In: Emre Bayamlıoğlu, Irina Baraluic, Liisa Janssens and Mireille Hildebrandt (eds), BEING PROFILED:COGITAS ERGO SUM. 10 Years of Profiling the European Citizen, 2018: Amsterdam University Press. 106-109. DOI 10.5117/9789463722124/CH18.

# Induction is not robust to search

*Clare Ann Gollnick**

**Abstract**

Data cannot be systematically examined and selected while still being compelling evidence. Selecting specific facts or observations that are consistent with a strategic goal is known colloquially as 'cherry picking' evidence. Frustratingly for those in pursuit of truth, it is still cherry picking even if what one attempts to select for is validity or quality. One cannot learn from data by iterative search. This creates a practical problem for scientists and data scientists, whose professions are centered on using data to understand and predict the world. Given the goal of identifying a valid model or hypothesis, how does one choose the correct one without effectively cherry picking from a set of poorly performing models? In this provocation, the reproducibility crisis in experimental science in biology and psychology is examined within the context of the limits of inductive logic. Scale is identified as a contributed cause in the reproducibility crisis, extending the reach of the crisis the not only from experimental science but into machine learning and automated intelligence as well.

**Keywords:** p-hacking, statistical inference, reproducibility, machine learning.

## Introduction

A data scientist's goal is one of translation: from data to knowledge to action. After defining a hypothesis but before making a decision, a critical step in the process is transforming data into evidence and assessing the quality of that evidence. Data does not speak for itself. The same observation in disparate contexts will support a disparate set of conclusions. Thus formalized inductive logic, including statistical inference and machine learning, require quantification of both the data and the context. Probabilities and probabilistic reasoning are used almost exclusively to quantify evidence as these frameworks combine observation with context into question-agnostic, cross-disciplinary metrics (1 in 100 chance, 95% confident, 99% accurate etc.).

### Automated intelligence (AI) is based on search and selection

As data analysis evolved from a means to an end to a profession in and of itself, there have been substantial efforts to automate and scale the inductive process. Basic statistical inference requires a hypothesis, observed data, and a probabilistic measure of evidence. Automated intelligence additionally requires: 1) searching, testing many hypotheses or candidate models at the same time or in quick succession and 2) selection, choosing among the candidate hypotheses the 'best' one to be used in decision making.

Search and selection make it difficult to accurately represent context. The role of context is best understood with an example. A scientist believes that a coin is fair. She performs an experiment in which the coin is flipped repeatedly, recording the outcomes. The scientist observes eight 'tail' outcomes consecutively. The data (eight consecutive tails) is raw observation. The total number of times the coin was flipped is relevant context. If the coin was flipped a total of ten times, the series of eight consecutive tails is substantial evidence the coin is not fair. However, if the coin was flipped ten thousand times, one series of eight consecutive tails is still consistent with the hypothesis of a fair coin. If the number of total flips is unknown, it is difficult to make any statement with respect to the hypothesis.

Automated intelligence uses probabilistic reasoning and inductive logic outside the confines of controlled experiments or defined context. Multiple hypotheses and uncontrolled variables are tested simultaneously. With coin-flip data, for example, the goal may be to predict the outcome of the next coin flip without knowing details of the experiment. Potential predicative hypotheses may include: 1) the flipping process is biased; 2) the coin is changed mid-experiment to a new coin at random; 3) both the coin and flipping process are biased but with different degrees and direction; 4) the coin was made of chocolate and bias was influenced by room temperature. The inclusion of bizarre hypotheses is used to drive the point: with sufficient contextual uncertainty, a historical data set is likely to be consistent with multiple contradictory explanations. To make decisions, it is necessary to define a selection criterion by which to choose a 'best' hypothesis or model (the model most likely to generalize into knowledge). These selection criteria also take the form of probabilistic estimates of evidence, requiring context of their own.

### Reproducibility in scientific literature

Experimental scientists, particularly in theory-poor and hypothesis-rich fields such as biology and psychology, have learned the hard way what happens when inductive logic is automated and scaled. Currently, most published, peer-reviewed studies in these fields describe a result that would not occur again if the experiment were repeated (Baker 2016; Ioannidis 2005). This problem has become known as the 'reproducibility crisis*'*. Reproducibility is a core tenet of the scientific method; excessive irreproducibility undermines the credibility and minimizes the impact of the output of scientific endeavours.

The reproducibility crisis is often attributed to the misuse of statistical hypothesis testing (Nuzzo 2014), but is better understood as an unavoidable outcome of scaling induction (search and selection of evidence). Briefly, a hypothesis test is an algorithm that calculates the probability that the differences between the experimental and control groups would have occurred if the experimental modulation had no effect (the null hypothesis). A common output metric is the 'p-value'. If a p-value is sufficiently small, the null hypothesis is rejected. The experimental result is considered statistically significant. Statistical significance has become the default selection criterion at which an experiment is published, thereby making its way into scientific literature.

The problem emerges as researchers try multiple similar experiments. A statistically significant result is unlikely to occur in one experiment, but is likely to occur eventually if an experiment is repeated. The scale required to produce a false-positive is smaller than one might imagine. A common scientific threshold of statistical significance is $p<0.05$ (less than 5% chance of occurring due to chance alone); using this threshold, the number of experiments needed to create where a scientist is more likely than not to observe a false-positive result is on the order of twenty experiments.[1] In practice, a scientist could see this false-positive result in an early iteration and perceive it as strong evidence of a true effect.

Scientists can exacerbate the problem by using search-based strategies within their own research process. This is known as p-hacking or data dredging. Scientists are incentivized to seek out unexpected anomalies and patterns. In fact, a scientific career is considered successful only if a scientist publishes statistically significant results regularly and repeatedly. There is external pressure to design a research process that maximizes the likelihood of finding a statistically significant result with minimal time or effort. For example, a researcher could design a method to screen hundreds of chemicals as drug candidates at the same time, increasing the likelihood that one or a few will have a statistically significant result. A researcher could collect many covariates and test every combination of covariates for combinatorial effects on an experimental outcome, performing thousands of hypothesis tests either explicitly or implicitly. Counter-intuitively, while a scientist is hired to run experiments, the more productive (by number of experiments) a researcher is when attempting to support a given hypothesis, the less reliable the evidence generated by any one experiment. At the extremes, data dredging methods can be used to support nearly any conclusion: including arguing that the mind of a dead salmon can be read using fMRI data (Bennett, Wolford, and Miller 2009) or that people can age in reverse become younger by listening to a Beatles' song (Simmons, Nelson, and Simonsohn 2011).

The reproducibility crisis is a tangible demonstration of the limits of induction. In fact, the degradation of the quality of scientific literature was predictable from an understanding of statistical inference and scale alone. In 2005, Dr. John Ioannidis of Stanford University demonstrated using a Bayesian framework that scientific literature would become more unreliable over time as many scientists repeatedly tested similar, but ultimately incorrect, hypotheses (searching) and only reported significant results (selection) (Ioannidis 2005). Importantly, an individual scientist may not be aware that the same experiment was performed in the past, is currently being performed in other laboratories, or that an analogous experiment was performed using other methods. Yet this invisible context is critical to accurately assess the quality of the probabilistic evidence provided by their study (Ioannidis 2005; Nuzzo 2014; Simmons, Nelson, and Simonsohn 2011). As such, the reproducibility crisis is a problem observed not by individual scientists or within a single study, but upon examination of the complete body of scientific literature or from the perspective of a population [2]. Unsurprisingly, pharmaceutical companies that depend on academic research to identify drug candidates were one of the first to quantify the magnitude of the reproducibility crisis in biology. Bayer reported less than 30% of attempts to reproduce findings resulted in successful replication (Prinz, Schlange, and Asadullah 2011). Amgen reported less than 11% (Begley and Ellis 2012).

**Saving machine learning from p-hacking**

Machine learning is not foundationally different from hypothesis testing. While p-values have been replaced with other probabilistic metrics of evidence, most machine learning models are still well approximated by the model of 'many hypothesis tests performed simultaneously'. The training phases of machine learning are highly iterative, relying on the same methods of statistical inference used in all types of induction (search). A hypothesis (candidate model) is generated, tested, updated and tested again until some stopping condition is met (selected). The model that most closely aligns with a previously chosen metric of success is selected, much in the same way that a particularly successful experiment is selected for publication based on the success of the experiment. The practice of running many iterative analyses on the same data and choosing the model that performs 'best' is indistinguishable from p-hacking, except that most of the steps are performed by a computer. In fact, due to automation, it occurs faster and more obviously than in traditional science experimentation. Rather than using the term p-hacking, the machine learning terminology is 'overfitting'. Overfitting is evidence of having too many degrees of freedom, considering through too many potential hypotheses (searching) to find (select) a model that appears to perform well.

Much of the manual work that goes into generating a machine learning algorithm is focused on mitigating the damage done by scaling induction. Cross validation (separating into training and testing) mimics the 'one hypothesis' to 'one experiment' standards of the ideal scientific method. Regularisation (penalising complex explanations) is meant to limit the number of models an algorithm implicitly considers, thereby reducing the amount of data dredging. Yet, just like well-meaning scientists seeking statistically significant results for their research projects, data scientists may break these protections using excessive search and selection in their own workflow. Repeated training, testing, training, and testing will create models that appear to work (perform better than chance), but are fitting noise. Much like publication of scientific experiments, the incentive structure around an individual data scientist's performance is often not aligned with success of a data science initiative overall.

If not already there, automated intelligence and machine learning will develop a reproducibility crisis of its own. Early research wins and models announced with much publicity will not generalize and eventually fail. Businesses will perceive their data science teams as underperforming, or not worth the investment. Practitioners and strategic leaders would benefit from understanding the limits of inference. Models built based on strong theoretical foundations (existing knowledge, context), based on rules that have already shown substantial predictive value, will outperform models developed largely by inference, based on excessive search and selection.

## Notes

***** Clare Ann Gollnick is a practicing data scientist, technologist and entrepreneur based in New York city.

[1] Author acknowledges an over-simplification. The exact number depends on number experimental variables such as variability within the population but often falls on this order of magnitude.

[2] An often-proposed solution to the reproducibility crisis is to publish all experiments regardless of outcome (negative or positive). This proposal solves a problem of selection, but also changes the nature and intent of scientific literature. Scientific literature would no longer represent a body of knowledge, but a public record of experiments. As such, it only pushes the problem of scaling induction to a later stage of the scientific inference process.

## References

Baker, Monya. 2016. "1,500 Scientists Lift the Lid on Reproducibility." Nature 533 (7604): 452–454. doi:10.1038/533452a.

Begley, C. Glenn, and Lee M. Ellis. 2012. "Raise Standards for Preclinical Cancer Research." Nature 483 (7391): 531–33. doi:10.1038/483531a.

Bennett, Craig M., George L. Wolford, and Michael B. Miller. 2009. "The Principled Control of False Positives in Neuroimaging." Social Cognitive and Affective Neuroscience 4 (4): 417–22. doi:10.1093/scan/nsp053.

Ioannidis, John P A. 2005. "Why Most Published Research Findings Are False." PLoS Medicine 2 (8): e124. doi:10.1371/journal.pmed.0020124.

Nuzzo, Regina. 2014. "Scientific Method: Statistical Errors." Nature 506 (7487): 150–52. doi:10.1038/506150a.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" Nature Reviews. Drug Discovery 10 (9). Nature Publishing Group: 712. doi:10.1038/nrd3439-c1.

Simmons, Joseph P., Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." Psychological Science 22 (11): 1359–66. doi:10.1177/0956797611417632.