1. **Title:** Outliers: obligations and opportunities
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 **Bernoulli** 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 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, I argue, 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, 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. So too are objections to each proposed solution. To navigate this landscape where land mines await every decision, I recast recommendations from philosophers and statisticians across time. The confluence of these insights converge on several pragmatic solutions for data scientists.

The forces that animate these decisions about outliers are a special case of a more fundamental problem first identified by philosophers - that reality is how it appears.

Finally, I summarize cases where inclusion and exclusion of outliers from data sets can introduce biases and conflicts between our fiduciary and moral responsibilities. From these premises, I argue that confrontation with outliers can challenge ethical principles that are not always obvious and demand critical examination, caution, and actions that may be at odds with near-term analytical duties. It is therefore incumbent upon us to be explicit about the value-laden decisions we use to navigate encounters with outliers and balance both analytical and moral obligations. To satisfy these responsibilities, recommendations are offered.

**References**

1. 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*, 48:3-13.