HDSR
HARVARD DATA SCIENCE REVIEW

Volume 0 Issue 0
DOI: 00.000
ISSN: 2644-2353

Cautionary notes for data-driven science policy

Jevin D. West[†,*]

[†] Center for an Informed Public, Information School, University of Washington

Abstract. Data can misinform as much as it informs. As science policy increasingly incorporates more data into its analysis and decisions, it is important that policymakers avoid data pitfalls. This perspective highlights four cautionary notes: (1) avoiding the over-dependence on outside organizations and institutions for core datasets, (2) investing in data and software audits, especially data that includes demographic information and entity resolution, (3) training policymakers to identify quantitative distortion in data graphics and statistics, and (4) designing metrics that encourage the kind of science that innovates rather than just build long curricula vitae.

Keywords: Science Policy, Science of Science, Data Science, Machine Learning

Media Summary

Science policy is increasingly being informed by data. This is a good thing for the field and for science, but as this trend continues, there are some cautionary notes: be wary of dependence on data, invest in data and software audits, train policymakers to spot and refute quantitative BS, and design the inevitable metrics that encourage successful science outcomes.

Data-driven science policy

Science is tenaciously empirical Strevens, 2020. Its persistent pursuit of quantifiable evidence is one of its most salient features of success. The reverence to new data and new evidence puts science in a constant sanity check. Ironically, the policies that steer science are far less data-driven. Things are improving with new data sets for exploring policy-related questions (e.g., the UMETRICS data initiative Lane et al., 2015), but gaps remain. As a thought exercise, think about a dataset, analysis or random control trial that created or eliminated a funding program. They are rare. Science policy is less data-driven and more shoot-from-the-hip.

There are many reasons for this data detachment. One, science is a social process, and social data is messy and difficult to collect, curate and analyze. Two, privacy issues, rightly, raise roadblocks. Three, measuring 'success' in science is like rating the best restaurant; everyone has a different opinion, not only of the best restaurants, but how the restaurants should be measured. And, four, the timelines needed to gather evidence supersede leadership cycles at funding agencies.

[*] jevinw@uw.edu

This isn't a paper, though, arguing that we should be informing policy with data. One would be hard-pressed to find anyone not in favor of such efforts. I will assume that data will increasingly drive many of the policy decisions and programs at funding agencies and policy centers. Instead, the goal of this perspective is to highlight some cautionary notes for the world when data dominates policy decisions. Data can misinform as much as it can inform. This can occur both at the collection phase and analysis phase. In this perspective, I will highlight two cautionary notes at the data collection and curation phase. I will encourage policymakers to be wary of dependence of core data sets on outside organizations and institutions. I will then reiterate what others have discussed regarding audits both of data and the software ingesting the data. I will then turn to the post-data-collection phase, where analysis and communication of the analysis occurs. Finally, I will end with a note about the natural cycle of data-to-metrics and how these metrics can influence behavior that encourages good science, but also bad science if we are not careful.

Science policy has not been at the forefront of data-driven decision-making when compared to other fields. However, if the field can learn from the mistakes of others, this data delay can be an asset rather than a liability.

1. Be wary of data dependencies

I have been working with bibliometric data for well over a decade. I have relied on these collections to ask questions about gender bias in science West et al., 2013, the cost effectiveness of open access journals West et al., 2014, and the narrowing of the literature Kim et al., 2020. The data is made available through proprietary sources (e.g., Clarivate's Web of Science [WoS]) and open sources (e.g., arXiv). Few are as clean and comprehensive as the WoS, but it comes with sharing restrictions and cost[1]. The open sources are less restrictive but come with a reliability cost. That is why, in 2016, the release of Microsoft's Academic Graph (MAG) created waves within the science of science community. Powered by Microsoft's Bing indexer MAG Sinha et al., 2015, MAG is a scholarly search engine and database with nearly 250 million articles and more than 1.6 billion citations. Because of the automated approach for indexing, there are problems, such as duplicate entries and missing data, but the entity matching of authors and institutions, while far from perfect, is still reasonably good. One of the key advantages of MAG, when compared to projects like Google Scholar, that shares none of its data, is its openness to the research community. And not just parts of the data but the entire graph, which is necessary for large-scale clustering Bae et al., 2017. It has provided this service to the research community at low cost (if any cost for most researchers). It almost seemed too good to be true...

Unfortunately, it was too good to be true. On May 4, 2021, MAG was officially killed[2]. Their rationale is to expand its mission 'to have intelligent agents gather knowledge and empower humans to gain deeper insights and make better decisions'—whatever that means. This graveyarding of projects is typical for the technology industry. The 'killedbygoogle' project highlights the Google graveyard, which is extensive[3]. My research also relies on social media data. My colleagues and I are at the whim of these companies. For social media, it is the only game in town. There is no

[1]Full disclosure: my lab receives access to the WoS at no cost so I don't have the cost restrictions that others may have.

[2]https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/

[3]https://killedbygoogle.com/

government sponsored social media sights that can be used to track the spread of misinformation in science West and Bergstrom, 2021 but for most science data, this is not the case. Science is funded by tax dollars and therefore in a position to collect, curate, and make available this data to researchers.

The death of MAG is a significant hit to the field. There have been hundreds and possibly thousands of research papers that have relied on MAG since its inception in 2016. I appreciate Microsoft's investment and contribution to the field, but as data becomes a larger part of science policy, the field cannot rely on industry to cover data needs. There do exist seemingly reliable sources from industry (e.g., WoS, Scopus) and new ones have come on board (e.g., Dimensions), but whenever possible, research agencies should invest in data sets critical to informing policy (e.g., UMETRICS). The most interesting research questions will likely depend on long-term, consistent collections that are available to researchers, just as they are for other fields that depend critically on these investments, such as the Sloan Digital Sky Survey (SDSS) in astronomy York et al., 2000. This democratization of science is preferable to the current state where only a select number of researchers have access to entire corpora.

2. Invest in and reward data/software audits

I teach an introductory Machine Learning (ML) course at the University of Washington where students learn about cross validation, gradient descent, regularization, neural networks, and many of the simple, time-tested algorithms such as k-nearest neighbors, naive bayes, and decision trees. Students come eager to learn. The job prospects are strong for these skills and hardly a day goes by without students hearing some 'game-changer' news related to ML. Many enter the class thinking the algorithm is the crucial part of ML. By the end of the class, they realize that the predictive success is as much, if not more, about the quality of the training data than the algorithmic minutiae.

Yet much of the research efforts in ML are about the algorithm, optimization and applications. Less sexy are the tasks of assembling, curating, and auditing the datasets used to train and evaluate the newest computer vision algorithm or natural language processing (NLP) model. This lack of attention has consequences. Northcut and colleagues identified a plethora of errors in the 10 most popular datasets in computer vision, NLP and audio classification Northcutt et al., 2021. For example, they find that 6% of the validation set for ImageNet, one of the core datasets used to train and benchmark computer vision algorithms Deng et al., 2009. The researchers then looked at the consequences of these errors. Interestingly, they discovered that the simpler models perform better than the more complicated ones when the errors were corrected. This could have consequences for your self-driving car and the omnipresent surveillance cameras found around the world.

ImageNet was developed in 2006 in response to the same problem we are seeing today. Everyone wants to invent the next, new algorithm rather than delve into the messiness and difficulty of labeling data and reviewing existing algorithms. Fei-Fei Li, the creator of ImageNet, wanted to change that with a large-scale, human-annotated data set Deng et al., 2009. This was a necessary step for advancing computer vision, but it is not enough to create these data sets, especially when they involve 14 million images annotated by error-prone, Mechanical Turk. Researchers have recently discovered that, not only are their consequential errors in these datasets for the algorithms, but also for the people in these datasets and the people affected by these algorithms Crawford and Paglen,

2021. Rayid Ghani, and others at this *Value of Science Conference*[4] will discuss auditing toolkits being developed to address some of these bias and fairness issues Saleiro et al., 2018.

One of the core principles of any ML class is Garbage In, Garbage Out (GIGO). Unfortunately, GIGO doesn't always translate outside the classroom. As science policy increases is data diet, it is critical to continually sanity-check and humanity-check the datasets in which consequential decisions are made. This includes, but certainly is not limited to, large-scale datasets with demographic features and entity resolution needs (e.g., author disambiguation). There is a hardly a day when I don't find some strange issue in the multi-million-row bibliometric data sets that I use to explore science of science questions. To some, big data means certainty and the end of theory Anderson, 2008; to me, big data just means more auditing and being in a permanent state of 'preliminary results.'

3. Train policymakers to call BS on data and machines

Governments are increasingly communicating to the public with data. The Georgia State Department of Public Health (GSDPH), like many health departments, has delivered tables, statistics and data graphics of COVID-19 case counts throughout the pandemic. Should businesses close? Should kids attend school? Should people wear masks? Data can help to answer these questions, but raw data is not the primary form of communication. More often data is communicated through data graphics and statistics. In July of last year, GSDPH posted two heatmaps (see Figure 1). The heatmap posted on July 2 showed three counties with more than 3000 cases per 100 thousand. GSDPH then posted a heatmap on July 17 and compared it to the map on July 2. A twitter-post glance showed little difference. Two of the three counties were still red, and the rest of the counties were mostly the same shades of blue as they were on July 2. There seemed to be little change over those two important weeks, when policy debates were taking place regarding the questions above. There was a problem, though. The bins on July 2 were very different than the bins on July 17. This was not a fair comparison. The bins on July 17 were all significantly higher in case counts. Things were not stable in Georgia over this two-week period; they were higher across almost all counties in the state.

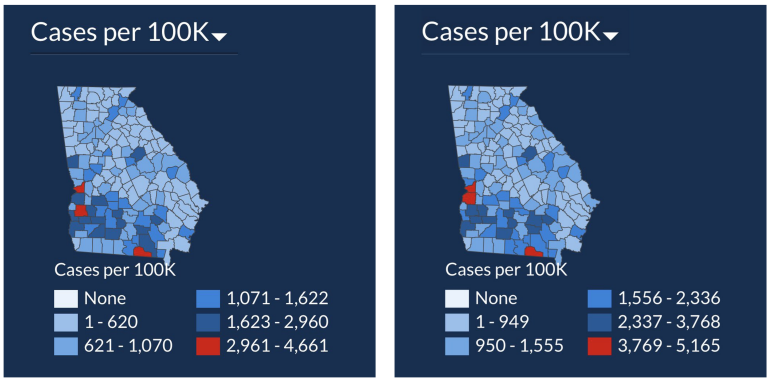

Figure 1. Heatmaps showing COVID-19 case counts per 100k in the counties across the state of Georgia on July 2, 2020 and July 17, 2020.

[4]https://coleridgeinitiative.org/value-of-science-conference/

Policies often intend to be data-driven but instead become visually distorted. Data graphics and statistics are mediums in which large, complex data are communicated. Policymakers need to be on the lookout for the ways which they can mislead as much as they can lead. Are the visual attributes of a data graphic, like the area of bar graphs, proportional to the quantitative values they are supposed to present? If not, they may be violating the *Principle or Proportional Ink* Bergstrom and West, 2017. This often occurs when presenters zoom in on a bar graph so far that even small differences look big. Conversely, distortion can occur when line graphs are zoomed out so far that trends essentially disappear. My colleague, Carl Bergstrom, and I discuss these pitfalls in our recent book Bergstrom and West, 2021 and the various ways in which visualizations can misinform.

Being on the lookout for data distortion is not just the job of consumers but also the producers of information graphics. In the Georgia State example above, the department was accused of pushing a political agenda based on several visualization mistakes. They defended themselves against this criticism, saying that the heatmaps were automatically binned by the software they were using. This may indeed be true, which is even more reason to invest in training to help them spot these errors. Just as technology has created autopilot for cars and airplanes, data acquisition, data cleaning, and data analysis will increasingly be automated. This isn't reason to nap; it is reason to pay attention even more. Automation can lead to complacency, onerous conclusions and overconfidence.

In early 2017, my colleague Carl Bergstrom and I released a new course at the University Washington titled, *Calling Bullshit: Data Reasoning in a Digital World*[5]. The course reflected our frustrations in seeing data being used to make specious arguments both in our professional and personal lives. At the same time, the crisis of misinformation was becoming a concern both in society and science West and Bergstrom, 2021. Our goal was to better prepare our students for this misinformation crisis, especially the kind wrapped in data. We teach students to not conflate correlation and causation. But who makes that mistake; everyone knows that correlation does not imply causation? Journalists and researchers make this mistake. When examining the top 50 most shared health-related studies, authors of a recent study found that journalists mis-attributed causation in 1/2 of the news articles about the studies Haber et al., 2018. Even worse, 1/3 of the journal articles themselves made the same mistake. A similar study should be conducted in science policy. I am guessing their will be more than zero examples.

Even less complicated than causation but even more prevalent is selection bias. If we ask attendees of this conference the importance of data investments, I am sure the average score would be higher than the average score for all academics, many of whom are critical of quantitative methods. Even though selection bias can be fairly straightforward when looking for data samples that are not representative of the full population, it can get trickier but just as important for science policy to be on the lookout for data censoring, the Will Rogers effect Feinstein et al., 1985, and Berkson's paradox Bergstrom and West, 2021. Berkson's paradox occurs when two traits are negatively correlated but appear positively correlated in the population. For example, why is it that the best hitters in baseball are often mediocre fielders and vice versa? There is selection against those that are mediocre at both traits, which gives the illusion of a negative correlation between these traits.

Most of what we talk about in the class are the ways that humans fool and are fooled by others. But machines are also prone to being fooled or look for signals that have nothing to do with the underlying task. Researchers at my university recently showed what machines are 'seeing' in chest radiographs when trying to predict whether a patient had COVID-19 or not DeGrave et al., 2021.

[5]https://www.callingbullshit.org/

They found that the algorithms used to make these predictions were taking spurious shortcuts. Instead of looking at the lungs in the images, the algorithms focused on irrelevant features, such as the patient position and text markers on the radiograph. These shortcuts limit the machine's ability to generalize and can even lead to dangerous diagnoses. As statistical learning and data further infuse science policy, it is important to be aware of the ways in which machines and statistical models can be fooled by tangential data features.

When teaching students to spot BS (Bad Science or Bullshit, whichever is most appropriate), we emphasize Hanlon's Razor: 'never attribute to malice that which is adequately explained by stupidity'. But even better, never attribute to malice that which is adequately explained by honest mistake... because we have all made them. The same applies to the Georgia State example and to fields of research. As data infuses science policy, spotting those pitfalls of big data will be most successful in a culture where calling BS is encouraged but civically minded.

4. Design metrics that encourage successful science

When data arises, metrics follow. When Eugene Garfield started collecting citations between journals, the infamous Impact Factor followed Garfield, 1999. Today, even after countless discussions about the misuses of this metric, journals, scholars, and universities still misuse and abuse the metric Simons, 2008. But this is not specific to Impact Factor. It applies to any marker that measures science success, production and innovation. When measures are invented that govern resources, humans respond. My colleagues and I wrote about this in a commentary in 2010 West, 2010. I noted then (and still believe it today) that 'giving bad answers is not the worst thing a ranking system can do—the worst thing is to encourage bad science' West, 2010 British economist, Charles Goodhart, recognized this many years before. He noted that 'any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes' Goodhart, 1984. Even more succinctly, Marilyn Strathern noted that 'when a measure becomes a target, it ceases to be a good measure' Strathern, 1997.

As we collect more data in science policy, there will be an natural blossoming of new metrics to inform funding, awards and hiring decisions. We need to be extra careful how this inevitable set of new metrics influences behaviors that counter the goals of science. If we are not careful, all this data collection will just be an exercise in plausible deniability.

Conclusion

The value of science can be realized more effectively with data. But data can also misinform. Going forward, it is important that policymakers protect its core data sets and communicate accurately their analyses with scientists and the public.

Disclosure Statement. I am the co-author of the book, 'Calling Bullshit: The Art of Skepticism in a Data-Driven World.'

Acknowledgments. I would like to thank Carl Bergstrom for comments and discussions on the topic of data.

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7), 16–07.

Bae, S.-H., Halperin, D., West, J. D., Rosvall, M., & Howe, B. (2017). *Scalable and efficient flow-based community detection for large-scale graph analysis*.

Bergstrom, C. T., & West, J. D. (2017). Vizualization: The principle of proportional ink [Online resource]. https://callingbullshit.org/tools/tools_proportional_ink.html

Bergstrom, C. T., & West, J. D. (2021). *Calling bullshit: The art of skepticism in a data-driven world*. Random House Trade Paperbacks.

Crawford, K., & Paglen, T. (2021). Excavating ai: The politics of images in machine learning training sets. *AI & SOCIETY*, 1–12.

DeGrave, A. J., Janizek, J. D., & Lee, S.-I. (2021). Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 1–10.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Feinstein, A. R., Sosin, D. M., & Wells, C. K. (1985). The will rogers phenomenon: Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *New England Journal of Medicine*, *312*(25), 1604–1608.

Garfield, E. (1999). Journal impact factor: A brief review. *Cmaj*, *161*(8), 979–980.

Goodhart, C. A. (1984). Problems of monetary management: The uk experience. *Monetary theory and practice* (pp. 91–121). Springer.

Haber, N., Smith, E. R., Moscoe, E., Andrews, K., Audy, R., Bell, W., Brennan, A. T., Breskin, A., Kane, J. C., Karra, M., et al. (2018). Causal language and strength of inference in academic and media articles shared in social media (claims): A systematic review. *PloS one*, *13*(5), e0196346.

Kim, L., Adolph, C., West, J. D., & Stovel, K. (2020). The influence of changing marginals on measures of inequality in scholarly citations: Evidence of bias and a resampling correction. *Sociological Science*, *7*, 314–341.

Lane, J. I., Owen-Smith, J., Rosen, R. F., & Weinberg, B. A. (2015). New linked data on research investments: Scientific workforce, productivity, and public value. *Research policy*, *44*(9), 1659–1671.

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

Simons, K. (2008). The misused impact factor. *Science*, *322*(5899), 165–165.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. (, & Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications, 243–246. https://doi.org/10.1145/2740908.2742839

Strathern, M. (1997).
improving ratings': Audit in the british university system. *European review*, *5*(3), 305–321.

Strevens, M. (2020). *The knowledge machine: How irrationality created modern science*. Liveright Publishing.

West, J. D. (2010). How to improve the use of metrics: Learn from game theory. *Nature*, *465*, 870–872.

West, J. D., & Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences*, *118*(15).

West, J. D., Bergstrom, T., & Bergstrom, C. T. (2014). Cost effectiveness of open access publications. *Economic Inquiry*, *52*(4), 1315–1321. https://doi.org/10.1111/ecin.12117

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PloS One*, *8*(7), e66212. https://doi.org/10.1371/journal.pone.0066212

York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J., Barkhouser, R., Bastian, S., Berman, E., et al. (2000). The sloan digital sky survey: Technical summary. *The Astronomical Journal*, *120*(3), 1579.