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, data ethics, value-laden
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 Bernoulli (1777)expressed 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”.

Here I argue that the answers to these questions demand 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 gone awry, holes in the ozone, and US Census data are discussed. The common denominator for all of these cases is their entailment with upstream 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 resolve these ethical puzzles? 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 methods for sensitizing ourselves to the consequences that can follow from our interactions with outliers; reframing the obligations incumbent upon data scientists as opportunities to learn something about the world through data; 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 about the decisions behind our treatment of outliers. Equipped with these tools and the wealth of technical aids available, I hope to help prepare data scientists for the unforeseeable challenges and opportunities that arise when learning from data.

Throughout this essay I argue that data scientists cannot make decisions about outliers without recourse to values. Even when we *outsource* these decisions to technical aids or precedence, values come along for the ride. This is not a problem per se. The problems arise when we confuse value-laden decisions for value-neutral ones. Fortunately, solutions to this problem are abound. Like other problems, the first step is to accept that we have a problem.

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>