Outliers: obligations and opportunities

David Linnard Wheeler

W231: Final Project

I. Summary

The objective of this essay is to show that attempts to define, detect, and deal with outliers require value-laden decisions. These decisions, in turn, entail consequences that can challenge ethical principles. To motivate this topic I will expose the subtle and pernicious consequences that arise from different definitions, detection, and handling procedures. To start, I will show that the different conceptions we use to categorize outliers arouse controversies that are hard to predict. Next I will describe the different sources of outliers and the mirages encountered when ascribing provenance. To underscore the imperatives introduced when confronted with outliers, I will show that the hubris and or expedience that often animates outlier omission is a special case of a more fundamental problem - that reality is how it appears. Further, I will describe how systematic outlier omission can snowball and stifle scientific advancements by suppression of anomalies - the things that, as Kuhn argued, often precipitate paradigms shifts. Finally, I will show that both inclusion and exclusion of outliers from data sets can introduce biases and conflicts between our fiduciary and moral responsibilities. From these points, I will conclude 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 analysts to be explicit about the value-laden decisions they use to navigate

encounters with outliers and balance both analytical and moral obligations. To satisfy these responsibilities, recommendations are offered.

II. Introduction:

Outliers arouse a diversity of emotions in people who analyze and interpret data. They are a source of fear, anxiety, suspicion, or even excitement. They can introduce bias, alter conclusions but also signify rare and important events (Anguinis et al. 2013). They are both the objects of interest and despair. Both noise and signal.

Emotional responses to outliers are likely as old as empirical inference itself. The first instances of “outlier” and its synonym “anomaly” date back at least as early as the 1500s in Google’s ngram viewer corpus (https://books.google.com/ngrams) (**Figure 1**).

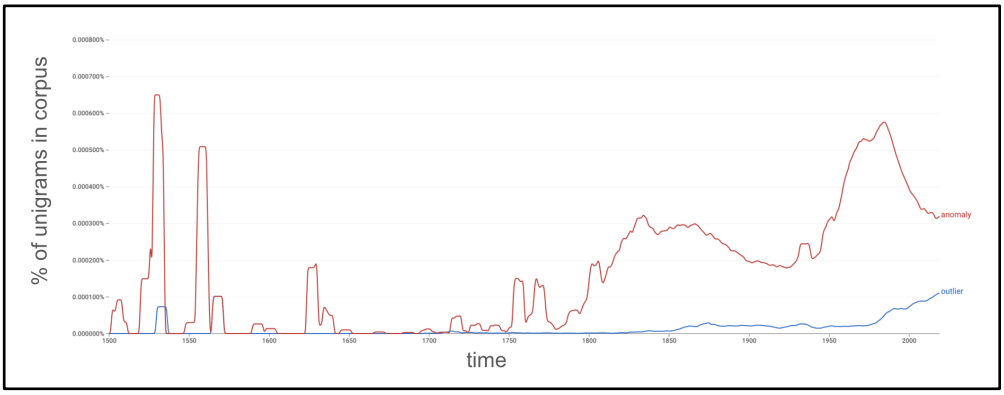

Figure 1. The percentage of instances of “outlier” (blue) and “anomaly” (red) in Google’s ngram viewer corpus over time.

In the earliest found written mention of outliers in the scientific literature, Bernoulli (1777) expressed a sense of frustration with the practice of outlier removal without a priori justification: “But is it right to hold that the several observations are of the same weight or moment, or equally prone to any and every error? Are errors of some degrees as easy to make as others of as many minutes? Is there everywhere the same probability? Such an

assertion would be quite absurd,... I think each and every observation should be admitted whatever its quality."

Further, Bernoulli (1777) sympathized with the rejection of outliers when observers encounter justifiable reasons before inspecting the actual data: "I do not condemn in every case the principle of rejecting one or other of the observations, indeed I approve it, whenever in the course of observation an accident occurs which itself raises an immediate scruple in the mind of the observer, before he has considered the event and compared it with the other observations." Thus, since the first published record of people wrestling with outliers, context seems to govern their valence. In some contexts, we welcome outliers as harbingers of fraud we want to detect. In other contexts, we disdain them as obstacles to overcome. In all cases, our feelings about outliers are conditional on our objectives. Outliers are contextual. Their definitions, identification, and treatment are all dependent on other observations (can outliers exist when n=1?) as well as the goals and capabilities of the analysts.

Thus, outliers are a contentious and potentially contested topic. The controversy likely dates back at least to Bernoulli in 1777 but shows up extensively in the 1800s. At one end of the spectrum are the "rejectors" who favored the rejection of outliers (Legendre, 1805). At the other end of the spectrum are the "retainers" who favored the retention of outliers (Bessel and Bauer, 1838). In between these two poles is a rich array of nuanced options.

What is the *right* thing to do?

This essay will not answer this question. Instead I will argue that our conceptions, detection strategies, and treatment of outliers require value-laden decisions that have

unforeseen consequences. Decisions about outliers are thus often fraught with unforeseen repercussions. My goal is just to expose the nontrivial reverberations that arise from seemingly trivial analytical decisions.

To begin, I will first show that attempts to define outliers require value-laden decisions. Further, these decisions have real world consequences.

III. What is an outlier?

Outliers, anomalies, aberrants, contaminants, discordant observations, and stragglers have all been used, sometimes interchangeably, to describe more or less the same thing: observations that are markedly different from other observations.

Over time, many have tried to define outliers. Some of these definitions are ambiguous and therefore difficult to operationalize. For example, Grubbs (1969) defines an outlier as an observation that "appears to deviate markedly from other members of the sample in which it occurs." Other definitions, for example one from Anscombe and Guttman (1960), are more technical, precise, and easy to operationalize: "[a]n observation with an abnormally large residual…". Notice, however, that this conception of outliers fails to include categorical data. Still other conceptions, like this one from the Dorland's Illustrated Medical Dictionary, define outliers with brazen and prescriptive language: "an observation so distant from the central mass of data that it is considered an obvious mistake that should be removed from the data whether or not a cause of deviation can be found" (Anderson et al. 1998).

And these conceptions only capture univariate numerical data types (**Figure 2**). Categorical outliers and those in multivariate spaces have received less attention. For the former case, outliers are generally conceived of as "rare data objects" (Pang et al. 2016;

Suri and Athithan, 2019) or data objects with frequency occurrences that are "exceptionally typical or un-typical within the distribution of frequencies occurrences of any other attribute value" (Angiulli et al. 2020).

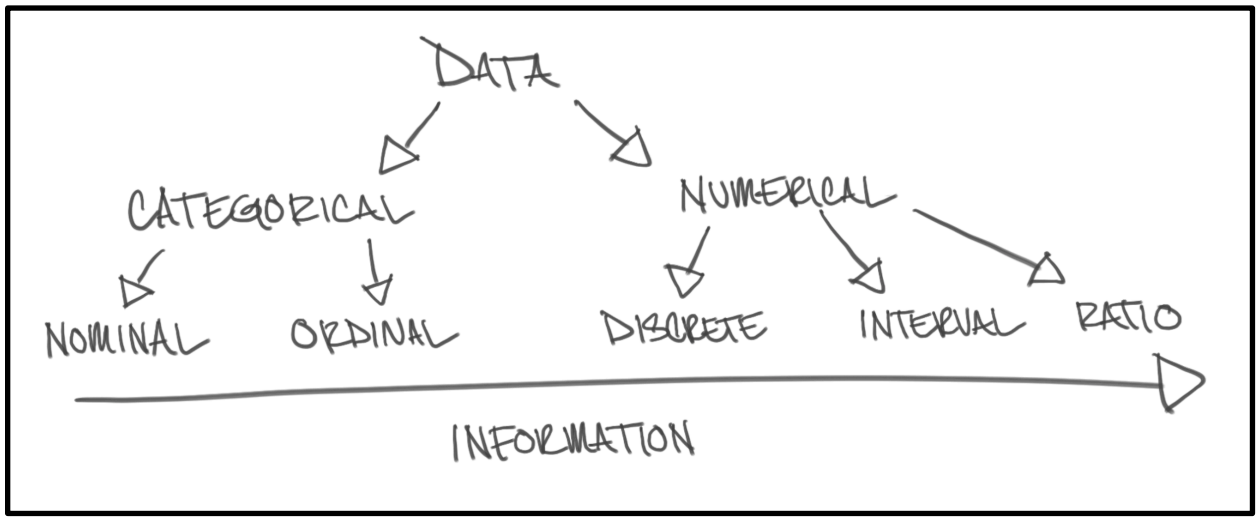

Figure 2. Types of data arranged along a continuum of information richness.

Hence, there is a diversity of conceptions of outliers. To be more precise, a literature review by Anguinis et al 2013 identified 14 mutually exclusive definitions of outliers. Similarly, a small (n=23), non-random, survey of Master of Information and Data Science (MIDS) students at Berkeley revealed conflicting conceptions of outliers (**supplemental document 1**).

The variability present in different conceptions of outliers may just seem like semantics, however; much more is at stake. Categorization of outliers and non-outliers is an essential value-laden judgement. Just as Kraemer et al. (2010) argued that algorithms often require value-laden decisions, one cannot demarcate outliers from non-outliers *a priori* without encountering ethical issues.

III.I Outlier Definitions Require Value-Laden Decisions

For example, take the case of *Hadlum* v. *Hadlum* (Barnett, 1978). Mr. Hadlum left home for military service in August of 1944. His wife, Mrs. Hadlum delivered a baby nearly a year later, 349 days after Mr. Hadlum left home. Upon his return, Mr. Hadlum filed for divorce. He argued that, since the baby was born weeks after the average gestation period of 280 days, Mrs. Hadlum *must* have committed adultery. Is Mr. Hadlum right? Did Mrs. Hadlum deliver another man's child? To answer this question we must define the gestation time of 349 days as an outlier or not (and assume no infidelity before Mr. Hadlum left for military service). To do this is to make a value-laden judgement, fraught with marital consequences.

Moreover, to argue for or against Mr. Hadlum's claim of adultery is to claim that we know how the data *should* be distributed. That is, to claim that the 349 day gestation period is an outlier is to claim that no babies can possibly be born beyond 348 days. Conversely, to claim it does not constitute an outlier is to claim that babies can be born beyond a 348 day gestation period. Each case invokes the problem with induction, as discussed by Hume (1779) and others. Just because most previous human gestations are less than 349 days does not mean that all future gestation periods are less than 349 days. In either case the stakes are high - one could make or break the marriage depending on one's definition of an outlier. The court ruled in favor of Mrs. Hadlum. The gestation period, they decided, was unlikely but biologically possible. The marriage was preserved.

Cases like *Hadlum v. Hadlum* underscore the primacy of outlier classification. This case shows that implicit but operational definitions about outliers (the courts later defined outlying gestation periods as those longer than 360 days (Barnett, 1978)) require value-

laden decisions. Moreover, this case illustrates how outlier classification systems conjure ethical issues. These issues, in turn, challenge the principles expressed by the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978) and the Menlo Report (Dittrich and Kenneally, 2012). For example, if we imagine that Mrs. Hadlum was unfaithful to Mr. Hadlum and still found not-guilty, justice, à la The Belmont Report, would not be served. Alternatively, even if Mrs. Hadlum was faithful, it is hard to imagine that their marriage was the same after the failed divorce case. In either case the distribution of benefits and burdens seems far from just. Moreover, the court also seemed to challenge the "Respect for Law and Public Interest" criteria proposed by (Dittrich and Kenneally, 2012). Although they ruled in favor of Mrs. Hadlum they were not, as far as I can tell, "transparent in [their] methods and results" used to arrive at their conclusion on innocence. But even if they were, who *is* a court to tell anyone that their child is the result of adultery just because of a "unlikely" gestation period?

The Hadlum case is not the only example of a case where our definitions of outliers matter and trespass ethical norms. Examples of ethical issues that follow from attempts to define numerical and categorical outliers are abound. Consider the case, for example, of the biased facial analysis softwares discovered by Joy Buolamwini (Hardesty, 2018). One putative source of the bias in these softwares is the dearth or absence of underrepresented minorities in the training sets (Hardesty, 2018). How does this relate to definitions of outliers? If truly a source of bias, then the failure to include certain demographics, for example black women, is tantamount to defining these demographics as outliers, even if implicitly. Again, if the absence of minority demographics truly

constitutes a source of bias, then the programmers who designed the algorithms implicitly drew a line in sand. All people beyond the line *are* outliers. The choice to include or exclude certain demographics from training sets for algorithms again constitutes a value-laden decision. The consequences, like the case of *Hadlum v. Hadlum*, include infringements on both beneficence and justice, as conceived by the Belmont Report.

The above cases illustrate that outlier definitions demand value-laden decisions. These decisions, in turn, can yield unforeseen consequences that violate ethical principles. This problem requires a solution. One solution could be to define outliers in theoretical terms that are universal across data types and operational terms within data types (**Figure 1**). This latter task, however, is not trivial. Given the diversity of data types and definitions of outliers in each, this task is beyond the scope of this project. For to operationalize outlier definitions for each type of data is to recast theoretical definitions as mathematical expressions that are deployable by the masses. Thus, for now I will just submit a candidate theoretical definition from (Barnett and Lewis, 1994): "we shall define an outlier in a set of data to be an observation (or a series of observations) which appears to be inconsistent with the remainder of that set of data."

III.II Application of Bowker and Stars Framework for Classification Systems

To test the performance of this definition in classifying outliers, I will now apply Bowker and Star's (1999) framework for classification systems. If the definition above furnishes comparability, visibility, and control across data types and domains then perhaps it will serve as a valuable classification system.

For a classification system to be comparable it must enable and facilitate communication, by regularization of semantics, across entities (Bowker and Star, 1999).

Thus the candidate definition must serve all data types and domains. From a bird's eye view, this definition seems to serve all data types. That is, it is easy to imagine instances of both categorical and numerical data that are inconsistent with the rest of the data. For example, imagine a dataset of people with eye color and height. From these data, categorical (nominal) outliers might constitute subjects with abnormal eye colors (e.g. black). Similarly, numerical (ratio) outliers might constitute subjects with abnormal or impossible heights (e.g. -2 feet or 11 feet).

So far so good, right? Maybe not. The fidelity of this candidate definition to outliers from all data types is likely a mirage. Upon closer inspection we can see that this definition fails to capture all categorical outliers, for example. Although it catches local cases of unexpected categories, like black eye color, it does not capture cases where the frequency occurrences of expected categories are anomalous. For instance, if we collected eye color data from an island populated with almost entirely brown eyed people but a fraction of blue eyed people, this definition would not catch the blue eyed subjects even though the frequency occurrences of these subjects might be anomalous. They would be "invisible" or residual. This brings us to the next criterion: visibility.

For categories within a classification system to be visible, they must be classifiable (Bowker and Star, (1999). To be invisible is to be unclassifiable or residual. Thus our candidate definition provides visibility for those outliers that are discernible within its scope. Some outliers, like those that differ not necessarily in the distance from other observations but in their frequency occurrences, will slip under the radar. They will be "invisible".

Finally, we have control. Objects of a classification system, like outliers, are subject to more or less control by those who analyze data. The candidate theoretical definition presented above does not exercise much control over instances of outliers in each data type or domain. That is, the classification system itself is not constrained by this definition. The onus is on the analyst to decide and justify what exactly it means for observations to be inconsistent. This is both good and bad. The ambiguity of the definition endows each analyst with the license to exercise their own due diligence. The cost, however, is of comparibility and visibility. It is possible, for example, for some analysts to implement a localized incarnation of the candidate definition that is not comparable across data types or domains and or invisible to consumers of their work.

In summary of this section about outlier definitions, I hope the following is clear. A diversity of definitions of outliers are present in the literature and among data scientists within the MIDS program at UC Berkeley. To define outliers is to erect classification systems that demarcate outliers from non-outliers. This act requires value-laden judgements that are often implicit. In turn, these value-laden decisions have ethical consequences. The severity of the ethical consequences can be gauged to some extent by the Belmont Report. For example, the cases of *Hadlum v Hadlum* and facial analysis softwares show us that, from our conceptions of outliers and the requisite value-laden judgments that animate them, arise consequences that are not obvious when the definitions are first conceived. These consequences can violate notions of justice and beneficence, as conceived by the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978). A theoretical candidate definition, formalized by Barnett and Lewis (1994), was presented.

This definition appears to enable comparability across most data types. However, this comparability property, as predicted by Bowker and Star (1999), comes at a cost of control.

Before I show that detection of outliers also requires value-laden decisions, I will briefly describe the origin of outliers.

IV. Where do outliers come from?

Outliers arise from several sources. Most authors who write about such topics (e.g. Anscombe 1960; Barnett, 1978; Grubbs 1969; Smiti, 2020) agree that outliers can be attributed to:

1. Inherent variability
2. Execution error
3. Measurement error

Inherent variability is just the variability "that would be observed in the population even if all measurements were perfectly accurate" (Anscombe 1960). This source of variability cannot be modified without intervention of the underlying population from which the samples were collected. An example of inherent variability includes Mrs. Hadlum's child that was born after a 349 day gestation period. In many cases, the parameters about the underlying population are fixed and unknown without exhaustive censuses.

Execution errors are introduced when there is a "discrepancy between what we intend to do and what is actually done, other than error in the use of measuring instrument" (Anscombe 1960). Some authors further parse execution errors into sub-categories containing (i) sampling error and (ii) mis-reporting errors (Smiti, 2020). For the sake of precedence, parsimony, and Bowker and Star's (1999) comparibility criterion, I

clump these sub-categories into the larger category of execution errors. Execution errors are absent in a world free from mistakes. An execution error would occur if they unintentionally asked another person, not Mrs. Hadlum, for their gestation period. To determine if an outlier arose from an execution error, we would likely need an exhaustive and searchable record of the procedures that proceeded data collection. For most researchers, this information is likely not available.

Errors that arise from “measuring instruments” are measurement errors (Anscombe 1960). If the things with which we collected data were precise and accurate, then measurement errors would not exist. For example, if Mrs. Hadlum’s gestation period was measured by an infallible tool that reflected her true gestation period at fine scale resolution, no measurement error would have occurred. Measurement errors are likely present in most data collection environments. They become outliers recognizable by the candidate definition above and detectable by the methods discussed below when they breach above baseline errors associated with other observations. Like execution errors, it can be exceedingly difficult to attribute an outlier to measurement error unless scrupulous records are preserved, searched, and cross-referenced.

The provenances of outliers as described above does not appear to be controversial. Most authors who write about the subject agree on the mechanisms by which outliers arise. The *attribution* of provenance, however, can be controversial.

How do we know the true source of the outliers? The question becomes harder to answer with less and less information. That is the accuracy and indeed our confidence about the answer to this question *should* be inversely proportional to the availability of information with which we can attribute provenance. In practice, answering this question

can be next to impossible. One can compile records to support hypotheses about provenance. However, such evidence is often a limiting resource. Moreover the absence of such evidence is not the evidence of absence of provenance (i.e. *argumentum ad ignorantiam*). And this is just the start of it.

To attribute provenance to outliers is again to confront value-laden decisions. Most examples of such value-laden decisions seem to involve accusations directed towards individuals on the basis of the assumption that the outlier(s) in question have arisen from inherent variability. Below are a couple familiar examples.

For example, take the familiar case of *Hadlum v. Hadlum* (Barnett 1978). The source of controversy here was not just the definition of an outlier but also the provenance of the outlier. In accusing Mrs. Hadlum of infidelity Mr. Hadlum assumed the outlying gestation period was due to inherent variability. This is a value-laden decision because, in making this assumption, he confronted an ethical issue: infidelity.

Other examples, where value-laden decisions are embedded in the attribution of provenance to outliers, are available in other domains. Another example is from Costanza-Chock's (2018) account of their experience with TSA agents after being flagged as anomalous by the millimeter wave scanner used in Whole Body Imaging (WBI). Constanza-Chock is a "nonbinary, transgender, femme presenting person" Costanza-Chock's (2018). While in TSA, their groin was flagged by the millimeter wave scanner as anomalous. Since millimeter wave scanners are designed to detect concealed objects the anomalous label warrants completion of further security protocols. Hence, they were then subjected to further searches by a TSA agent.

This encounter required layers of value-laden decisions, some of which will be discussed further below. For now, I want to emphasize the value-laden decision entailed by the accusation of concealment. The accusation assumed that the object(s) that triggered the anomalous label were due to inherent variability. In this case, the assumption is likely valid. Nonetheless, the attribution of provenance to the outlier(s) here involved value-laden decisions because one could not make such a decision without introducing one's values.

Lastly, these value-laden decisions illicit ethical implications. In the first case of *Hadlum v Hadlum*, it is hard to see how justice, as conceived by the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978), is satisfied by assuming that the outlying gestation period is due to inherent variability. To be just the burdens and accusations that arise should be uniformly distributed across all candidate sources of outliers, not just inherent variability. To use Bayesian language, our prior for the putative provenance of the outlier in question should be uniform until and unless evidence to the contrary is discovered. Deviations from the uniform prior are not just because, in this case, they result in accusations of infidelity that might also be explained by other sources.

Likewise, for the TSA agents to be just to Constanza-Chock (2018) they should have assigned tentative blame to each potential source of outliers before accusing Constanza-Chock of something that could have been explained by other phenomena. To explore all explanations for an anomalous label is different from first accusing someone of something without due diligence.

The cases above appear to represent a convergence of Bayesian, deontological, and consequentialists arguments. If we accept the claim that both Mrs. Hadlum and Constanza-Chock were harmed by injustice then each of these frameworks arrived at the same destination. This convergence of disparate philosophies to the same conclusion might create the illusion that they are all equally *right.* Since we cannot predict outcomes *a priori*, however, it follows that the conservative framework - the one that minimizes harms by exploration of all possible sources of outliers and the value-laden decisions they entail - is the Bayesian-deontological hybrid.

Up until now I have shown that value-laden decisions are implicit in both defining and attributing provenance to outliers. I will next argue that value-laden decisions seep into detection of outliers.

V. How do we detect outliers?

When the diversity of definitions used to classify outliers above are expressed mathematically they become methods for detection. Like the definitions from which spring, these detection methods are designed to capture outliers of specific data types. Documentation of all outlier detection methods is beyond the scope of this project and is the subject of many existing publications (e.g Anguins et al. 2013; Beckman and Cook, 1983; Smiti, 2020; Suri and Athithan, 2019; Taha and Hadi, 2019).

Just as definitions for outliers contain value-laden decisions, so too detection methods. Value-laden decisions survive the translation from natural language to mathematical notation.

For example, let’s revisit one of the cases where WBI with a millimeter wave scanner flagged a transgender individual as anomalous (Constanza-Chock, 2018). Since

we cannot look under the hood of the algorithm(s) used to detect anomalous objects on individuals (Waldron and Medina, 2019), we cannot determine with certainty the location and form of the value-laden decision(s) implicit in the algorithm(s). However, given that these algorithms specify parameters under which anomalous labels are assigned to individuals, it follows that, if empirical parameter estimates are not available, they cannot be specified without introducing ones priors or values.

Let's first assume, for example, that Constanza-Chock's (2018) diagnosis is correct and anomalies are triggered by millimeter-sized deviations from the statistical norm of the gender assigned by the TSA officer. If this is the case then how does one define the anomaly thresholds without recourse to unrepresentative estimates (because they don't appear to reflect transgender individuals) or value-judgements?

Alternatively, let's assume that TSA's press secretary was correct when she said that screening is completed "without regard to a person's race, color, sex, gender identity, national origin, religion or disability." (Waldron and Medina, 2019). If this is the case then, again, how does one decide where to draw the line between normal and anomalous individuals without introducing one's own values? And how does one deploy algorithms that consistently confuse penises and breasts for concealed objects in female and male presenting individuals, respectively, without accounting for gender?

To be charitable, let's finally assume that there are other ways that these algorithms can arrive at the same results without value-judgements. Okay, fine. But then what accounts for the systematic bias towards transgendered individuals? Regardless of who is correct here it is hard to escape value-laden decisions. Moreover, it is hard to escape the ethical principles they challenge.

For example, the principle of beneficence (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978) - to do no harm and maximize benefits - was violated for Constanza-Chock (2018) and others documented by (Waldron and Medina, 2019). Justice, and the distribution of benefits and burdens, was biased against transgendered individuals. They bore a disproportionate amount of the burdens (Waldron and Medina, 2019). Finally, TSA appears to have violated the respect for law and public interest (Dittrich and Kenneally, 2012) for their failure to reveal what is going on under the hood of the algorithm.

From this section I hope it is clear that value-laden decisions can be embodied into outlier detection algorithms. As illustrated above, these decisions can challenge ethical principles and elevate risks of privacy harms. What's more, value-laden decisions pop up after we detect outliers and are faced with decisions about how to treat them.

VI. How do we deal with outliers?

From the perspective of a data analyst, one usually encounters decisions about how to deal with outliers after they are detected. Once detected, we generally dig deeper and attempt to ascribe provenance. For example, if we can uncover evidence of measurement or execution error then maybe the outlier(s) in question do not represent inherent variability and thus *should* be omitted. But what if we don't discover evidence consistent with either type of error? What then?

In lieu of evidence or error we have to entertain the possibility that the outlier(s) in question arose from inherent variability. Strategies for handling these types of outliers include accommodation, incorporation, identification, rejection (Barnett 1978), and others

described by Anguinis et al. (2013). More broadly, there are two strategies: to include or exclude outliers from data. Value-laden decisions are implicit in each.

Let's start with exclusion.

VI.I Risks entailed by outlier exclusion.

Once we resign ourselves to the possibility that the outliers in questions may have arisen from inherent variability we might think that "these outlier(s) are *too* different to account for inherent variability". This was Mr. Hadlum's response to his wife's long gestation period. Under the influence of our intuitions it is tempting to argue that the outlier(s) *should* be removed under these circumstances because, afterall, values *this* different could not have arisen from inherent variability, right?

Not necessarily. This argument confuses appearance with reality. Just because something looks *off* does not mean it is off. Moreover, oftentimes we do not know how the data *should* be distributed. This is often why we collect data - to understand the processes that generate it. In this way we are not unlike the prisoners in Plato's allegory of the cave. We just see the shadows cast by the *real* objects. We see the samples, not the population from which they were collected. Thus, to omit outliers with the defense that they were *too* different to represent inherent variability is to claim that we know how data *should* be distributed. But if this is true then why collect data at all?

If this sounds too highbrow to warrant consideration, please think again. Consider the case of the Challenger disaster discussed here (https://blogs.ischool.berkeley.edu/w231/). One of the reasons the "outlying" temperature values were not considered in the final discussion between NASA and Morton Thiokol

was that some executives did not think that they contributed any information (Presidential Commission on the space shuttle Challenger Accident, 1986).

The cost of this tragic oversight, of course, was catastrophic. Violations of respect for persons, justice, beneficence, and respect for law and public interest were abound. Respect for persons was challenged because the crew was not debriefed about the risks of O-ring failure and consent was not solicited. The distribution of burdens were far from just since some people - the crew and their bereft families - suffered more than others involved. Similarly, beneficence was violated since harms were not minimized and benefits not maximized - for this the launch should have been postponed. Likewise, law and public interests were not respected since the process by which the decision-makers came to a conclusion was not transparent until after the tragedy. Moreover, since some of the decision-makers were privy to the risks of O-ring failure under cold conditions, the recommendation that the launch proceed as planned constitutes a deceptive act, as diagnosed by the three-part test conceived by the Federal Trade Commission (Federal Reserve, 2016). At the risk of oversimplification, this disaster was, at least in part, due to value-laden decisions masquerading as positivism. The decision-makers excluded the outliers because they did not *look* like they contributed to the decision before them. They confused *their* 'view from nowhere' (à la Jorgenson, 2014 and Nagel 1989) with their actual views from somewhere.

After witnessing the failure of appeals to some Platonic aesthetic of how data *should* behave we might arrive at the conclusion that we should employ a rejection rule to justify the retention or removal of the outliers. Numerous retention rules are available (Anscombe 1960; Dean and Dixon 1951; Ferguson 1961; Grubbs 1950). But, just

because we have test results to guide our decisions does not mean that we have escaped all value-laden decisions. We just outsourced them to someone else. The burden has just been shifted. Positivism and the “view from nowhere” appear to exist, until we look close enough, then they disappear.

For example, let’s take the case of the hole in the ozone layer. In the 1980’s, NASA was studying ozone levels with the Nimbus satellites in Antarctica (Newman, 2018). The resultant data they recorded were expressed in Dobson Units where 1 DU is the same as a 0.01mm thick layer of pure ozone under standard pressure and temperature (Newman, 2018). As the DU data were collected, observations were retained if > 180 DU and flagged as anomalous otherwise (Real Climate, 2017). Thus, by flagging these values NASA had inadvertently blinded itself from detecting the depletion of the ozone layer above Antarctica (Sparling, 2001).

Fortunately, the British Antarctic Survey and, later, NASA itself realized that ozone levels were dropping above Antarctica (Farman et al. 1985; Sparling, 2001). “No harm, no foul” might be the response of a Millian utilitarianist. Since no one was harmed there is no reason to belabor the point, they might say. But this perspective misses the larger point, one not unique to NASA. The action itself, not only the consequence(s), matter according to the deontological argument offered by Kantians.

When we flag or exclude observations from a dataset from which we plan to make inference, we can introduce systematic bias (Gress et al., 2018). For example, imagine if NASA had not course-corrected and corroborated BSA’s results. There is a non-zero chance that we might still be contributing to ozone depletion at scales like those prior to the Montreal Protocol. If true, then the value-laden decision to flag DU values below 180

might have challenged ethical principles like beneficence (if people were harmed) and justice (if the distribution of harms was not uniform). Thus, caution needs to be exercised making decisions about how to flag and potentially omit observations.

Moreover, consequences from these types of value-laden decisions impact more than ethics. For example, consider Thomas Kuhn's celebrated conception of the structure of scientific revolutions (Kuhn, 1970). Kuhn argues that scientific revolutions result from the accumulation of anomalous results that "subvert[s] the existing tradition of scientific practice" and precipitate paradigm shifts (Kuhn, 1970). If this model of science is accurate then it seems reasonable to expect that systematic outlier omission might constipate the pace of scientific revolutions.

From this perspective we can now see that, in some contexts, outliers can be disproportionately valuable. That is, they can contain *the* signal of interest. Examples of these disproportionately valuable outliers are abound in domains like surveillance, privacy, and fraud, fault, and intrusion detection (Chandola et al. 2009; Huberman et al. 2005). In these cases, the outliers can become the primary objects of interest and non-outliers are only peripherally interesting because they provide contrast against which we can see the outliers. Under these circumstances exclusion is tantamount to sabotage in the broadest sense of the word.

To see how the value-laden decisions around outlier exclusion can elevate risks and precipitate sabotage, consider the thought experiment presented by Jonas Lerman (2013). Two hypothetical individuals are posited. One contributes to big data by engagement with the devices that collect it. The other does not. The result is an imaginary database that is populated with data from the first person but with very little data from the

second person. In this way, the second person is an outlier in the space of information richness - his/her row in the database is sparsely populated compared to her compatriots. Moreover, she/he is *excluded* insofar as her/his data, being mostly NAs, does not contribute as much as the first person. The value-laden decision here is that data retrieved from electronic devices captures a representative picture of the subjects of interest. The consequences that follow from this value-laden decision include economic, civic, and political penalties (Lerman, 2013).

Lastly, for a more concrete case, recall the racial bias detected in facial analysis by Buolamwini (Hardesty, 2018). If one of the primary sources of bias was, as it seems, the exclusion of black faces from the training datasets (Hardesty, 2018) then it follows this value-laden decision challenges the respect for law and public interest criteria since the programmers were not transparent about the processes by which the algorithms were designed.

So the value-laden decisions that motivate outlier exclusion can introduce unanticipated risks and harms. This is not to say that we should never omit outliers. Indeed the same can be said about inclusion.

VI.II Risks entailed by outlier inclusion.

Examples of cases where outlier inclusion biases parameter estimates are abound (**CITATION**). Empirically documented cases where outlier inclusion challenges ethical principles are harder to find.

One example comes from historical cases where census data were used to characterize population demographics. Seltzer (2006) describes several instances where categorical outliers from population demographics are targeted and abused. For example,

in the 1940 United States census Japanese-Americans were counted like other demographics. This is the value-laden decision: to understand the population, we must count *everyone*. Unfortunately, this information was then misused as Japanese-Americans individuals were forced to migrate and retained in internment camps (Seltzer and Anderson, 2000; 2003). Thus, from the value-laden decision to include most of the population's subgroups in the census **arose** several ethical issues.

In this case, the consequences of these individuals being captured by the census entailed violations of respect for persons, justice, beneficence, and respect for law and public interest. Respect for persons is violated because consent was not solicited and autonomy was not respected since migration and internment were forced. Justice was violated because the distribution of burdens and benefits was not uniformly or fairly distributed across subgroups. The Japanese-Americans were targeted. Beneficence was violated because benefits to Japanese-Americans were not maximized and harms not minimized. Finally, respect for law and public interest was violated because the motivation for inclusion of Japanese-Americans was not transparent in real-time, but only once it was too late. From this example, we can see that seemingly innocuous, albeit value-laden decisions - to count and characterize the population of a country - can yield consequences that infringe on large groups of people. Lastly, it is not clear that we learned our lesson. The current president of the United States, Donald Trump, has attempted to use census data to reapportion the House (Wines and Bazelon, 2020).

From the above discussion, one can see that the value-laden decisions involved in both outlier inclusion and exclusion can challenge ethical principles. Other examples surely exist. Additionally, surely there are many cases where outlier inclusion and

exclusion do not engender sticky moral questions. Because of selection bias, we mostly know about cases where inclusion and exclusion go *wrong* - not where they go *right.* Finally, there are many technical remedial solutions we can use to handle outliers (Anguinis et al. 2013; Beckman and Cook, 1983). Inclusion and exclusion, as noted by Kruskal (1960) does not exhaust the space of possible solutions: “it is a dangerous oversimplification to discuss apparently wild observations in terms of inclusion in, or exclusion from, a more or less conventional formal analysis.” The availability of remedial solutions to handle outliers does not, however, exculpate us from the consequences entailed by our value-laden decisions. That is, technical solutions don’t free us from the ethical principles we sometimes inadvertently violate. So what do we do? How should we proceed?

Various technical answers to this question come to us from times past. Bernoulli (1777), for examples, offers a mostly useless piece of advice: “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”. Others, like Kruskal (1960) offer procedural solutions: “ My own practice in this sort of situation is to carry out an analysis both with and without the suspect observations. If the broad conclusions of the two analyses are quite different, I should view any conclusions from the experiment with very great caution”. Still others, like Rider (1933), offer a more pragmatic solution: “In the final analysis it would seem that the rejection or the retention of a discordant observation reduces to a question of common sense”. Finally, Kruskal (1960) advises us to look at outliers as an opportunity to learn something new: “An apparently wild (or otherwise anomalous) observation is a signal that says: "Here is something from which we may learn a lesson, perhaps of a kind not

anticipated beforehand, and perhaps more important than the main object of the study" (Kruskal, 1960). This last framing, offered by Kruskal, to look at outliers as opportunities to learn something new and not as obstacles, is perhaps the most inspiring and constructive.

When framed through the lens of opportunity, outliers become something to look forward to, not an obligation to be avoided. This is perhaps *the* best way to deal with outliers.

VII. Conclusions and recommendations

Platitudes aside, I hope this essay clearly conveys my diagnosis of the problems we encounter when we try to define, detect, and deal with outliers without explicit recognition of our assumptions and their consequences. When faced with questions about outliers, value-laden answers often appear as first responders, even with recourse to empirical evidence. Sometimes there is a mirage and we think something *is* objective, a view from nowhere, when in fact, at bottom, we find a decision made by someone. The decision is generally only exposed when something goes wrong. As such we have a biased sample of value-laden decisions. The ones we see are the ones that violate established ethical principles. By the time this happens, it is generally too late to prevent issues from materializing.

So how do we predict which value-laden decisions will cause trouble? I am not sure if we can, however, strategies are available to minimize harms. One strategy is just to exercise awareness that outliers arise from multiple sources, may contain useful information, and may affect future consequences in ways that are not always obvious at present. This strategy serves as a reminder that risks are out there.

Another strategy is to exercise negative visualization. Popularized by Seneca and other stoic philosophers, negative visualization recommends that we imagine the worst-case future scenarios (Seneca, 1607). By brainstorming dystopic outcomes, we can visualize the possible outcomes of our actions. At best this might help us anticipate harms before they transpire. This exercise seems like a good strategy to envision how bad things could be. Unfortunately, worse harms, beyond the bounds of our imagination, are still possible.

Once we become aware of potential harms that could arise from our value-laden decisions we can test the products of negative visualization off the (i) ethical principles established by the Belmont report and Menlo report, and (ii) criteria for fair and deceptive acts and privacy harms. If the outcomes of our negative visualization challenge the ethical or privacy principles we can update our priors and proceed with commensurate caution.

Lastly, when faced with the potential ethical consequences of our value-laden decisions we can apply the golden rule, as recommended by (Vitak et al, 2016). “Would we do X to another person?” That is the question to which we must respond.

References

1. Anderson DM, Keith J, Novak PD, Elliot MA. 1998. Dorland’s Illustrated Medical Dictionary. Saunders, Philadelphia, PA.

2. Angiulli F, Fassetti F, Palopoli L, and Serrao C. 2020. Detecting and Explaining Exceptional Values in Categorical Data. SEBD. http://ceur-ws.org/Vol-2646/29-paper.pdf
3. Anguinis H, Gottfredson RK, Joo H. 2013. Best-practice recommendations for defining, identifying, and handling outliers. 16:270-301.
4. Anscombe FJ, and Guttman I. 1960. Rejection of outliers. Technometrics. 2: 123-147
5. Barnett V. 1978. The study of outliers: purpose and model. Journey of the Royal Statistical Society. 3:242-250.
6. Barnett V and Lewis T. 1994. Outliers in statistical data. John Wiley & Sons. 3rd Edition
7. Beckman RJ and Cook RD. 1983. Outlier…….s. Technometrics. https://www.jstor.org/stable/1268541?seq=1#metadata_info_tab_contents
8. 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.
9. Bessel FW and Baeuer JJ. 1838. Gradmessung in Ostpreussen und ihre Verindung mit Presussischen und Russischen Dreiecksketten. Berlin. Reprinted in Adhendlungen von FW Bessel
10. Bowker GC and Star SL. 1999.. What a Difference a Name Makes - The Classification of Nursing Work. Chapter 7 of *Sorting Things Out: Classification and its Consequences*. MIT Press. https://github.com/UC-Berkeley-I-

School/w231/blob/master/Readings/Bowker%20and%20Star.%20Sorting%20things%20Out%20ch7.pdf

11. Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Comput Surv. 2009;41(3):15.
12. Constanza-Chock S. 2018. Design Justice, A.I., and the Escape from the Matrix of Domination. JoDs. https://doi.org/10.21428/96c8d426
13. Dean RB and Dixon WJ. 1951. Simplified Statistics for Small Numbers of Observations. Anal. Chem., 1951, 23 (4), 636–638.http://depa.fquim.unam.mx/amyd/archivero/ac1951_23_636_13353.pdf
14. Dittrich D and Kenneally E. 2012. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research, Tech. Report, U.S. Department of Homeland Security, Aug 2012. https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/
15. Farman JC, Gardiner BG, and Shanklin JD. 1985. Large losses of total ozone in Antarctica reveal seasonal ClO*x*/NO*x* interaction. Nature. https://www.nature.com/articles/315207a0
16. Federal Reserve. 2016. Federal Trade Commission Act Section 5: Unfair or Deceptive Acts or Practices. Consumer Compliance Handbook. https://www.federalreserve.gov/boarddocs/supmanual/cch/ftca.pdf
17. Ferguson TS. 1961. Rules for Rejection of Outliers. Review of the International Statistical Institute.29: 29-43. https://www.jstor.org/stable/pdf/1401948.pdf
18. Gress TW, Denvir J, Shapiro JI. 2018. Effect of Remo ect of Removing Outliers on Statistical Inf ving Outliers on Statistical Inference: Implications tence: Implications

to Interpretation of Experimental Data in Medical Research. Marshall Journal of Medicine. https://mds.marshall.edu/mjm/vol4/iss2/9/

19. Grubbs FE. 1950. Sample criteria for testing outlying observations. *Annals of* Mathematical Statistics. **21** (1): 27–58. doi:10.1214/aoms/1177729885.
20. Grubbs FE. 1969. Procedures for Detecting Outlying Observations in Samples. Technometrics. 11: 1-21.
21. Hardesty L. 2018. Study finds gender and skin-type bias in commercial artificial-intelligence systems. MIT News. https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212
22. Huberman BA, Edar E, and Fine LR. 2005. Valuating privacy. EEE Security and Privacy. https://ssrn.com/abstract=488324
23. Hume D. 1779. An enquiry concerning human understanding. In D. Hume, Essays and treatises on several subjects, Vol. 2. Containing An enquiry concerning human understanding, A dissertation on the passions, An enquiry concerning the principles of morals, and The natural history of religion (p. 3–212). Unknown Publisher. https://doi.org/10.1037/11713-001
24. Jurgenson N. View From Nowhere. October 2014. https://thenewinquiry.com/view-from-nowhere/
25. Kertzer and Arel D. 2002. Censuses, Identity Formation, and the Struggle for Political Power. Chapter 1 of *Census and Identity: The Politics of Race, Ethnicity, and Language in National Censuses*. Cambridge University Press. https://github.com/UC-Berkeley-I-

School/w231/blob/master/Readings/Kertzer%20and%20Arel.%20%20Census%20and%20Identity%20ch1.pdf

26. Kraemer F, van Overveld K, Peterson M. 2011. Is there an ethics of algorithms? Ethics Inf Technol. 13: 251-260.
27. Kruskal WH. 1960. Some remarks on wild observations. Technometrics. https://www.jstor.org/stable/pdf/1266526.pdf?refreqid=excelsior%3A0062c3ba25755fe0e14547aab8ce7fe3
28. Kuhn, Thomas S. The Structure of Scientific Revolutions. Chicago :University of Chicago Press, 1970.
29. Legendre AM. 1805. Nouvelles Méthodes pour la Determination des Orbits des Cométes. Courcier, Paris.
30. Lerman J. Big data and its exclusions. Stanford Law Review. https://www.stanfordlawreview.org/online/privacy-and-big-data-big-data-and-its-exclusions/
31. Nagel T. 1989. The view from nowhere. Oxford University Press.
32. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1978. *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. [Bethesda, Md.]: The Commission.
33. Newman P. 2018. History of the Ozone Hole. NASA Ozone Watch. https://ozonewatch.gsfc.nasa.gov/facts/history.html

34. Pang G, Cao L, and Chen L. 2016. Outlier Detection in Complex Categorical Data by Modelling the Feature Value Couplings. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 1902-1908.
35. Real Climate. 2017. What did NASA know? and when did they know it? http://www.realclimate.org/index.php/archives/2017/12/what-did-nasa-know-and-when-did-they-know-it/#ITEM-20924-1
36. Rider PR. 1933. Criteria for rejection of observations. Washington University Studies - New Series, Science and Technology - No. 8, St. Louis.
37. Ranga Suri N.N.R., Murty M N., Athithan G. 2019. Outlier Detection: Techniques and Applications. Intelligent Systems Reference Library, vol 155. Springer, Cham. https://doi.org/10.1007/978-3-030-05127-3_5
38. Seltzer W. 2006. The Dark Side of Numbers: Updated. Mackensen R. (eds), *Bevölkerungsforschung und Politik in Deutschland im 20. Jahrhundert. VS Verlag für Sozialwissenschaften.*
39. Seltzer W and Anderson M. 2000. After Pearl Harbor: the proper role of population statistics in the time of war.https://margoanderson.org/govstat/newpaa.pdf
40. Seltzer W and Anderson M. 2003. Government Statistics and Individual Safety: Revisiting the Historical Record of Disclosure, Harm, and Risk.
41. Seneca LA. 1607. Ad Lucilium epistolarum liber M. Antonii notis, Ferdinandi Pinciani castigationibus, Erasmi Roterodami annotationibus, Joannis Obsopoei collectaneis, Jani Gruteri et Fr. Jureti animadversionibus illustratus (in Latin). Foillet.

42. Smiti A. 2020. A critical overview of outlier detection methods. Computer Science Review. https://doi.org/10.1016/j.cosrev.2020.100306
43. Sparling B. 2001. Ozone depletion, history, and politics. NASA. https://www.nas.nasa.gov/About/Education/Ozone/history.html
44. Taha A and Hadi AS. 2019. Anomaly detection methods for categorical data: a review. ACM Comput. Surv. 52, 2, Article 38.https://doi.org/10.1145/3312739
45. United States. 1986. Report to the President. Washington, D.C.: Presidential Commission on the Space Shuttle Challenger Accident.
46. Vitak J, Shilton K, and Ashktorab Z. 2016. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*, 941–953. https://doi.org/10.1145/2818048.2820078
47. Waldron L and Medina B. 2019. When Transgender Travelers Walk Into Scanners, Invasive Searches Sometimes Wait on the Other Side. ProPublica & Miami Herald. https://www.propublica.org/article/tsa-transgender-travelers-scanners-invasive-searches-often-wait-on-the-other-side
48. Wines M and Bazelon E. 2020. Flaws in Census Count Imperil Trump Plan to Exclude Undocumented Immigrants. New York Times. https://www.nytimes.com/2020/12/04/us/census-trump.html