Scientific Data Quality as Contrasted with Other Sources of Uncertainty in Accurately Describing Nature

D.L. Marrin[1]

[1]Water Sciences & Insights

November 21, 2022

Abstract

Data quality is only one of many uncertainties involved in our attempt to understand, model and predict complex nonlinear systems and to identify the properties emerging from them. Perhaps not surprisingly, less than half of published scientific studies can be successfully reproduced, with the earth sciences occupying an intermediate position among the major disciplines. The uncertainty of data quality is only one factor that contributes to the relatively poor reproducibility of research studies and the resulting hypotheses. Potentially more important factors include unsuitably low thresholds for assessing statistical relevance, inappropriate data manipulation, inadequate research design, and outright fraud. Data quality for earth science data can be communicated via measurement errors (e.g., instrument accuracy, technician practices) and addressed by employing measurement quality codes, whereas the other aforementioned factors can be identified or controlled through rigor, methods selection and research competency. A more difficult source of uncertainty includes biases, assumptions, and environmental or social influences that affect the judgment and perceptions of scientists themselves. The difficulty is that many of these factors operate beyond one's awareness or control, residing within automatic processes of the human brain. From continually seeking, interpreting and projecting patterns (spatial and temporal) to confabulating answers and relying on heuristics, we are largely at the mercy of what the brain has evolved to do, which is not necessarily to accurately perceive the natural world. Peer reviews, precision instruments, mathematical abstractions and digital computers certainly assist us in probing nature; nevertheless, we ultimately perceive the world as we are, rather than as it is. Consequently, a scientifically "objective" view of nature is being questioned in research ranging from quantum mechanics to human consciousness. To what extent are the peculiarities of human brain responsible for the uncertainty and nuances that we believe exist in nature? Are scientists fooled by a brain that constructs, rather than simply observes, the world around us? If so, do we have the tools to deal with this kind of uncertainty?

Scientific Data Quality as Contrasted with Other Sources of Uncertainty in Accurately Describing Nature

D.L. Marrin

Water Sciences & Insights

CONTROLLABLE FACTORS

Data quality is only one of many uncertainties involved in our attempt to understand, model and predict complex nonlinear systems and to identify the properties emerging from them. Perhaps not surprisingly, less than half of all published scientific studies can be successfully reproduced, with the earth sciences occupying an intermediate position among the major disciplines [1]. Uncertain data quality is only one factor that contributes to the relatively poor reproducibility of research studies.

The uncertainty of data quality may be due to unsuitably low thresholds for assessing statistical significance, inappropriate manipulation of data sets, and inadequate research design. Data quality for earth science is communicated via measurement errors (instrument accuracy, technician practices) and could be improved by employing measurement quality codes [2]. Other sources of uncertainty are identified or controlled via scientific rigor, method selection and research competency.

Aspects of Scientific Uncertainty

Lacking knowledge of the value of quantities or their relationships.

- Defining and Evaluating Data: parameter selection; reliability of numbers; data quality; measurement errors
- Interpreting Complex Data: scale-dependent, spatial and temporal factors; reliance on heuristics; nonlinearity
- Standard Scientific Perspectives: assumptions; conceptual models; applicability to real world conditions
- Inherent Limitations of Scientists: biases; beliefs; prejudices; unconscious search for patterns and causes

Example: Measuring Water Quality

- Uncertainty lies in knowledge (inadequate site understanding) and measurements (variability)
- Measurement uncertainty is due to sampling, analytical and heterogeneity factors:

#1 Heterogeneity of sample matrix and distribution of target analytes within the matrix

#2 Variations in sampling methods, equipment or procedures

#3 Sample preparation and lab techniques/instrumentation

Addressing Measurement Uncertainty

- Reduce or Limit Uncertainty
  - Select optimal sampling/preparation methods for matrix
  - Increase size/number of samples and mix thoroughly
  - Maintain consistency in sampling procedures/equipment
- Communicate or Quantify Uncertainty
  - Precision addressed by replicate sample collection and measurement, followed by statistical analysis
  - Accuracy addressed by method proficiency or comparison
  - Expressed as a standard deviation or combined uncertainty
  - Result plus uncertainty lies above/below compliance limit

LESS CONTROLLABLE FACTORS

A more difficult source of uncertainty includes the biases, assumptions, and environmental or social influences that affect the judgment and perceptions of scientists themselves. The difficulty is that many of these factors operate beyond one's awareness or control, residing within automatic processes of the human brain. Peer reviews, logic analyses, mathematical proofs and digital computers undoubtedly assist us in probing nature. Nonetheless, these kinds of abstractions may actually lead us further from the world we seek to comprehend as we continually refine our models, concepts and ideas about it.

Addressing Data Interpretation Uncertainty

- Reduce or Limit Uncertainty
  - Describe temporal patterns of relevant parameters
  - Map spatial distribution of parameters in critical areas
  - Identify major mechanisms affecting observed processes
  - Collect duplicate samples from a subset of locations
- Communicate or Quantify Uncertainty
  - Uncertainty: interpretation > sampling > measurement
  - Identifiable vs. unknown sources of variability/anomalies
  - Compare data for different scale-dependent scenarios
  - Contrast data interpretation for different data sets

UNCONTROLLABLE FACTORS

Anais Nin noted that we don't see things as they are, but as we are. Accordingly, a scientifically "objective" view of nature is being questioned in research ranging from quantum mechanics to consciousness. To what extent are the nuances of the human brain responsible for the uncertainty and anomalies that we believe exist in nature? Are scientists often fooled by a brain that constructs, rather than simply observes, the world around us? If so, do we have the tools to deal with this kind of uncertainty? While we can become more aware of our propensity to generate such uncertainty, can we somehow compensate for it?

Unconscious and Automatic Brain Processes

- Pattern Seeking, Interpreting & Projecting: the primary activity of the brain in recognizing, understanding and making predictions about reality (evolutionary-based)
- Simplifying & Separating: often unable to recognize complex nonlinear patterns, reducing them to component subpatterns yields linearity via oversimplification
- Guessing & Confabulating: perception involves as much guessing as observing, and explanations are based largely on post hoc inferences about cause-and effect
- Maintaining Consistency & Continuity: strong bias to interpret similar experiences in consistent ways and to infer causality in preserving continuity of knowledge

AVAILABLE TOOLS

It has been observed that science is a race between inventing ways to fool and not to fool ourselves. Actions that minimize our self-deception include the avoidance of: collecting or interpreting data to support a single hypothesis, selecting the most desired or interesting data, checking unexpected results more than expected ones, devising post hoc theories to explain results, excluding collaboration with rival or adversary labs, restricting scientists in related fields to initially review data, and discouraging journals from accepting papers before research is conducted [3]. Whereas these precautions may reduce some of the uncertainty related to self-deception, our brains have an evolutionary bias for heuristics over accuracy.

REFERENCES

[1] M. Baker. Is there a reproducibility crisis? *Nature*, 533, p. 452 (2016).

[2] R. Cumming. Measurement uncertainty and the development of data quality codes for hydrological data. *Proceedings of 2012 AHA Conference*, Melbourne, Australia, 39 pp.

[3] R. Nuzzo. Fooling ourselves. *Nature*, 526, p. 182 (2015).

CONTACT

D.L.'West' Marrin, Ph.D.

Water Sciences & Insights

POB 232125, Encinitas, California 92023

marrin@watersciences.org