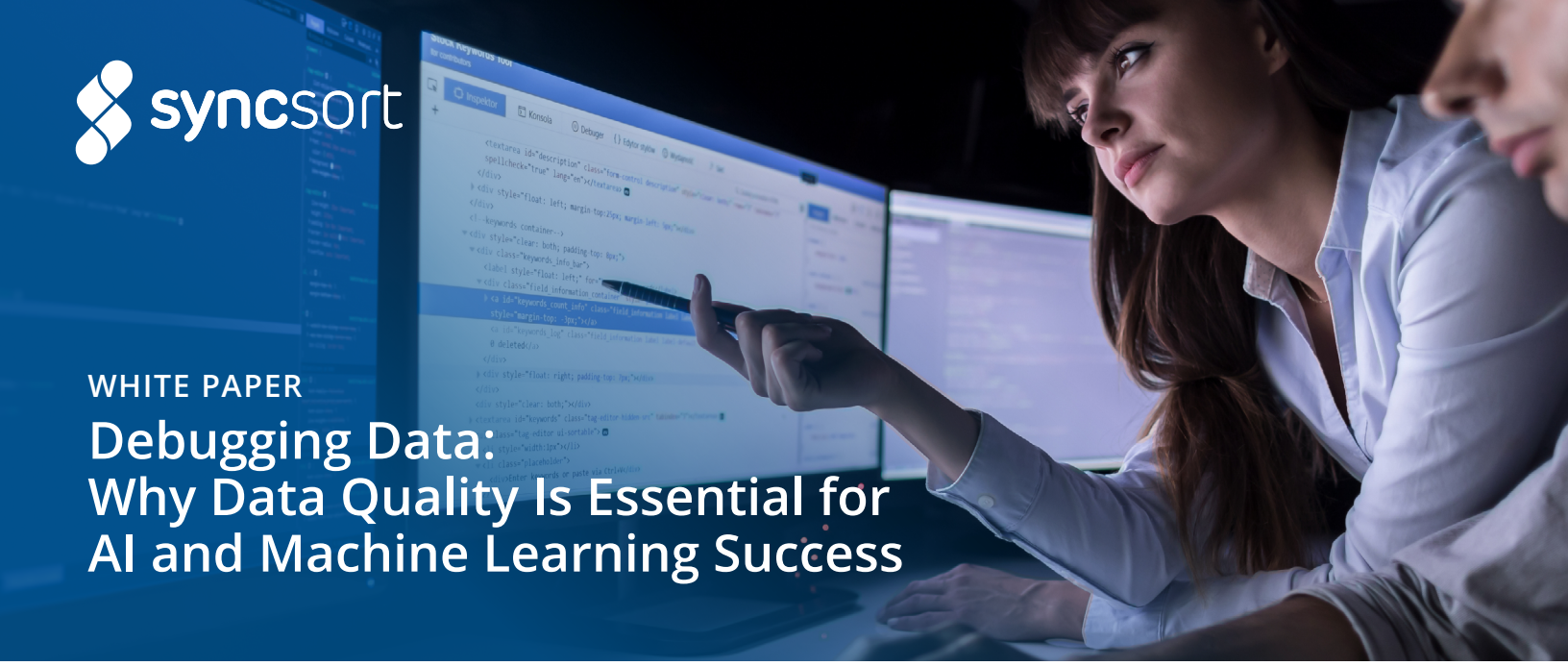

WHITE PAPER

Debugging Data: Why Data Quality Is Essential for AI and Machine Learning Success

Introduction

In most applications we now use, data is retrieved by the source code of the application and is then used to make decisions. The application is ultimately affected by the data, but source code determines how the application performs, how it does its work, and how the data is used.

But today, in the world of AI and machine learning, data has a new role – becoming essentially the source code for machine-driven insight. With AI and machine learning, the data is the core of what fuels the algorithm and drives results. Without a significant quantity of good quality data related to the problem, it's impossible to create a useful model.

The algorithms find signals in the data that are then used to make predictions and take actions. If the model is trained on different data, the predictions and actions will be different.

In addition, AI and machine learning are about unlocking the secrets of complex data. Frequently, AI and machine learning techniques find signals that are hidden inside variations and patterns that no human could ever detect on his or her own.

When you consider the role of data in AI and machine learning, a thorny problem emerges. On the one hand, it is clear that having as much data as possible that is of high quality will make AI and machine learning algorithms work better. But it is also clear that because the signal is hidden deep inside the data and can only be revealed by algorithms, it is not always straightforward to see how we can clean such data to improve its quality without obscuring the signal.

Additionally, there are concerns about how well you can reproduce your results or explain them. AI and machine learning algorithms are complex making it difficult to demonstrate why a prediction was made – making the users of the predictions skeptical of their accuracy or fairness. And if you adjust the data, it may dramatically change the outcomes and introduce bias or error.

Since data is now the instruction set for predictions – not code – another way to state this problem is: How can we debug our data?

This white paper examines the numerous challenges to having quality data for AI and why companies should be aware of these issues when trying to create accurate AI and machine learning algorithms.

The Problem

AI and machine learning need clean data to function properly. Yet companies have to figure out how to ensure their data is of the highest quality and is clean, without damaging the data patterns in the process.

To get at the heart of this challenge, companies cannot be passive with how they approach the data fed into their AI and machine learning. They have to deliberately work through a number of steps to get the results they want:

1. Identify the problem the business is trying to solve by using AI and machine learning.
2. Determine the hypothesis for that problem.
3. Find the data needed to solve the problem.
4. Examine whether the data is biased, accurate, or missing key information, and if there is enough data that shows the desired pattern.

This is not always an easy task. Context and understanding of the industry or domain from which the data originated is crucial to ensuring data quality. Entity resolution, and de-duplication when combining multiple data sets, are particularly important challenges. For instance, in the realm of healthcare, a John Doe entry is not usually the name of a person, but an identifier for an unknown male who has received actual services for an actual condition. It's thus not as simple as just purging these entries from the records (let alone linking numerous John Does together).

As another example, in the IoT realm, a sensor may display a reading of -200 degrees Fahrenheit. But this may be an error code rather than a temperature. Having the context to understand the data is thus instrumental to ensuring it is accurate.

We usually think of data quality processes as a method to remove defects from data. But in a machine learning and AI context, how can you tell what's a defect and what's not? In advance and often even in retrospect, it's difficult to know which features of the data were actually useful to the algorithms.

As a result, one should be careful about applying standard data quality methods before knowing where the extracted signal is coming from.

Missing Data

Companies have to be hyper-aware of data that might possibly be missing from a data set they are going to use. Zip codes could be missing from shipping orders. Time of purchase might be missing from in-store receipts. Regardless of what is missing, companies must be aware of how they normalize or correct this missing information – as soon as a decision is made on how to move forward with or without the missing data, bias may be introduced into the AI and machine learning algorithm results based on the decision-maker's understanding of the problem. Users of the algorithms thus must be vigilant about how they are influencing data and affecting data quality. That is true whether the missing data is corrected or not – bias can be part of the data set either way.

Being Aware of Bias

When dealing with data quality, companies must be on guard because there are so many ways to introduce bias into AI and machine learning:

- **Innate to the data:** There can be bias in the data set from the get-go. An example of this is using police data about past high-crime areas to predict where crime will occur in the future. Training AI on this data could inadvertently reinforce any racial or socioeconomic prejudices that caused those areas to be the focus for so many arrests.
- **Data that is incorrect in a consistent way:** For instance, data entry staff at one facility may have left marital status blank when entering orders for people whose salutation is Mr. but included it for those who used Mrs. or Ms. If such data is used for training AI or machine learning, preferences for certain products might be falsely correlated with marital status.

Each of these instances must be taken on a case-by-case basis to determine what the appropriate remedy is for improving the data quality.

Accurate, but Irrelevant

Additionally, companies must be aware that a data set could be entirely accurate but completely irrelevant to their desired hypothesis. For example, to return to the issue of crime statistics, data about past arrests for white collar and blue collar crime would largely lead to predictions that future crimes would occur in a city's richest and poorest locations, respectively. This is entirely accurate information, but it does nothing to inform the analyst about where future crimes will occur. Or, when examining the response to a disaster, if AI used social media content to determine which areas were hardest hit by a cataclysmic event, the results would largely be misleading, as no data about areas without power, where people could not post to social media, would have been captured.

Thus, when determining data quality for AI, the **completeness** of the data set (in the broad sense of completeness as the potential pool or population of data) is essential to determining whether the AI can produce accurate results. Users must be sure that their data sets have all the signals they need.

Deliberate Introductions of Bias

In the cases covered thus far, data quality has been an issue despite the good faith efforts of the users involved to have the cleanest data possible. But there are also cases where bias is deliberately introduced to influence the AI, and users and companies must be aware of this type of problem.

A notable example of this is the prevalence of bots on social media. Hackers and even nation states are using bots disguised as the accounts of real people to influence information on social media. If AI then relied on the most popular comments from social media to make predictions, the results would be highly inaccurate.

Past Isn't Always Precedent

Using historical data can introduce bias into predictions about the future. For instance, if a company were examining the best people to hire based on past data sets about successful employees, there might be a significant bias towards men, as hiring practices in the past often discriminated against women.

What's key to all this is that companies must do everything they can to have clean, accurate data. One clear lesson of the big data era has been that relying on bad data means you end up with unreliable results. That's why even in a time when AI and machine learning can greatly streamline the ability of a company to find signals in big data, companies must focus even more on ensuring they have quality data than they have in the past. No one would expect code to just naturally be without bugs. Similarly, good data quality doesn't just happen. It must be consciously and carefully considered and created.

The Road to Quality Data

Once a company has identified the many ways in which the data they're using for AI could be compromised, biased, missing information, or inaccurate, the next step is to decide what can be done to the data set to make it as accurate as possible. There are six key things companies can do to improve their data quality.

Apply Standard Techniques

As a rule, improving data quality is more likely to boost real signals than to increase noise. Standardizing data so that it appears in the correct fields, for instance, provides a much better signal. Entity resolution can mean the difference between machine learning looking at Bob Smith, Dr. Robert Smith, and Rob Smith as three different people versus having a richer set of information to analyze about Dr. Robert Smith.

Data Profiling

Profiling data sets can help companies determine whether using the data for machine learning will produce accurate results. Profiling means exploring and understanding the available data. This is when companies should look to see if there are missing values, clear indications of bias, or signs the data may be irrelevant to the issue at hand. Profiling can also provide indications of how complete the data set is as a whole.

Dependency and Correlation Analysis

To determine the accuracy of a data set prior to using AI on it, companies can also run dependency and correlation analyses. They can then investigate the dependencies and correlations to uncover those that are spurious and not actually signs of causation, or correlations that may produce overfitting of training models.

Matching and Deduplication

Companies can run analyses to determine whether they have sets or groups of data that may be comparable. This can indicate applying matching techniques, where you can then incorporate more, correlated data into the machine learning, or deduplication, where data that is repetitive and thus might exaggerate results is excised prior to use.

Introspection

Prior to changing any data sets, companies should stop and consider how those changes will affect the overall output of the data. Will they add in bias? If so, in what way? On the other hand, will data standardization help eliminate existing bias? Companies should ask themselves whether they are really testing what they intended to test with the data they have available for machine learning.

Enrichment

Once a data set has been profiled, and the level of its completeness determined, companies can then decide whether to enrich it with additional data (e.g., demographics, firmographics, geospatial). For instance, a company may know that using zip code data on a particular set of orders does not provide all that much insight. However, if they can incorporate latitude and longitude for the delivery addresses into the original data set, this type of enrichment could enable location analysis and identification of geospatial patterns. Enrichment gives companies the ability to see the data in new ways.

Bringing It All Together

Ultimately, there's no way around the need for companies to ensure they have the highest quality data possible to get the best results from their AI and machine learning. While the process of identifying biases present in the data, as well as the completeness of it, can be complex, it's an essential step towards debugging the data that underlies machine learning predictions. And most importantly, improving data quality is worth the effort – it will likely boost important signals that can guide business decisions.

About Syncsort

Syncsort is the global leader in Big Iron to Big Data software. We organize data everywhere to keep the world working – the same data that powers machine learning, AI and predictive analytics. We use our decades of experience so that more than 7,000 customers, including 84 of the Fortune 100, can quickly extract value from their critical data anytime, anywhere. Our products provide a simple way to optimize, integrate, assure and advance data, helping to solve for the present and prepare for the future. Learn more at syncsort.com.

syncsort