S&P Global

Avoiding Garbage in Machine Learning

October 11, 2019

Key Takeaways

- "Garbage" includes poorly labeled or inaccurate data, data that reflects underlying human prejudices, and/or incomplete data.
- Every dataset inevitably contains some bias. Bias in a dataset only matters if it is germane to the question being answered.
- Identifying "garbage" in your output requires both a general skepticism when evaluating results and a knowledge of best practices in data science.
- To use machine learning effectively, you need to embrace the potential for garbage and anticipate it.

"Garbage in, garbage out" — it's a cliché in machine learning circles. Anyone who works with artificial intelligence (AI) knows that the quality of the data goes a long way toward determining the quality of the result. But "garbage" is a broad and expanding category in data science — poorly labeled or inaccurate data, data that reflects underlying human prejudices, incomplete data. To paraphrase Tolstoy, great datasets are all alike, but all garbage datasets are garbage in their own, unique and horrible ways.

People believe in machine learning. Israeli philosopher and historian Yuval Noah Harrari coined the term "dataism" to describe a blind faith in algorithms. This faith extends beyond machine learning's ability to analyze data. Many people believe machine learning is auto-magically able to predict the future. This is inaccurate. Machine learning is excellent at identifying patterns in large well-labeled datasets. In certain cases, those patterns will continue to unfold in the future. In other cases, they won't. However, the machine-learning researcher must assume responsibility for the perception that AI is predictive. Mitigating the effects of garbage data becomes a moral imperative when your algorithm is being used to make parole decisions or to invest billions in hard-earned pension dollars.

Faith in machine learning isn't misplaced. The reason machine learning is considered more objective than other methods of data analysis is because it is demonstrably superior. Machine learning is able to produce models that can analyze larger, more complex datasets than conventional methods, and deliver more accurate results, all at scale. When used properly, machine learning can help companies to identify profitable opportunities and avoid risks. But even a technology so advanced can't perform the alchemy of turning garbage into gold. Learning to identify "garbage" data is the first step in unlocking machine learning's vast potential.

Faith in machine learning isn't misplaced

"Machine learning is not magic," cautions Nick Cafferillo, chief data and technology officer at S&P Global. "It's about augmenting our thought processes to help prove — or disprove — a hypothesis we have; a hypothesis that is founded on real business questions that we understand in the abstract."

Garbage data is a concern when you're working with machine learning because there are two opportunities for the garbage to mess up your results. First, if you train your machine learning model with garbage data, you have baked bad data into the underlying algorithm. Feed good data

into a neural network trained on garbage and your results may be inaccurate. Alternately, you could train your neural network on great data and then run garbage data through the well-trained algorithm. Either way, your output will be questionable.

Mikey Shulman is the head of machine learning for Kensho Technologies, acquired by S&P Global last year. He insists that his teams develop a deep understanding of the datasets they use.

“A lot of this comes down to the machine-learning practitioner getting to know the data and getting to know the ways in which it’s flawed,” he says.

“Recently we were using natural-language processing with a neural network to identify companies from text. The problem is that every time an article mentions General Electric, it mentions the company name and then the ticker symbol — General Electric (GE). If you use this data to train the algorithm, well, that’s ‘garbage in’ because now it doesn’t count any mentions of General Electric that don’t have a ticker symbol next to them,” Mr. Shulman explains. ”To make matters worse, the algorithm will fail to identify any private company in the world. It’s not garbage per se. It’s still good data. But the person looking at the data missed this and put it into the neural network and then the neural network learns to look for that pattern because it’s lazy.”

Unstructured, Semi-structured, Badly Structured

One of our machine-learning projects at S&P Global and Kensho is to use natural-language processing to pull financial data and sentiment from investor calls. While these calls typically to follow a set format, the jargon and acronyms tend to vary greatly by sector. We have learned to train different algorithms for each sector and accommodate for cultural differences depending on the speaker’s country of origin. To take a single example, in some sectors the word “units” refers to money, in others to goods sold, and in still others it can refer to divisions. An algorithm trained on a single sector would badly misinterpret this data.

Poorly labeled or unstructured data is the most common form of garbage

Poorly labeled or unstructured data is the most familiar form of garbage in any organization. The problem with badly labeled or unstructured data is that machine learning can only distinguish the signal from the noise when it already has a good idea what the signal looks like. To train a neural network to recognize patterns you need data that is well-labeled enough to identify success. If I am attempting to train a neural network to identify companies that are likely to beat their earnings guidance, I need (at the very least) a dataset that consistently identifies companies, guidance, results, and the myriad of contributing factors that may help the network detect patterns.

Sometimes data can look highly structured date but suffer from inconsistencies. S&P Global’s “Codex Project” attempts to significantly enhance how our clients interact with textual content — including publicly available findings — on our Market Intelligence platforms. At first glance, these types of filings look highly structured and quite similar. But the underlying encoding of the data isn’t structured. Each company is using very different formats or even different terminology for the same data. To pull the data successfully, we’ve had to develop a range of algorithms.

Sometimes the “garbage in” is the query itself. A dataset may be perfectly adequate to answer queries within its scope, but the machine-learning specialist frequently finds herself being asked to answer questions for which the dataset has no labels. A good rule of thumb: If your query includes terms for which there is no consistently labeled category in the dataset, it’s a garbage query. Always make sure that objects and entities are tightly defined. Machine learning is amazing technology, but it can’t find things that aren’t there.

Mr. Shulman cautions against any simple definition of garbage data.

“There is sometimes badly labeled data that is not garbage and well-labeled data that is garbage. So much of this is application dependent,” he says. “You can’t just say data is garbage in a vacuum — you have to say that the data is garbage for this application.”

Bias is Blind

There’s also the issue of bias. Media coverage of machine learning tends to focus on rare

examples of obvious bias yielding poor outcomes. Unfortunately, amusing anecdotes aside, bias is rarely apparent or straightforward. Every dataset inevitably contains some bias — not necessarily the overt bias of racism or sexism, but rather the subtle bias of existing conditions. Bias in a dataset only matters if it is germane to the question being answered. A dataset of information on America's prison population may contain race and class-based bias, but that won't matter if you're just using it to predict which desserts are likely to be the most popular among prisoners. The garbage in only produces garbage out if you're asking about the garbage.

Bias isn't limited to what is included in a dataset. Bias may also be the things that are left out. A group of inter-disciplinary researchers at S&P Global recently began looking at how women in the C-suite and on company boards affect performance. While the results have been illuminating, the team first needed to overcome the fact that, despite the incredible scope of our biographical data on corporate officers, we had no information on their gender. As we begin to use machine learning methods to track gender in the C-suite, we are identifying performance advantages to gender parity that we would never have seen in a "genderless" dataset.

Chase Ponti is the senior engineering manager of the data team at Kensho. He believes that bias is inevitable if you don't put in the upfront work on your data.

"By definition, bias is blind," he says. "You can create bias by thinking the data is structured enough. You can create bias because you don't have enough domain experts reviewing the data. You can create bias because the context under which the data was collected has changed."

Garbage in Quantity

But garbage isn't a matter of quality alone. The secret pain of everyone who works with machine learning is constantly being asked to extrapolate from insufficient data. Garbage is also a function of quantity. A given dataset may be free of bias, well-labeled and accurate. But if you are only working from a few hundred datapoints, you will struggle to train a neural net effectively.

Returning to the issue of women's research at S&P Global, our teams have had to greatly expand the number of companies we look at in order to correlate gender to performance for the simple reason that there aren't many women in the C-suite. When you consider that there are only 24 female CEOs in S&P 500 companies, it's easy to see how a single star performer or a single weak link could invalidate the results.

"Garbage Out"

Given the work and resources that machine-learning methods demand, it's tempting to accept the output uncritically. This is a mistake. It's important to understand the types of input errors and query errors that can lead to false or biased results. Identifying "garbage" in your output requires both a general skepticism when evaluating results and a knowledge of best practices in data science.

General skepticism and a knowledge of best practices are essential

Correlation Without Causation

In academia, combing through vast data sets in search of spurious correlations is known as "fishing." Yet, frequently the media reports on these naïve correlations as serious insights. Any dataset of sufficient size, will be littered with meaningless correlations that lack any causal relationship.

According to Mr. Shulman: "A lot of what machine learning comes down to in practice, is making sure that we haven't done this. Especially for us, especially in financial services where we have these rich data sets."

Machine learning isn't a truth engine. It might just spot meaningless correlations faster than humans could (or should). How then can you know when a correlation is spurious? The simplest way is to ask how many variables were evaluated for correlation. A meaningful correlation is the result of testing a hypothesis. If you begin with a belief that a causal relationship may exist between two factors and then test that belief, your results will be more meaningful. That doesn't mean that all correlations resulting from a hypothesis are causal or that all unforeseen

correlations aren't. It just means that you are statistically less likely to be fooled if you follow this simple guidance.

Unfalsifiable Results

If a correlation is judged to be causal, then necessarily its inverse must be both uncorrelated and not causal. Otherwise, you're not really proving anything. Let's assume that our sentiment analysis of quarterly investor calls showed a correlation between certain words and phrases and future company performance. If the phrase "we anticipate a change in consumer sentiment" is correlated with poor performance in the coming quarter, then necessarily the phrase "we anticipate that consumer sentiment will remain the same" (or similar) shouldn't be correlated with poor performance. If both phrases are correlated with poor performance then you might assume that any mention of consumer sentiment is correlated with poor performance. But what if not mentioning consumer sentiment was also correlated with poor performance? If everything is correlated with poor performance, the only thing your algorithm has discovered is a recession. (Please Note: this is only a hypothetical example and should never applied to investment decisions by anyone, ever.)

This concept isn't new. Famed philosopher of science Karl Popper introduced the concept of falsifiability as a cornerstone of scientific truth and knowledge in the 1950s. It is doubly useful for machine-learning users. First, it allows us to easily measure a correlation against its opposite. Not in the naïve sense of simply adding a negation to the original correlation. But by actively considering the totality of circumstances in the original correlation and defining its inverse accurately, given conditions. Thus, the inverse of "up" may be "down" or "flat" or "under" depending on circumstances.

Karl Popper introduced the concept of falsifiability

Second, falsifiability is useful because certain correlations involve conditions so practically meaningless that they can't be falsified. If I establish that mentioning consumer sentiment in an investor call is correlated with future movement in the stock price, I have said nothing. In nearly 100% of cases, there is movement in stock prices over time. This sounds like a correlation, but there is no way to test its inverse — that mentioning consumer sentiment in an investor call was correlated with future stasis in stock price — because that outcome can't happen.

General Skepticism

Despite being superior to often-biased human judgement or conventional data analysis, machine learning still demands a healthy degree of skepticism. Often, the time to apply this skepticism is at the outset of a project.

As Mr. Pont says: "The first task on most machine-learning projects is getting subject matter experts to analyze the data before you apply machine-learning methods. This is the opportunity to identify the errors in the data or the biases that will lead to a bad result. Only someone who really understands the data and the circumstances under which the data was collected will be able to identify issues before you begin."

Mr. Shulman has a practical exercise for applying human judgement to a data set.

"When you train neural networks, you need to do a few of the tasks you're asking of the algorithm to really understand it. Because if you don't understand it, it's going to be really hard for you to teach a machine to do it," he says.

Mr. Cafferillo played an active role in bringing the Kensho team into S&P Global. He sees the need for human judgment and expertise as a huge advantage for the organization.

"As a company, we not only have the data in a highly organized and clean format, but we also have the business knowledge, curiosity and expertise to truly use the data in interesting ways," he says, adding that the commercial application of machine learning at S&P Global involves both machine-learning engineers and subject matter experts in different markets and sectors to evaluate the data.

Despite a world of good intentions and careful review of data by subject matter experts, there

are times when the output of machine learning just seems like garbage. It doesn't pass the smell test. The discomfort engendered by the difference between expected results and actual results is sometimes dismissed as a product of human cognitive bias — "you can't handle the truth."

This general skepticism should be used cautiously in evaluating your results. It's a backstop against massive error and not an excuse to dismiss inconvenient results. But if the data feels wrong, that feeling shouldn't be ignored. If it looks like a duck and quacks like a duck don't allow your faith in an algorithm to convince you it's not a duck. Ask questions. Dig deeper. Maybe it isn't a duck, but you will have more confidence in your results if you begin from a place of skepticism.

Beyond Dataism — Learn to Love the Garbage

"Deriving meaning out of data with machine learning isn't a purely automated process yet. Maybe some day it will be. I hope so. But right now, it is as much a human process as a technological one," says Chase Ponti.

The errors people make with machine learning — the funny ones that become cautionary tales for the industry — are usually the result of an exclusive faith in the technological side of the process.

"You will never anticipate all of the ways in which your data will go wrong," says Mr. Shulman. "You think you have caught every edge case that your data providers will give you, but you haven't. If you come at it from that mentality, you realize that you need to understand your data first, before you rush your machine-learning algorithm."

This doesn't mean that machine learning isn't valuable. It's a crucial tool that S&P Global is using to derive insights on multiple sectors and markets. It just isn't flawless. To use machine learning effectively, you need to embrace the potential for garbage and anticipate it.

"This is a big portion of the 'garbage in, garbage out' thing — machine learning algorithms are lazy. That's what they're supposed to do," Mr. Shulman says. "They're going to take advantage of the data that you give them."

The views and opinions expressed in this piece are those of the author(s) and do not necessarily represent the views of S&P Global.

These materials have been prepared solely for information purposes based upon information generally available to the public and from sources believed to be reliable. No content (including index data, ratings, credit-related analyses and data, research, model, software or other application or output therefrom) or any part thereof (Content) may be modified, reverse engineered, reproduced or distributed in any form by any means, or stored in a database or retrieval system, without the written permission of S&P Global or its affiliates (collectively, S&P Global). The Content shall not be used for any unlawful or unauthorized purposes. S&P Global and any third-party providers, (collectively S&P Global parties) do not guarantee the accuracy, completeness, timeliness or availability of the Content. S&P Global Parties are not responsible for any errors or omissions, regardless of cause, for the results obtained from the use of the Content. THE CONTENT PROVIDED ON "AS IS" BASIS. S&P GLOBAL PARTIES DISCLAIM ANY AND ALL EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, ANY WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE OR USE, FREEDOM FROM BUGS, SOFTWARE ERRORS OR DEFECTS, THAT THE CONTENT'S FUNCTIONING WILL BE UNINTERRUPTED OR THAT THE CONTENT WILL OPERATE WITH ANY SOFTWARE OR HARDWARE CONFIGURATION. In no even shall S&P Global Parties be liable to any party for any direct, indirect, incidental, exemplary, compensatory, punitive, special or consequential damages, costs, expenses, legal fees, or losses (including, without limitation, lost income or lost profits and opportunity costs or losses caused by negligence) in connection with any use of the Content even if advised of the possibility of such damages.

S&P Global's opinions, quotes, and credit-related and other analyses are statements of opinion as of the date they are expressed and not statements of fact or recommendations to purchase, hold, or sell any securities or to make any investment decisions, and do not address the suitability of any security. S&P Global assumes no obligation to update the Content following publication in any form or format. The Content should not be relied on and is not a substitute for the skill, judgement and experience of the user, its management, employees, advisors and/or clients when making investment and other business decisions. S&P Global keeps certain activities of its divisions separate from each other in order to preserve the independence and objectivity of their respective activities. As a result, certain divisions of S&P Global may have information that is not available to other S&P Global divisions. S&P Global has established policies and procedures to maintain the confidentiality of certain non-public information received in connection with each analytical process

Copyright © 2019 by S&P Global Inc. All rights reserved.