**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:

1. Be realistic to explore within 1 day of brainstorming and 3 days of data science work
2. Be realistic to address with the data provided
3. Not be at risk due to issues with data sharing, e.g., ethics, technical restrictions, privacy constraints, or data quality.
4. 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
5. Focus on analytics and AI, rather than on rote tasks such as data munging, data curation, or data scraping.
6. Be appealing to participants, by real world impact, potential long-term project, or the “right” level of data scientific challenge
7. 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:

1. The challenge scope is mistakenly set up as a 1-year research project rather than a 3-day exploration.
2. The data provided is unrelated to, or insufficient to answer the challenge questions - the challenge owner expects “magic”.
3. Data gets stuck due to last-minute problems, or is shared in terrible shape.
4. Questions are phrased too vaguely, or are without exception extremely ambitious. The participants have no idea what to do or where to start.
5. 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.
6. 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.
7. 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:

1. This is a crucial part of scoping done by the challenge PI: concretizing 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.
2. 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., representativity 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.

1. 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.
2. 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).
3. 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.
4. 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.
5. 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 organize multiple follow-on activities, including follow-on workshops, spring/summer projects, project organization etc, subject to an organisational partnership with the challenge owner.

**Addendum, for DSG organizing 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 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.  
  A related, smaller version of the problem is lack of data scientific competence while data storage infrastructure is present. In this case, organisatorial 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 no 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.

--- general thoughts/bullet points

* Trade offs
  + Scientific quality, Impact, Data Quality, Stakeholder
* Interdisciplinary challenges
* Data readiness

A good challenge suitable for the data study group is more than just the data or the question

* Data:
  + quality (collection mechanism )& cleanliness - is there missing or inaccurate data?,
  + sensitivity (Tiers),
  + Size # data points, physical size
  + Accessibility - how easy is it to get the data ready and transferred to the Turing
  + richness (always fun, within reason) -
* The Question:
  + can be tackled with data provided,
  + can be progressed on in 5 days,
  + Question is challenging, interdisciplinary or slightly non-standard
    - (i.e., while we have some ideas on how to tackle it we don’t know for sure what will work, you won’t have the full answer in five days & there is room for follow up: DSG is a proof of concept)
* Stake holder buy-in
* Challenge Owners: someone (or > 1) competent from the company willing to collaborate with Turing and academics on data & question all the way through — note that there is often more preparation that expected, ideal combo: someone high level + people acquainted with details of data, availability during event itself
* Follow up: desired outcome post DSG