportunity to review the report prior to publication, for minor redactions relating to confidentiality.

You can read some of the reports from previous DSGs here and a blog on a good DSG report here.

Code

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

The only action on the part of the PI is to declassify code if the data was tier 1 or above.

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).

What's in it for you

As a PI you will have the opportunity to work with industry, government or third sector on real data science problems. This 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
- Find out how research areas of interest to you can actually be used to generate real world impact
- Learn to work at the cross section of industry and academic (a territory of interest to many!)
- Grow your scientific and industrial network
- Learn to identify and shape collaborative opportunities
- Help foster relationships between the Turing and industry (these may develop into programme collaborations or even strategic partnerships)
- Help shape and lead any potential follow up projects (e.g., paper, internship, etc)

We note that, should any follow up project come out of the DSG, the PI is given the option of first refusal to continue to be involved with the challenge.

"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 DSP PIs: Firstly, I think that it is very important to meet with the challenge owner 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

FAQs

When is the DSG?

DSG is held around 3 times a year at the Turing offices in London. We are also piloting a DSG Network event, which is taking the DSG model to the University Partners. A single DSG event usually hosts around 5 – 6 challenges.

How is the DSG week structured?

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

Day 2-4: Brainstorming, modelling and problem solving

Day 5: Progress and recommended routes forward are presented

Is the DSG a cheap consultant?

No – Whilst the CO dictates the scope of the challenge, it is up to the PI to shape the challenge into an interesting research question. And 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.

IP?

Any 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 license.

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 day (group forming and presentations). Each group has a facilitator with whom you will be connected with 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 data set 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 into this?

For challenge scoping we estimate perhaps 2 or 3 meetings (phone or physical) with the CO to draft the title, short description and long description, with an additional couple of hours to finalise.

The ethics form should then be less than an hour as primarily you will be able to copy from the long description.

Data evaluation could take a little longer depending on the size and complexity of the data set.

For report review we estimate a minimum 6 hours over the course of a month – to review and feedback to the report writer.

What should be included in the report?

Please see here for a complete description.

Can the report be used for my university REF?

The report would 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).

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 that we try and continue as many of the projects that we have started at the DSG in some form of follow on work. This could be from writing a paper to extending the research, working closely with the CO continuing to investigate the results from the DSG more closely, to forming longer term partnerships.

As the PI for the project, in these kinds of situations, you would have the first right of refusal to be involved in such follow-on projects.

The DSG team will also ask participants for an expression of interest in any of the week's challenges, creating an instant cohort of researchers to choose from.

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 experiment with a multitude of approaches. The DSG PI does not need to be affiliated with the Turing but should be affiliated with one of Turing's Partner Universities.

Will I get paid?

No – this is a voluntary role. We want to give early career researchers who may not have had the opportunity to work with 3rd party organisations the experience and skills to sit at the cross section of academia and industry.

How do I get involved?

The Turing will post specific challenge calls in the weekly bulletin when available. Alternatively, you can register your interest by sending an introductory email to datastudygroup@turing.ac.uk listing the types of projects in which you would be interested.

Those registered will be contacted first prior to advertising the challenge in the weekly bulletin.

Appendix 1
What makes a good 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 PI and challenge owner team should make sure that a challenge is presented which works well in the 1-week setting of the DSG.

Concretely, challenges should:

a. Be realistic to explore within 1 day of brainstorming and 3 days of data science work

b. Be realistic to address with the data provided

c. Not be at risk due to issues with data sharing, e.g., ethics, technical restrictions, privacy constraints, or data quality.

d. 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

e. Focus on analytics and AI, rather than on rote tasks such as data munging, data curation, or data scraping.

f. Be appealing to participants, by real world impact, potential long-term project, or the "right" level of data scientific challenge

g. 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 looks as follows:

i. Participants with the right skills read the challenge 2-pager, and during week self-assign to the challenge since they are interested and feel they can contribute

ii. During the week, the challenge team of owner representatives, PI, and participants produces proof-of-concept solution for the "Low hanging fruit" challenges, and brainstorms a series of approaches for the wider context

iii. On 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.

iv. Results of the follow-on project get published in major scientific venues, and/or lead to disruptive business innovations which in turn inform further collaborative research projects, embedded in a long-term partnership network

Common failure points that create a "bad" challenge arise from the negation of "good practice" should be avoided:

a. The challenge scope is mistakenly set up as a 1-year research project rather than a 3-day exploration.

b. The data provided is unrelated to, or insufficient to answer the challenge questions - the challenge owner expects "magic".

c. Data gets stuck due to last-minute problems, or is shared in terrible shape.

d. Questions are phrased too vaguely, or are without exception extremely ambitious. The participants have no idea what to do or where to start.

e. The data is not ready-for-analysis, or in such a bad shape, that participants spend the entire week cleaning it up rather than doing anything interesting.

f. The challenge is boring or impact-free, e.g., by having a narrow business or academic focus, by being addressable by rote technology, or by likely entailing analytic tedium.

g. The challenge owner has not thought about what to do with proofs-of-concept or results of scientific brainstorming, e.g., conversion into follow-on or scale/ scope expansion.

Much of this can be mitigated or entirely prevented with due diligence in preparatory work:

a. This is a crucial part of scoping done by the challenge PI: concretising and narrowing down the question, so it is still interesting, representative of the application area's broadness, while defined enough to be tackled in one week productively.

b. This is also a crucial part of scoping done by the challenge PI. Applying basic data scientific principles, it needs to be checked whether the data supports the question - e.g., rep-resentativity of the sample, presence of an intervention/ instrument (for causality), etc.

   If the data is insufficient, one can see whether one can weaken the questions, select an area which the data can support, select reasonable proxies, or recommend collecting different data and return to DSG at a later stage.

c. This can be mitigated by ensuring the data is shared at least 2 month in advance of DSG, and before publication of the challenge. Note that, assuming due diligence on the side of DSG team, only having the data already, and having it in good quality, is a 100% guarantee for having it in time for the DSG, and having it in good quality.

d. Again, this is subject to scientific judgment in the scoping phase, by the challenge PI.

   A good rule of thumb for is to have one or two simple off-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 allow to identify a technical solution (as opposed to non-solutions).

e. For data quality, the challenge PI should iterate with the challenge owner, 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. Participants should be able to understand the data from its documentation alone.

   Regarding quality, all major issues should be resolved before DSG. It is possible that the DSG week will turn up further issues, but these should no longer be "obvious" problems that bring the work to a halt in the first couple hours.

f. The "interestingness" of the challenge is perhaps the most subjective judgment call to be made jointly by challenge PI and challenge owner, among those listed. Here it helps to measure likely outcomes against common qualifiers of impact, i.e., knowledge transfer success, business disruption, academic excellence - as well as considering the participants' common motivators on why to get involved: decent technical challenge level, impactful applications, research potential, follow-on opportunities.

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

   It is recommended that challenge owner and challenge 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 offer 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 challenge owner.

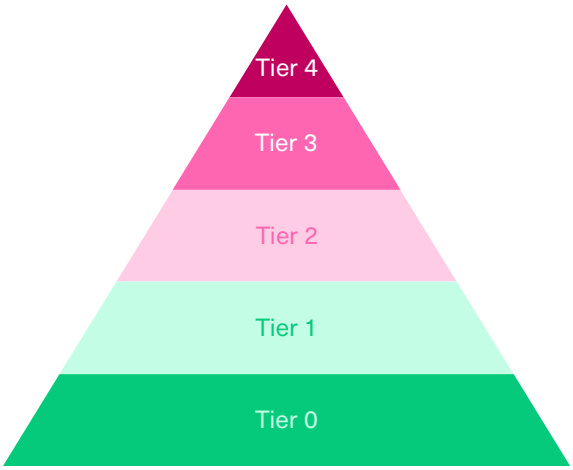
Addendum, for DSG organising teams only (not to be shared with challenge teams)

In addition, one may want to keep in mind the following "special" considerations which may or may not apply to a given challenge:

- A challenge owner may not be ready, organisationally, to specify a DSG challenge yet. This happens often if an organisation has just started to adopt data science infra-structure. 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.
- A related, smaller version of the problem is lack of data scientific competence while 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 challenge owner, outside the DSG, since participants are not data cleaners, see (e).
- Some challenge owners 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 usually poses the danger of reputational damage since it is usually the same challenge owners who are completely uninterested in valid outcomes or scientific rigour, as opposed to making grandiose marketing claims. Warning signs are vague claims or vague 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.

Appendix 2
Turing Data Safe haven tiers and flow diagram



Our model goes from Tier 0 – publicly available, open information – to Tier 4 – personal data where disclosure poses a substantial risk to safety.

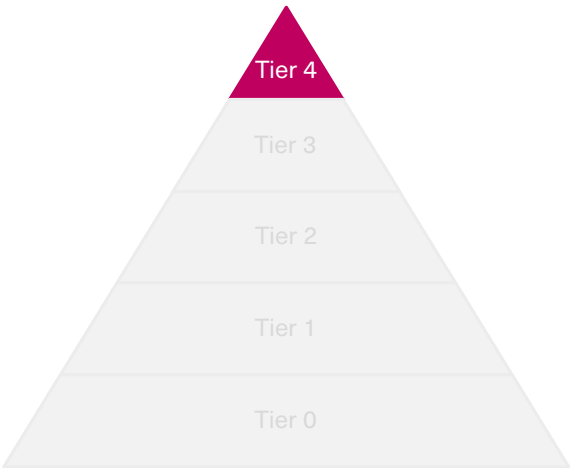
This guidance document should give you an idea of how to classify your data for the Data Study Group.

It should be referenced in conjunction with the classification flowchart.

Tier 4 environments are for:

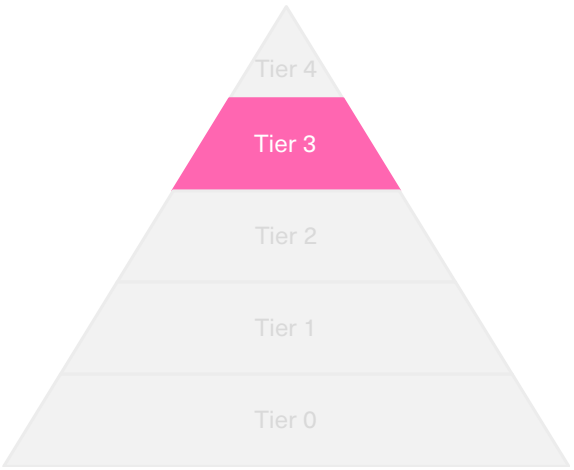
- Personal data where disclosure poses a substantial threat to safety, security or health
- Commercial or governmental data which could be subject to attack by sophisticated, well-resourced and determined actors such as nation states

Tier 4 data is **not** appropriate for Data Study Group use.



Tier 3 environments are for:

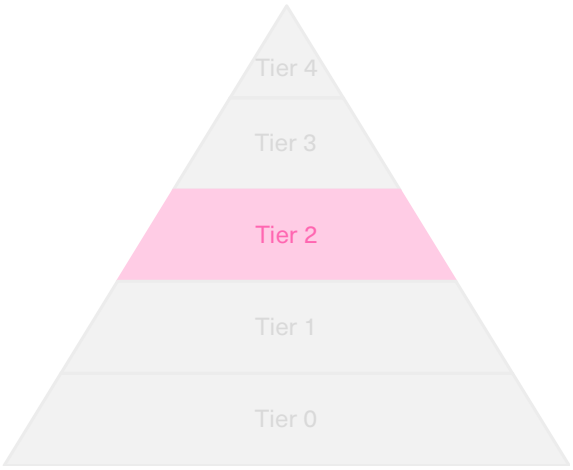
- Pseudonymised or synthetic data where confidence in the quality of anonymisation is weak
- Commercial data which is sensitive
- Commercial or governmental data which could be subject to attack by attackers with bounded capabilities such as hackers



Tier 2 environments are for:

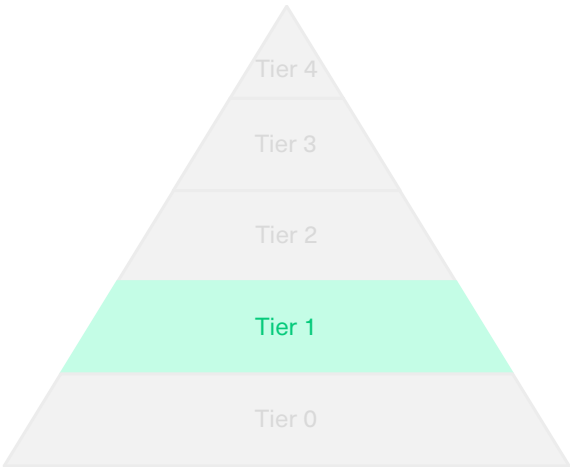
- Pseudonymised or synthetic data where confidence in the quality of anonymisation is strong
- Commercial data where risks from disclosure are low

Tier 2 data should be very unlikely to be subject to targeted attack.



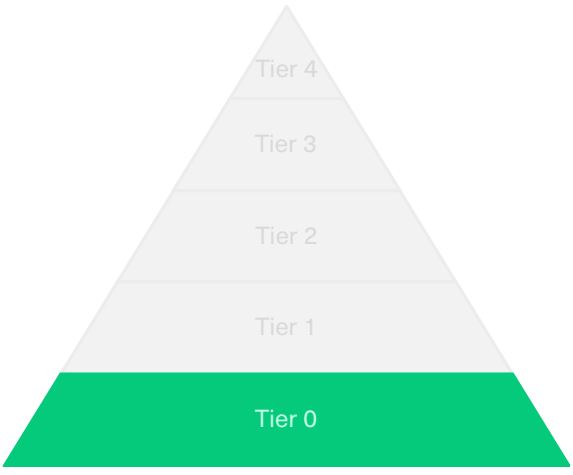
Tier 1 environments are for:

- Data intended for eventual but not immediate publication
- Datasets where the only risks of disclosure are to the researchers' competitive advantage
- Pseudonymised or synthetic data where confidence in the quality of anonymisation is absolute
- Commercial information where the consequences of disclosure are so low as to be trivial

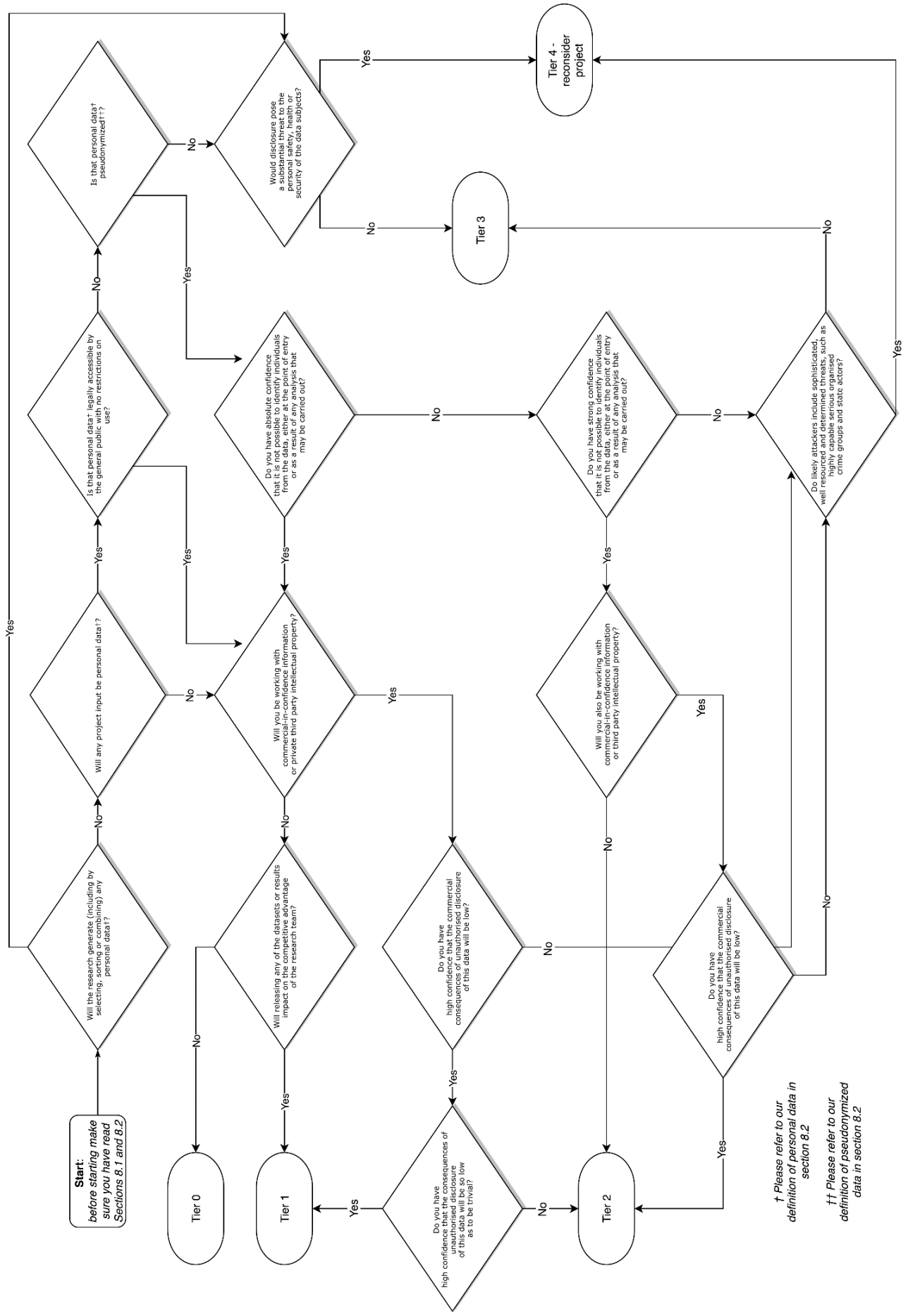


Tier 0 environments are for:

- Publicly available, openly published information
- Data which is intended for immediate publication



How to assess the tier of the projects



Please share this opportunity with your research teams

turing.ac.uk
@turinginst