TMCF 2024 Provocation: Cook

This is written in response to ii) Inference and replicability in data-driven science.

"*Weedy sea dragons are predominantly sighted during warmer months during peak tourist season*", "*COVID-19 caused a burst in sightings of trapdoor spiders around Maroochydore*", "*Milder temperatures and gentler rain beckoned more sightings of mountain pygmy possums*". **Faulty conclusions arise from interactive graphical analysis conducted in the absence of clearly delineated data collection procedures measuring the population of interest.**

From the trenches of teaching, fallacious conclusions are routinely reported in exploratory data analysis assignments. The source of the data and the way it is collected are of primary importance for making inferences, and routinely ignored by the analyst. The examples above arise from observations taken from the Atlas of Living Australia, where sightings are reported by human observers. To have an observation recorded (almost always) requires a human to be present. So the question is whether the animal is only active then, or is the human only active then. We can often calibrate numbers by adding missing information. Counts of COVID-19 per region re-computed per 1000 people using population statistics. Failure of appliances reported in warranty claim databases can be calibrated by adding sales counts. The Atlas data cannot be augmented with human occurrence data so easily.

A long time ago, I was analysing multivariate liver function measurements of patients in clinical trials at a pharmaceutical company. The sample measured was quite different from the normal range provided. It turns out that my sample were measurements on women, but normal range was conventionally computed on males of a restricted age. This lack of coverage of the full population in samples used for training complex models is a contemporary concern in the media and research today. Other issues are dependencies in the data such as repeated measurements, temporal and spatial context, hierarchical survey instruments, and weighted data.

A few years ago, in the summer before the pandemic started, the east coast of Australia was on fire, raging bushfires destroying vast areas of native bushland and coastal towns evacuated by sea because all roads were blocked by fire. Social media was rampant with accusations of arsonists starting the fires. I was curious about the cause of fires and what might be done to prevent them in the future. Using data collected by the Vic government on fire ignition causes, along with newly available satellite hotspot data recording potential fires, we collated data on weather, distance to human activities and vegetation with the goal of modelling fire cause. Interestingly, there was a big difference in the Vic government data with fire ignition cause and locations of fire ignition as given by hotspot data. So mostly our predictions were that the horrific bushfires were caused by lightning, because most ignition spots were in remote locations. The Vic government data was recorded by county fire officers investigating fires, so was much more closely connected to humanly-accessible locations.

To make inference we need to clearly specify the population of interest, including context and dependencies, and how the observed data relates to this. The infrastructure for making useful interactive graphics enabling inference needs

- tidy mapping from population to sample
- clear definition of the full data and uncertainties
- proper grammar for rendering plots and interactions
- methods to break dependencies and approximate populations
- capabilities for comparisons, what-if scenarios
- support for multivariate, high-dimensional visualisation
- robust software implementations

This perspective is guided by my background as a statistician, and my work in three areas: high-dimensional data visualisation, inference for data graphics and data pipelines. I have worked for many years on high-dimensional visualisation, especially as it can be used for exploring data, and to break open black-box models to gain some understanding of where and why they work, or don't work. In my work on inference for data graphics, we see a data plot as a statistic summarising some aspect of the data, that can be compared to what other samples from a population might look like. My work on data pipelines focuses on mapping from observed data to linked plots, that provide some insight under-the-hood, to see the whole data and avoid cherry-picking.