MITSloan
Management Review

The Machine Learning Race Is Really a Data Race

Organizations that hope to make AI a differentiator need to draw from alternative data sets — ones they may have to create themselves.

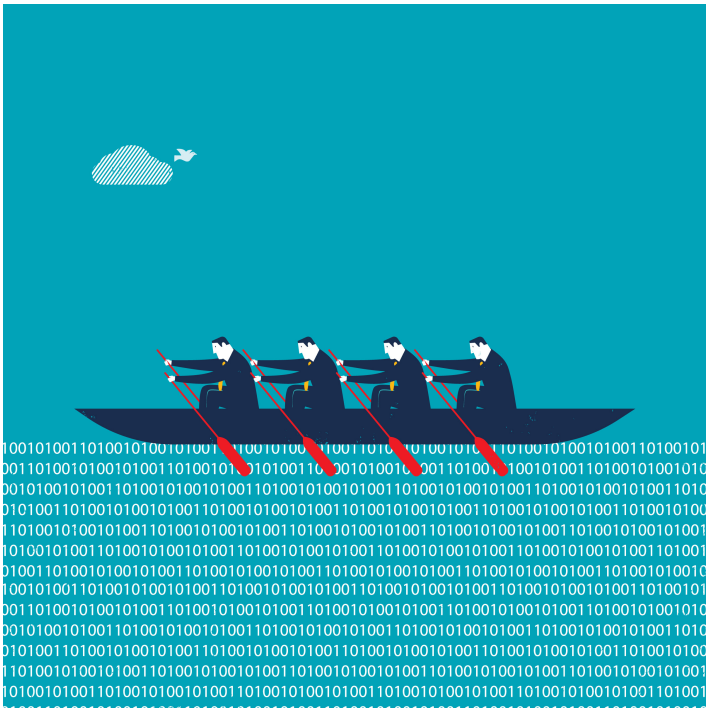
Megan Beck
Barry Libert

• Reprint #W63499 • https://mitsmr.com/2ZTfQU6

The Machine Learning Race Is Really a Data Race

Megan Beck and Barry Libert

Organizations that hope to make AI a differentiator need to draw from alternative data sets — ones they may have to create themselves.



Machine learning — or artificial intelligence, if you prefer — is already becoming a commodity. Companies racing to simultaneously define and implement machine learning are finding, to their surprise, that implementing the algorithms used to make machines intelligent about a data set or problem is the easy part. There is a robust cohort of plug-and-play solutions to painlessly accomplish the heavy programmatic lifting, from the open-source machine learning framework of Google's TensorFlow to Microsoft's Azure Machine Learning and Amazon's SageMaker.

What's not becoming commoditized, though, is data. Instead, data is emerging as the key differentiator in the machine learning race. This is because good data is uncommon.

Useful Data: Both Valuable and Rare

Data is becoming a differentiator because many companies don't have the data they need. Although companies have measured themselves in systematic ways using generally accepted accounting principles for decades, this measurement has long been focused on physical and financial assets — things and money. A Nobel Prize was even awarded on capital asset pricing in 2013, reinforcing these well-established priorities.

But today's most valuable companies trade in software and networks, not just physical goods and capital assets. Over the past 40 years, the asset focus has completely flipped, from the market being dominated by 83% tangible assets in 1975 to 84% intangible assets in 2015. Instead of manufacturing coffeepots and selling washing machines, today's corporate giants offer apps and connect people. This shift has created a drastic mismatch between what we measure and what actually drives value.

The result is that useful data is problematically rare. There is a growing gap between market and book values. Because of this gap, companies are racing to apply machine learning to important business decisions, even replacing some of their expensive consultants, only to realize that the data they need doesn't even exist yet. In essence, the fancy new AI systems are being asked to apply new techniques to the same old material.

Just like people, a machine learning system is not going to be smart about any topic until it has been taught. Machines need a lot more data than humans do in order to get smart — although, granted, they do read that data a lot faster. So, while there is a visible arms race as companies bring on machine learning coders and kick off AI initiatives, there is also a behind-the-scenes, panicked race for new and different data.

In finance, for instance, alternative data reaches beyond the traditional Securities and Exchange Commission reports and investor presentations that influence investment decisions. Alternative data, such as social media sentiment or number of patents awarded, is essential for two important reasons. First, traditional data focuses on traditional assets, and that isn't expansive enough in the age of intangible assets. Second, there's no reason to bother using machine learning to study the same data sets that everyone else in the market is analyzing. Everyone who is interested has already tried to correlate industry trends, profit margins, growth rates, earnings before interest and taxes, asset turnover, and return on assets — along with the more than 1,000 other commonly reported variables with shareholder return.

Looking for connections among the same sets of material that everyone else has isn't going to help companies win. Instead, organizations that want to use AI as a differentiator are going to have to find relationships between *new* data sets — data sets they may have to create themselves to measure intangible assets.

Curate Carefully: What Do You Want to Know?

Data creation is more complex than simply aggregating point-of-sale or customer information and dumping it into a database: Most organizations mistakenly believe that an expedient path involves gathering every scrap of possible data and painstakingly combing through it all hoping to find a glimmer of insight — the elusive feature that predicts or categorizes something they care about.

While machine learning can occasionally surprise us with a flash of rare brilliance that no one has yet to discover, the technology isn't capable of providing these insights with consistency. This doesn't mean the tool is broken. It means we have to apply it wisely. This is easier said than done: For instance, in our research of the alternative data market, we found that more than half of new data providers are still focused on measuring physical and financial assets.

The step that many organizations omit is creating a hypothesis about what matters. Where machine learning really excels is taking an insight that humans have — one based on rules of thumb, broad perceptions, or poorly understood relationships — and developing a faster, better understood, more scalable (and less error-prone) method for applying the insight.

In order to use machine learning in this way, you don't feed the system every known data point in any related field. You feed it a carefully curated set of knowledge, hoping it can learn, and perhaps extend, at the margins, knowledge that people already have.

Insightful Machine Learning Comes From Different Data

All this has three specific implications for companies wanting to create impactful and valuable machine learning applications:

- **Differentiated data is key to a successful AI play.** You won't uncover anything new working with the same data your competitors have. Look internally and identify what your organization uniquely knows and understands, and create a distinctive data set using those insights. Machine learning applications do require a large number of data points, but this doesn't mean the model has to consider a

wide range of features. Focus your data efforts where your organization is already differentiated.

- **Meaningful data is better than comprehensive data.** You may possess rich, detailed data on a topic that simply isn't very useful. If your company wouldn't use that information to help inform decision-making on an ad hoc basis, then that data likely won't be valuable from a machine learning perspective. An expert machine learning architect will ask you tough questions about which fields really matter, and how those fields will likely matter to your application of the insights you get. If these questions are difficult to answer, then you haven't put in the thought needed to produce practical value.

- **What you know should be the starting point.** Companies that best use machine learning begin with a unique insight about what matters most to them for making important decisions. This guides them about what data to amass, as well as what technologies to use. An easy place to begin is to scale and grow a piece of knowledge that your team already has and that could create more value for the organization.

It's clear that software has eaten the world (a phrase coined by software entrepreneur Marc Andreessen). But it is still hungry! Software needs a steady diet of new data combined with new technologies to continue adding value.

You don't want to be left behind by this shift in insights, machines, and alternative data. Start looking internally to identify your unique perspective and the valuable, alternative data you could and should produce. It's from those steps that you'll discover the related insights to keep your organization competitive.

About The Authors

Megan Beck (@themeganbeck) is cofounder and chief product officer of OpenMatters, a machine learning company. Barry Libert (@barrylibert) is CEO of OpenMatters and a senior fellow at Wharton's SEI Center.

MITSloan Management Review

PDFs ■ Reprints ■ Permission to Copy ■ Back Issues