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 are often confronted with decisions about outliers. What are outliers? How should we define, detect, and deal with them? Where do they come from?

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, and the actions they can provoke, 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 feature common to all of these cases, I argue, is that they hinge on value-laden decisions that, at face-value

After review of these cases, a diagnosis is needed. What is the *cause* of these incidents? I argue that decisions that seem objective and sometimes trivial are in fact subjective, non-trivial, and, in the absence of supporting evidence, depend on implicit values.

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.

Further, I describe how systematic outlier omission might snowball and stifle scientific advancements by suppression of anomalies - the things that Thomas Kuhn argued often precipitate paradigms shifts.

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. 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](https://doi.org/10.1214%2Faoms%2F1177729885). [*Annals of* Mathematical Statistics](https://en.wikipedia.org/wiki/Annals_of_Mathematical_Statistics). **21** (1): 27–58. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1214/aoms/1177729885](https://doi.org/10.1214%2Faoms%2F1177729885).
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](https://books.google.com/books?id=s4pXAAAAcAAJ&pg=PT25&dq=%22Futurorum%20malorum%20pr%C3%A6meditatio%22) (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