Data Science and Data Abuse

Mengyu Jackson

The “Reproducibility Crisis” in the scientific community is a topic I’ve heard mentioned a few times in the past couple of years, but I didn’t dive deeply into it until this past week. It is surprising to me just how big and serious the crisis is. If you aren’t familiar with it, this Stanford article is a very comprehensive overview of the crisis, but it’s also not a quick read. I’d like to share some of the most interesting points from this article as well as discuss some lessons we can learn in Data Science.

One large scale replication attempt by Reproducibility Project: Psychology involved 270 crowd sourced researchers in 64 different institutions in 11 different countries, and found that only 36% of results were successful in replicating the same finding (or lack) of statistical significance. The Effect Size in replication studies was only half of the effect of the original studies. This is a result that shocked me. More than half of published research didn’t have the same result when replicated, thanks to publication bias and “Questionable Research Practices” such as, P-Hacking, and HARKing. What interested me most is that some of these “questionable research practices” seem very similar to standard practice in data science.

In Data Science, exploratory data analysis is often done without any particular hypothesis, and we find correlations in the data without any prior bias. Once we find a relationship, we build models to predict outcomes based on those correlated variables. The fact that “Hypothesising After Results are Known” is a QRT that makes it more likely to publish irreproducible results is very surprising, because results in Data Science are very often reproducible.

Null Hypothesis Significance Testing (NHST) is diagnosed as the root cause in several papers referenced in the Stanford Article, largely because the “dichotomous nature of NHST facilitates publication bias”. I don’t fully understand how this (and other factors surrounding research publication) make practices that work perfectly every day in Data Science problematic in this domain, but I plan to continue digging into this topic and will be sure to share any interesting results back here!