Provocation (a data-driven social scientist weighs in)

TCMF: Navigating the garden of forking paths: theoretical foundations for interactive data analysis in data-driven science

AUTHOR
Rachel Franklin

> *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.

Long before researchers reach the gate to the garden of forking paths, where interactive analysis and visualisation are concerned, they will have traversed myriad other intersections of alternate analytical possibilities. Occasionally intersections will be well-marked: for example, problem spaces and research questions informed by theory or previous research, but nevertheless representing researcher degrees of freedom. Other times crossroads are less apparent, as with data production, encoding, structure, or variable creation, where decisions were likely made upstream of an individual researcher and resulting data are treated as commandments on stone tablets. And then, of course, we have data cleaning and wrangling, obviously results of researcher choices, although still typically treated as anticipatory to analysis, whether exploratory or confirmatory. Studies of migration or demographic change offer useful illustrations.

Although I'm betting the sharp divide between data prep and analysis isn't as clear as it's made out to be.

My point is *all* inferential and replicable analysis is human-in-the-loop. Probably, we can mostly agree on this. The evolution of research and science is like a whack-a-mole game where, just as one area of weakness is identified and addressed, another emerges in its place. Research communities respond with processes or systems designed to mitigate impacts of whatever new research crisis has emerged. Hence the creation of peer review, "route maps", pre-registration, and our current fixation on open science and reproducibility. Don't get me wrong: each of these innovations in support of robust science is, in theory, immensely useful. In practice, however, each can impart false security to researchers and consumers of research in a form of validity-washing.

What does this imply for data-driven science and, especially, interactive data analysis? Tools that nudge, encourage, or force a degree of explicit inference (e.g., lineups for visual inference) promise improvement on the status quo but may also 1) hard-wire some forking paths into commonly used tools or packages and 2) aid and abet further conflation of analytical complexity with quality. At the same time, I am excited by the analytical and pedagogical utility of approaches—whether narrative, procedural, or tool-based—that emphasise the importance of confirmation of findings and the sneaky influence of implicit inference. This includes new toolkits, but extends to wider embrace of a cross-validation (and counter-factual) mindset in the social sciences, renewed attention to social science theory, or more intentional focus on replication.

Theory can provide a form of pre-registration that clearly delimits what is being tested and how.

I'm using `replication` a bit loosely here to include formal replications, but especially research that asks the same question but in different settings, with different data, or with different models. For example: the effect of class size on student learning. For this to work, researchers need to *read* the literature.

Big picture: I am both sanguine and sadly resigned with regard to data-driven knowledge production. I have confidence that knowledge grows through repetition: researchers hammering away at a topic in a not-too-standardised fashion. From repetition emerges a distribution of findings which, hopefully, signals consensus. That suggests tools should embed openness and flexibility alongside inference assistance, lest we inadvertently constrain the output distribution. I also look at the mechanics of data-driven analysis and occasionally (not always!) feel we substitute analytical fixes for what are ecosystem-level problems: lack of integrated training, pressure to produce and find “insights” quickly, and incentives for innovation and easy wins over substantive research that builds on existing knowledge. Bright spotlights on only one piece of analysis production, like interactive visualisation, can blind researchers to other decisions hidden in the shadows.