Themes – provocations and questions – that TMCF will explore

We outline here ***three themes*** that we would like to use as vantage points and to structure discussions during the week. Under each theme, you will see a list of provocations and a few questions that we would like to explore.

i) ***Modelling*** *paradigms for data-driven science.*

Provocations

- *Heuristics trumps theory in data-driven research.* Model-development and knowledge-building is best approached by consulting data and domain praxis rather than isolated theory (see Wolf, 2023 on background and consequences of this position).
- *Models are exploratory artefacts.* Model-building – and visual methods – in data-driven analysis is about exploring different structure and outcomes that might have been generated from the observed data processes – for example, we don't use null hypotheses to mark out confidence intervals, but to simulate from our data to explore different outcomes (see Hullman and Gelman, 2021).

Questions to explore

What is distinctive about, or what characterises, statistical modelling in data-driven science? And what aspects of modelling practice (including ideation, development, selection, refinement, and evaluation) often get under-emphasised or forgotten?

Some clarifying notes: Here it might be useful to think of "data-driven science" in opposition – in its context, goals and practice – to statistical modelling in traditional designed experiments, but also the sorts of predictive modelling workflows in supervised machine learning – e.g. recipe of train under cross-validation with basket of ML algorithms, tune and select ML algorithm, validate through out-of-sample prediction.

ii) ***Inference*** *and* ***replicability*** *in data-driven science*

Provocations

- *Claims to knowledge can only be made through out-of-sample significance tests*. Data-driven analysis, and especially visual methods, induce false discovery via unchecked multiple comparisons.
- *Pre-registration locks researchers into facile statistical tests.* Without extensive exploratory analysis and visualisation, data-driven analysis will lead to weak or straw-men hypotheses.
- *Human-in-the-loop is incompatible with inferential and replicable analysis*. Exploratory and data-driven practices are associated with high researcher degrees of freedom – iterative, flexible and informal workflows – threatening replicability, formality and rigor and making inferential claims impossible.

Questions to explore

- How do we ensure that the inferences and claims we make from data-driven analyses are properly contextualised?
- Can pre-registration study designs be developed for data-driven science? How would they differ from pre-registration designs in experimental science?
- Can we formulate principles and guidelines for evaluating research findings claimed from informal, data-driven analyses?

iii) *From **analysis** to **communication***

Provocations

- *Visualizations are limited as evidence.* Many (subjective) decisions go into the design and generation of visualizations and they are open to the 'interpretation' – so they are never objective artefacts that can be trusted and shared as evidence in data-driven science.
- *There is no formal beginning, process or an end to an interactive data analysis session, it is all context-dependent.* Without a formal research design, decisions on when to "stop" analysis and communicate findings in data-driven science are arbitrary and ad hoc.
- *Provenance of exploratory data analysis processes are too complex and ad hoc to be useful.* Multiverse analysis for addressing forking-paths sounds great, but in data-driven analysis "decision points" are many and seldom clear – and fundamental problems of data and construct validity can soon be forgotten – it soon becomes overly complex and substantively problematic for practical use.

Questions to explore

- How can the context behind analytic decisions be recorded and how to balance informational complexity (context) with incisiveness of claimed findings?
- What are the limits to *multiverse analysis and other attempts to expose hard-to-quantify sources of uncertainty?* What constraints / frameworks can we impose on multiverse for practical, data-driven research?
- How can visualisations be designed and presented to facilitate rigour, depth and richness in data-driven science (Meyer and Dykes. 2019)?
- What future is there for open source tools for multiverse analysis (c.f. Sarma et al. 2023)?