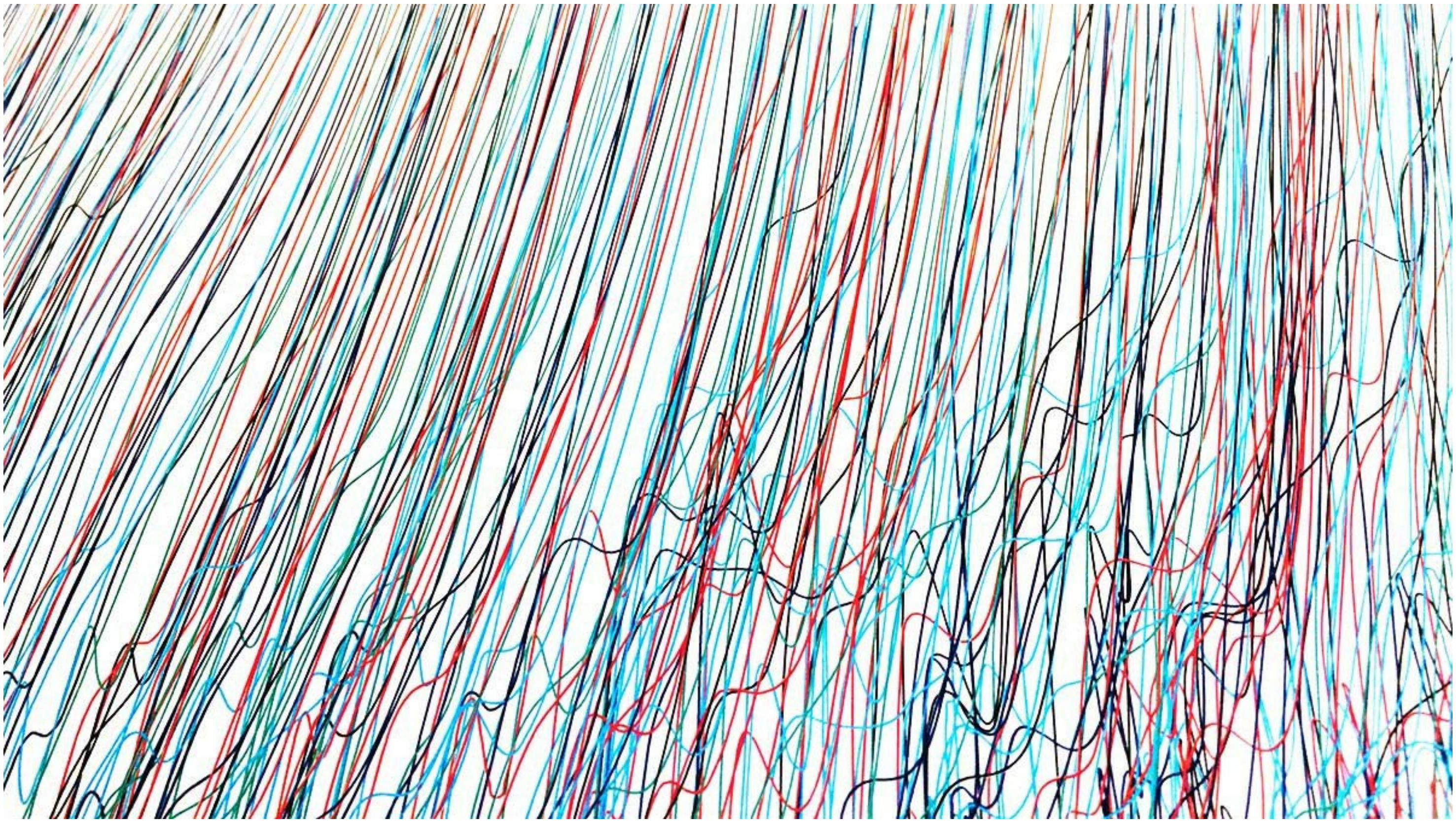
Why You’re Not Getting Value from Your Data Science

hbr.org/2016/12/why-youre-not-getting-value-from-your-data-science

12/7/2016



Businesses today are constantly generating enormous amounts of data, but that doesn’t always translate to actionable information. Over the past several years, my research group at MIT and I have sought answers to a fundamental question: What would it take for businesses to realize the full potential of their data repositories with machine learning?

As we worked to design machine learning–based solutions with a variety of industry partners, we were surprised to find that the existing answers to this question often didn’t apply. Why?

First, whenever we spoke with machine learning experts (data scientists focused on training and testing predictive models) about the most difficult part of their job, they said again and again, “the data is a mess.” Initially, taking that statement literally, we imagined it referred to well-known issues with data — missing values or a lack of coherence across databases. But as we dug deeper, we realized the problem was slightly different. In its rawest form, even clean data is too overwhelming and complex to be understood at first glance, even by experts. It has too many tables and fields and is often collected at a very high granularity (for example, online clickstreams generate new data with every click, and sensor data is collected at 125 observations per second). Machine learning experts are used to working with data that’s already been aggregated into useful variables, such as the number of website visits by a user, rather than a table of every action the user has ever taken on the site.

Insight Center

At the same time, we often heard business experts complain that “we have a lot of data and we are not doing anything with it.” Further investigation revealed that this was not strictly correct either. Instead, this frustration stems from two problems. For one thing, due to the time it takes to understand, formulate, and process data for a machine learning problem, machine learning experts often instead focus on the later parts of the pipeline—trying different models, or tuning the hyperparameters of the model once a problem is formulated, rather than formulating newer predictive questions for different business problems. Therefore, while business experts are coming up with problems, machine learning experts cannot always keep up.

For another, machine learning experts often didn’t build their work around the final objective—deriving business value. In most cases, predictive models are meant to improve efficiency, increase revenue, or reduce costs. But the folks actually working on the models rarely ask “what value does this predictive model provide, and how can we measure it?” Asking this question about value proposition often leads to a change in the original problem formulation, and asking such questions is often more useful than tweaking later stages of the process. At a recent panel filled with machine learning enthusiasts, I polled the audience of about 150 people, asking “How many of you have built a machine learning model?” Roughly one-third raised their hands. Next, I asked, “How many of you have deployed and/or used this model to generate value, and evaluated it?” No one had their hand up.

In other words, the machine learning experts wanted to spend their time building models, not processing massive datasets or translating business problems into prediction problems. Likewise, the current technological landscape, both commercial and academic, focuses on enabling more sophisticated models (via Latent variable models), scaling model learning algorithms (via distributed compute), or fine-tuning (via Bayesian hyper optimization)—essentially all later stages of the data science pipeline. However, in our experience, we found this focus to be misplaced.

If companies want to get value from their data, they need to focus on accelerating human understanding of data, scaling the number of modeling questions they can ask of that data in a short amount of time, and assessing their implications. In our work with companies, we ultimately decided that creating true impact via machine learning will come from a focus on four principles:

Stick with simple models: We decided that simple models, like logistic regression or those based on random forests or decision trees, are sufficient for the problems at hand. The focus should instead be on reducing the time between the data acquisition and the development of the first simple predictive model.

Explore more problems: Data scientists need the ability to rapidly define and explore multiple prediction problems, quickly and easily. Instead of exploring one business problem with an incredibly sophisticated machine learning model, companies should be exploring dozens, building a simple predictive model for each one and assessing their value proposition.

Learn from a sample of data—not all the data: Instead of focusing on how to apply distributed computing to allow any individual processing module to handle big data, invest in techniques that will enable the derivations of similar conclusions from a data subsample. By circumventing the use of massive computing resources, they will enable the exploration of more hypotheses.

Focus on automation: To achieve both *reduced time to first model* and *increased rate of exploration*, companies must automate processes that are normally done manually. Over and over across different data problems, we found ourselves applying similar data processing techniques, whether it was to transform the data into useful aggregates, or to prepare data for predictive modeling—it’s time to streamline these, and to develop algorithms and build software systems that do them automatically.

This acute understanding of how data scientists interact with data and where the bottlenecks are led us to launch “The Human-Data Interaction Project” at MIT, focusing on the goals listed above. Our target is rapid exploration of predictive models, and to actually put them to use by solving real problems in real organizations. These models will be simple, and automation will enable even naive users to develop hundreds if not thousands of predictive models within hours—something that, today, takes experts entire months.