8 types of bias in data analysis and how to avoid them

Analytics can exhibit biases that affect the bottom line or incite social outrage through discrimination. It's important to address those biases before problems arise.

George Lawton

Published: **26 Oct 2020**

There are a variety of ways bias can show up in analytics, ranging from how a question is hypothesized and explored to how the data is sampled and organized.

"The need to address bias should be the top priority for anyone that works with data," said Elif Tutuk, associate vice president of innovation and design at Qlik. "If not careful, bias can be introduced at any stage from defining and capturing the data set to running the analytics or AI/ML [machine learning] system."

Although data scientists can never completely eliminate bias in data analysis, they can take countermeasures to look for it and mitigate issues in practice.

"Avoiding bias starts by recognizing that data bias exists, both in the data itself and in the people analyzing or using it," said Hariharan Kolam, CEO and founder of Findem, a people intelligence company.

THIS ARTICLE IS PART OF

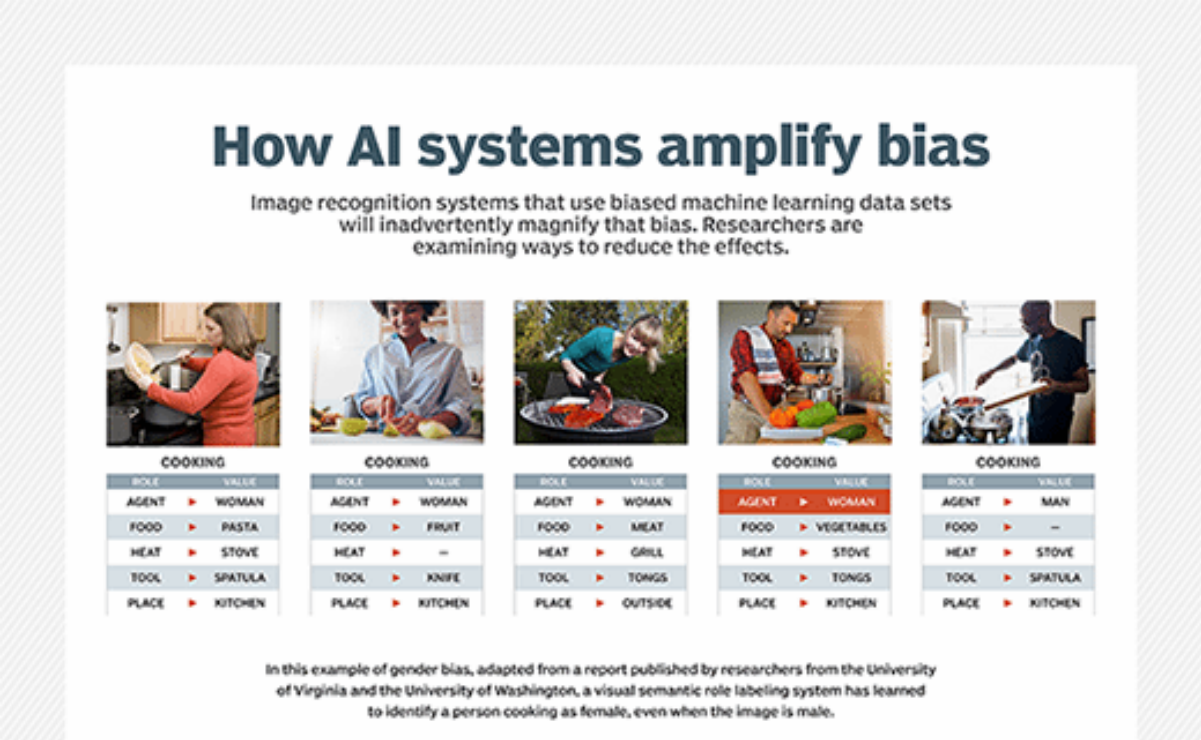
What is data science? The ultimate guide

Which also includes:

- **8 top data science applications and use cases for businesses**
- **8 types of bias in data analysis and how to avoid them**
- **How to structure and manage a data science team**

There are many adverse impacts of bias in data analysis, ranging from making bad decisions that directly affect the bottom line to adversely affecting certain groups of people involved in the analysis.

A root cause of all these problems is a lack of focus around the purpose of an inquiry. Kolam recommended data scientists get consensus around the purpose of the analysis to avoid any confusion because ambiguous intent most often leads to ambiguous analysis.



Bias can lead to misrepresentation in analytics and AI algorithms.

How does bias show in analytics?

"If you ask a data scientist about bias, the first thing that comes to mind is the data itself," said Alicia Frame, lead product manager at Neo4j, a graph database vendor.

Bias in data analysis can come from human sources because they use unrepresentative data sets, leading questions in surveys and biased reporting and measurements. Often bias goes unnoticed until you've made some decision based on your data, such as building a predictive model that turns out to be wrong.

Medical data tends to overrepresent white patients, particularly in new drug trials. As a result, the experiences and reports of new drugs on people of color is often minimized. This bias has urgency now in the wake of COVID-19, as drug companies rush to finish vaccine trials while recruiting diverse patient populations, Frame said. A lack of diversity is why Pfizer recently announced they were recruiting an additional 15,000 patients for their trials.

"Unfortunately, bias in analytics parallels all the ways it shows up in society," said Sarah Gates, global product marketing manager at SAS.

It is also a moving target as societal definitions of fairness evolve. A recent example reported by Reuters occurred when the International Baccalaureate program had to cancel its annual exams for high school students in May due to COVID-19. Instead of using exams to grade students, the IB program used an algorithm to assign grades that were substantially lower than many students and their teachers expected.

In business, bias can also show up as a result of the way data is recorded by people.

For example, "Salespeople updating CRM data rarely want to point to themselves as to why a deal was lost," said Dave Weisbeck, chief strategy officer at Visier, a people analytics company. By being more thoughtful about the source of data, you can reduce the impact of bias.

Here are eight examples of bias in data analysis and ways to address each of them.

1. Propagating the current state

One common type of bias in data analysis is propagating the current state, Frame said.

Amazon's (now retired) recruiting tools showed preference toward men, who were more representative of their existing staff. The algorithms didn't explicitly know or look at the gender of applicants, but they ended up being biased by other things they looked at that were indirectly linked to gender, such as sports, social activities and adjectives used to describe accomplishments.

In essence, the AI was picking up on these subtle differences and trying to find recruits that matched what they internally identified as successful. Frame said a good countermeasure is to provide context and connections to your AI systems.

2. Trained on the wrong thing

Arijit Sengupta, founder and CEO of Aible, an AI platform, said one of the biggest inherent biases in traditional AI is that it is trained on model accuracy rather than business impact, which is more important to the organization.

The root cause is that the algorithm is built with the assumption that all costs and benefits are equal. But in business, the benefit of a correct prediction is almost never equal to the cost of a wrong prediction. What if the benefit of winning a deal is 100 times the cost of unnecessarily pursuing a deal? You might be willing to pursue and lose 99 deals for a single win. An AI that only finds 1 win in 100 tries would be very inaccurate, but it also might boost your net revenue.

"Data scientists need to clarify the relative value of different costs and benefits," he said.

3. Under-representing populations

Another big source of bias in data analysis can occur when certain populations are under-represented in the data. This kind of bias has had a tragic impact in medicine by failing to highlight important differences in heart disease symptoms between men and women, said Carlos Melendez, COO and co-founder of Wovenware, a Puerto Rico-based nearshore services provider.

Bias shows up in the form of gender, racial or economic status differences. It appears when data that trains algorithms does not account for the many factors that go into decision-making. Melendez said good practices to mitigate this include using a diverse data science team, providing diversity training to data scientists and testing for algorithm bias.

4. Faulty interpretation

"When we approach analysis looking to justify our belief or opinion, we can invariably find some data that supports our point of view," Weisbeck said.

Medical researchers address this bias by using double-blind studies in which study participants and data collectors can't inadvertently influence the analysis. This is harder to do in business, but data scientists can mitigate this by analyzing the bias itself.

Weisbeck said Vizier conducted an internal study to understand the pay differences from a gender equity perspective. One technique was to segment the sample into data populations where they expected bias and where they did not. They then compared different outcomes by looking at pay adjustment for women who had male or female managers.

A second technique was to look at related results where they would expect to find bias in in the data. For pay equity, one example they tested was the statement: "If women face bias in compensation adjustments, then they also face bias in performance reviews." The latter technique takes advantage of the fact that bias is often consistent.

5. Cognitive biases

Cognitive bias leads to statistical bias, such as sampling or selection bias, said Charna Parkey, data science lead at Kaskada, a machine learning platform. Often analysis is conducted on available data or found in data that is stitched together instead of carefully constructed data sets.

Both the original collection of the data and an analyst's choice of what data to include or exclude creates sample bias. Selection bias occurs when the sample data that is gathered isn't representative of the true future population of cases that the model will see.

It's useful to move from static facts to event-based data sources that allow data to update over time to more accurately reflect the world we live in. This can include moving to dynamic dashboards and machine learning models that can be monitored and measured over time.

"Reminding those building the models as they build them -- and those making decisions when they make them -- which cognitive bias they are susceptible to and providing them with ways to mitigate those biases in the moment has been shown to mitigate unintentional biases," Parkey said.

6. Analytics bias

Analytics bias is often caused by incomplete data sets and a lack of context around those data sets.

> **Understanding the data that isn't part of the data set may tell as important of a story as the data that is feeding the analytics.**
>
> **Elif Tutuk**
> Associate vice president of innovation and design, Qlik

"Understanding the data that isn't part of the data set may tell as important a story as the data that is feeding the analytics," Tutuk said.

Static data is inherently biased to the moment in which it was generated. To handle these challenges, organizations need to use associative data technologies that can access and associate all the data.

Business is always in a constant feedback loop. Analytics must operate in real time, which means the data has to be business-ready to be analyzed and re-analyzed due to changing business conditions. Data managers need to work with IT to create contextualized views of the data that are centered on business view and use case to reflect the reality of the moment.

7. Confirmation bias

A confirmation bias results when researchers choose only the data that supports their own hypothesis.

"Most often, we carry out an analysis with a preconceived idea in mind, so when we go out to search for statistical evidence, we tend to see only that which supports our initial notion," said Eric McGee, senior network engineer at TRG Datacenters, a colocation provider.

Confirmation bias is found most often when evaluating results.

"If the results tend to confirm our hypotheses, we don't question them any further," said Theresa Kushner, senior director of data intelligence and automation at NTT Data Services. "However, if the results don't confirm our hypotheses, we go out of our way to reevaluate the process, the data or the algorithms thinking we must have made a mistake."

Kushner recommended developing a process to test for bias before sending a model off to users. For example, NTT Data Services applies a governance process they call AI Ethics that works to avoid bias in all phases of development, deployment and operations.

8. Outlier bias

Another common cause of bias is caused by data outliers that differ greatly from other samples.

"Including Jeff Bezos in an effort to analyze mean American incomes, for example, would drastically skew the results of your study because of his wealth," said Rick Vasko, director of service delivery and quality at Entrust Solutions, a technology solutions provider.

Outlier biases can be corrected by determining the median as a closer representation of the whole data set. If out of 10 people, one person has $10,000 in their bank account and the others have under $5,000, the person with the most money is potentially an outlier and should be removed from the survey population to achieve a more accurate result.