Numeracy
Advancing Education in Quantitative Literacy

Volume 15 | Issue 1 Article 1

2022

An Introduction to *Calling Bullshit*: Learning to Think Outside the Black Box

Jevin D. West
University of Washington, jevinw@uw.edu
Carl T. Bergstrom
University of Washington, cbergst@uw.edu

Recommended Citation

West, Jevin D., and Carl T. Bergstrom. "An Introduction to *Calling Bullshit*: Learning to Think Outside the Black Box." *Numeracy* 15, Iss. 1 (2022): Article 1. DOI: https://doi.org/10.5038/1936-4660.15.1.1405

An Introduction to *Calling Bullshit*: Learning to Think Outside the Black Box

Abstract

Bergstrom, Carl T. and Jevin D. West. 2020. *Calling Bullshit: The Art of Skepticism in a Data-Driven World.* (New York: Random House) 336 pp. ISBN 978-0525509202.

While statistical methods receive greater attention, the art of critically evaluating information in everyday life more commonly depends on thinking outside the black box of the algorithm. In this piece we introduce readers to our book and associated online teaching materials—for readers who want to more capably call "bullshit" or to teach their students to do the same.

Keywords

algorithms, statistics, critical thinking

Cover Page Footnote

Jevin D. West is Associate Professor and Director of the Center for an Informed Public in the Information School at the University of Washington.

Carl T. Bergstrom is Professor of Biology at the University of Washington.

Imagine you own a company that employs 1.3 million people. On average, the cost to hire each employee is $4,000—billions of dollars to sort through resumes, interview candidates, and make decisions. One of your leading data scientists argues for automating the process. Save billions! Plus, you have the millions of resumes needed to train a machine-learning algorithm. You house some of the most talented engineers in the world to build the system. You could sell the technology to other large companies. And you have dominated nearly every industry you have touched, from e-commerce to cloud computing to digital streaming. Why not human resources?

The pitch made to Amazon executives around 2014 may have been something like this. The argument sounded compelling. Shortly thereafter, a dozen or so high-paid engineers from Amazon's Edinburgh office began designing and testing automated hiring.[1] There must have been excitement in the air. Billions to be saved and bonuses to follow.

Unfortunately for those engineers and Amazon, the project crashed and burned. The machine had a fatal flaw: it didn't like women. The machine disproportionately ranked male candidates higher. The results were so biased that executives, to their credit, canceled the project and disbanded the engineering team.

So, what went wrong? How could the world's leading AI company fail so miserably? You might think that the answer requires advanced computer science degrees and decades of machine-learning experience. It doesn't. When we tell this real-world story, students in our class are able to explain with relative ease what went wrong using the black box schema (Fig. 1). They immediately focus their attention on what data was used to train the algorithms.

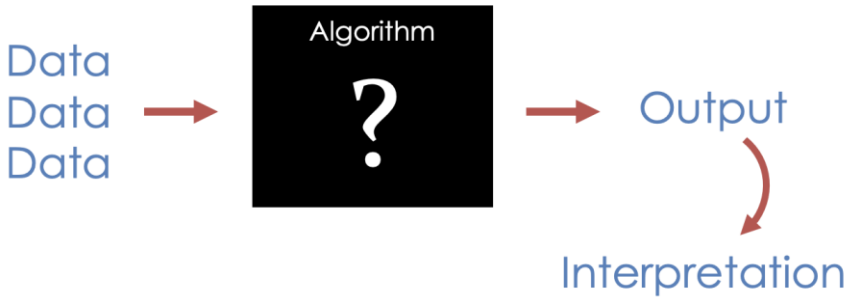


Figure 1. Black Box Schema. We often talk to our students about the black box. Even when you don't know how an algorithm or statistical test works, you can effectively spot quantitative BS much of the time by simply questioning what goes in and what comes out of the black box.

[1]https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

One of the central principles in our class and in our recent book is to ignore, or at least spend less time with, the black-box algorithms and complicated statistical procedures. Instead, focus one's attention on the input data and interpretation of the output data.

In the example above, Amazon trained their machines on existing CVs and hiring decisions from the previous ten years—a ten-year period when Amazon had disproportionately hired men. Though programmed likely to ignore gender and its most obvious correlates, the machine picked up on more subtle cues. Even if names and gender pronouns were removed in the CVs, it found secondary variables linked to gender (e.g., it downgraded candidates from all-women's colleges) and used these to preferentially choose male candidates. Trained on a biased data set, the machine learned—despite the engineers' best efforts—to replicate those biases.

It doesn't require a PhD in statistical learning to understand what went wrong. (It shouldn't have required millions of dollars-worth of engineers to see this problem either, but that discussion is for another time.) It simply requires a focus on the input data used to train the algorithms. This is where we concentrate our students' attention. Is the data a representative sample? What biases exist in the labels? What groupings in the data are likely to affect the interpretation of the results? Etc. These simple questions can help identify the obvious problems of research papers claiming an ability to automatically identifying criminals[2] and whether someone is gay[3] from photographs of human faces. In short, these papers are flawed and can do no such things.

Back in early 2017, we created a new class at the University of Washington: Calling Bullshit.[4] The goal of the class was to empower students and the public to question numbers, statistics, and overly confident AI. This is something that numeracy educators and professionals have been doing for decades and something that is needed more than ever. We build on this work with new case studies, tools, and open source lectures. We do this with a focus on current events such as the pandemic and the recent U.S. election. Over the years, we have found that students are exceptionally good at running python and R libraries, replicating code, and applying statistical procedures. They are less effective in questioning data and interpreting results. We want to help change this.

We have over 60 hours of lectures that we make freely available to the public, but if we could only teach one lecture for 5 minutes, it would be about the bullshit found outside the black box.

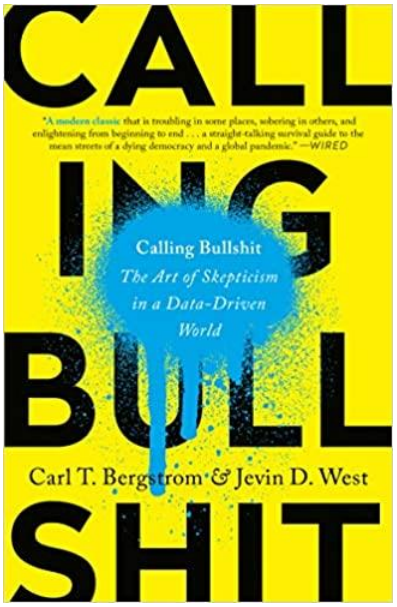
It is something we teach students from across campus, in the humanities, social sciences, and the sciences. It is also something we apply in our professional lives. We have reviewed thousands of science papers, reports, and grant proposals. We

[2] https://www.callingbullshit.org/case_studies/case_study_criminal_machine_learning.html
[3] https://www.callingbullshit.org/case_studies/case_study_ml_sexual_orientation.html
[4] https://www.callingbullshit.org/

DOI: https://doi.org/10.5038/1936-4660.15.1.1405

find that the majority of the problems are indeed outside the black box. It is also the place that requires the least amount of specialized knowledge to ask pertinent questions. Hopefully, these habits of mind translate into a new Edinburgh team in the future that immediately sees the flaws in using CVs as training data and instead saves millions of dollars for the next *Numeracy* conference.

Excerpt from *Calling Bullshit*[5]



Available from Random House

According to Latour, scientific claims are typically built upon the output of metaphorical "black boxes," which are difficult if not impossible for the reader to penetrate. These black boxes often involve the use of specialized and often expensive equipment and techniques that are time-consuming and unavailable, or are so broadly accepted that to question them represents a sort of scientific heresy. If I were to write a paper claiming that specific genetic variants are associated with susceptibility to bullshit, a skeptic might reasonably argue my choice of sample population, the way I measure bullshit susceptibility, or the statistical method I use to quantify associations. But the biotechnology used to derive the DNA sequences from blood samples would typically be treated as a black box. In principle a skeptic could question this as well, but to do so she would be challenging the scientific establishment and, more important for our purposes, she would need access to advanced equipment and extensive technical expertise in molecular genetics.

Latour is not saying that these aspects of academic science make the entire enterprise bullshit, and neither are we. He is saying only that science is more than a dispassionate search for the truth, a theme to which we will return in chapter 9. The important thing about Latour's black box idea is that we see a powerful analogy here to what speakers do when they bullshit effectively. Outright lies are often straightforward to catch and refute. But effective bullshit is difficult to fact-check. Bullshit can act like one of Latour's black boxes, shielding a claim from further investigation.

[5] Excerpted from *Calling Bullshit: The Art of Skepticism in a Data-Driven World* by Carl T. Bergstrom and Jevin D. West. Copyright © 2020 by Carl T. Bergstrom and Jevin D. West. Used by permission of Penguin Random House LLC. All rights reserved.

Suppose a friend tells you, "You know, on average, cat people earn higher salaries than dog people." It's easy to call bullshit on that statement when it stands by itself. And when you do so, perhaps your friend will simply laugh and admit, "Yeah, I made that up."

But suppose instead she doubles down and starts filling out—or making up—details to support her claim. "No, really, it's true. I saw this TED Talk about it. They explained how cat owners value independence whereas dog owners value loyalty. People who value independence are more likely to have NVT...no...NVS...I can't remember, but some kind of personality. And that makes them better able to rise in the workplace."

This is full-on bullshit, and it functions like one of Latour's black boxes. If you want to dispute your friend's claims, you now have substantial work to do. This is where lies and bullshit come together: In our view, a lie becomes bullshit when the speaker attempts to conceal it using various rhetorical artifices.

Now imagine that she points you to a research study that makes this claim. Suppose you track down the study, and read something like the following:

> We observe a statistically significant difference in cat- and dog-lovers' earnings, based on an ANCOVA using log- transformed earnings data ($F = 3.86$).

If you don't have a professional background in statistics, you've just slammed head-on into a particularly opaque black box. You probably don't know what an ANCOVA is or what the F value means or what a log transformation is or why someone would use it. If you do know some of these things, you still probably don't remember all of the details. We, the authors, use statistics on a daily basis, but we still have to look up this sort of stuff all the time. As a result, you can't unpack the black box; you can't go into the details of the analysis in order to pick apart possible problems. Unless you're a data scientist, and prob-ably even then, you run into the same kind of problem you encounter when you read about a paper that uses the newest ResNet algorithm to reveal differences in the facial features of dog and cat owners. Whether or not this is intentional on the part of the author, this kind of black box shields the claim against scrutiny.

But it doesn't need to. The central theme of this book is that you usually don't have to open the analytic black box in order to call bullshit on the claims that come out of it. Any black box used to generate bullshit has to take in data and spit results out.

Most often, bullshit arises either because there are biases in the data that get fed into the black box, or because there are obvious problems with the results that come out. Occasionally the technical details of the black box matter, but in our experience such cases are uncommon. This is fortunate, because you don't need a lot of technical expertise. You just need to think clearly and practice spotting the sort of things that can go wrong.

DOI: https://doi.org/10.5038/1936-4660.15.1.1405

If the data that go into the analysis are flawed, *the specific technical details of the analysis don't matter*. One can obtain stupid results from bad data without any statistical trickery. And this is often how bullshit arguments are created, deliberately or otherwise. To catch this sort of bullshit, you don't have to unpack the black box. All you have to do is think carefully about the data that went into the black box and the results that came out. Are the data unbiased, reasonable, and relevant to the problem at hand? Do the results pass basic plausibility checks? Do they support whatever conclusions are drawn?

Being able to spot bullshit based on data is a critical skill. Decades ago, fancy language and superfluous detail might have served a bullshitter's needs. Today, we are accustomed to receiving information in quantitative form, but hesitant to question that information once we receive it. Quantitative evidence generally seems to carry more weight than qualitative arguments. This weight is largely undeserved—only modest skill is required to construct specious quantitative arguments. But we defer to such arguments nonetheless. Consequently, numbers offer the biggest bang for the bullshitting buck.

References

Bergstrom, Carl T. and Jevin D. West. 2020. *Calling Bullshit: The Art of Skepticism in a Data-Driven World*. New York: Random House.