1. **Title:** Misadventures with outliers
2. **Corresponding Author:** David Linnard Wheeler1,2

**Affiliations:**

1 University of California, Berkeley, School of Information, Berkeley, CA

2 Washington State University, Department of Plant Pathology, Pullman, WA

1. **Keywords:** outliers, anomalies, value-laden judgements, ethics
2. **Executive summary**

Data scientists have long been confronted with decisions about outliers. What are outliers? Where do they come from? How should we define, detect, and deal with them?

These questions have occupied the minds of many, at least since Bernoulliexpressed his frustration in 1777: “I see no way of drawing a dividing line between those [observations] that are to be utterly rejected and those that are to be wholly retained” (Bernoulli, 1777).

Here I argue that the answers to these questions require value-laden judgements.

To motivate this topic, I first expose the pernicious consequences that can arise from our misadventures with outliers. For this, I review several consequential historical examples. Cases of marital litigation, space shuttle disasters, TSA searches, ozone holes, and US Census data are discussed. The common denominator for all of these cases is that they hinge on value-laden decisions. At face value these decisions can seem objective and even trivial. Consequences of these decisions, however, are non-trivial. They can trespass ethical norms and stifle scientific advancement.

So, what *should* data scientists do to avoid these quandaries? Several reflexive responses to this question are presented along with respective objections. To help navigate this landscape, I recast recommendations from philosophers and statisticians across time.

The confluence of these insights converge on several pragmatic solutions for data scientists. Solutions include increasing awareness and sensitizing ourselves to the consequences we confront when interacting with outliers; reframing the obligations incumbent upon data scientists as opportunities to learn something about the world; invoking the Golden Rule to remind ourselves that we should *treat* outliers as we would expect to be treated if we *were* outliers; maintaining transparency with stakeholders about the decisions behind our treatment of outliers. Equipped with these solutions…

For example, we can start by increasing awareness about the consequences of the ways he define, detect, and deal with outliers. Review of historical cases and negative visualization, a practice popularized by Seneca and other stoic philosophers, can help sensitize data scientists to the minefields we traverse when analyzing data. From this discussion, data scientists might feel trepidation when embarking on an analysis. To assuage this fear, animate curiosity, and instill a sense of duty, we can also reframe these obligations incumbent upon data scientists as opportunities to learn something interesting about the data in question. We should invoke the Golden rule: we *should* treat outliers as we would expect to be treated if we were outliers. Lastly the logic used to arrive at decisions about outliers should like other practices in data science be transparent to the stakeholders.

1. **References**
2. Bernoulli, D., (1777). The most probably choice between several discrepant observations and the formation therefore of the most likely induction. In C.G. Allen (1961), *Biometrika*, 48(1-2), 3-18. <https://doi.org/10.1093/biomet/48.1-2.3>