



What Managers Need to Know About Big Data

Foster Provost & Jim Euchner

To cite this article: Foster Provost & Jim Euchner (2017) What Managers Need to Know About Big Data, Research-Technology Management, 60:3, 11-17

To link to this article: <http://dx.doi.org/10.1080/08956308.2017.1300991>



Published online: 28 Apr 2017.



Submit your article to this journal [↗](#)



Article views: 12



View related articles [↗](#)



View Crossmark data [↗](#)



What Managers Need to Know About Big Data

An Interview with Foster Provost

Foster Provost talks with Jim Euchner about the accelerating growth of big data and how to profit from it.

Foster Provost and Jim Euchner

A convergence of trends—mobile data, the Internet of Things, and advances in machine learning, among others—is driving a surge in interest in big data. The tools and techniques for exploiting big data streams are applicable across a wide range of domains, from manufacturing and predictive maintenance to marketing and fraud detection. In this interview, Foster Provost, a professor of data science at New York University and author of *Data Science for Business*, discusses trends and challenges in big data and what R&D managers need to know to manage it.

JIM EUCHNER [JE]: You've been working with machine learning and large data sets since before people were calling it big data. What is it that's making data analytics so hot now?

FOSTER PROVOST [FP]: The interest has been building. Ten years ago, data analytics was an awful lot hotter than it was back in the '90s; now it's so much hotter than it was 10 years ago. There is a triumvirate of factors driving this, which include the greatly increased availability of data; the fact that computers are now fast enough to analyze it; and the maturing of the analytical technology itself.

Foster Provost is Professor of Data Science, Professor of Information Systems, and Andre Meyer Faculty Fellow at New York University's Stern School of Business. He is coauthor of *Data Science for Business*, which Fortune Magazine called "Required Reading" for MBAs. His data science research is broadly read and cited and has won many awards, including the INFORMS Design Science Award and best paper awards at top journals and conferences across three decades. He has co-founded several companies based on his research, most notably digital advertising pioneers Distillery and Integral Ad Science. Foster previously was editor-in-chief of the journal *Machine Learning* and program chair of the top data science research conference, ACM SIGKDD. fprovost@stern.nyu.edu

Jim Euchner is editor-in-chief of *Research-Technology Management* and vice president of global innovation at Goodyear. He previously held senior management positions in the leadership of innovation at Pitney Bowes and Bell Atlantic. He holds BS and MS degrees in mechanical and aerospace engineering from Cornell and Princeton Universities, respectively, and an MBA from Southern Methodist University. euchner@iriweb.org

DOI: 10.1080/08956308.2017.1300991

Copyright © 2017, Industrial Research Institute.

Published by Taylor & Francis. All rights reserved.

Back in the early '90s, you had to carry data in the trunk of your car from one place to another; you had to find someone who could read that sort of tape, and then find someone who had a powerful enough computer to be able to analyze it. Then you had to build your own algorithms. The Internet has gone a long way to solving the first problem, as systems are increasingly interconnected. Of course, computers are increasingly faster and faster, following Moore's Law, and analytic technology has followed these developments.

The truth is that many of the key algorithms were already there, even back in the '90s. As the data has become available to analyze, people have worked on making the analytic technology more mature. The exponential increase in the availability of data is really the main reason big data is hot right now.

JE: There is a lot of data available just because so much is now online: enterprise systems in corporations, click streams on the Internet, and now, the Internet of Things. It's not just that data is more conveniently available; it's that there is a lot more of it.

FP: These are two different things. It is certainly more convenient to access data now: data that previously would have been stored in an inventory database or at the point of sale is now available in enterprise systems.

But everything is also more instrumented now. The activities of people are instrumented through web-based systems and their mobile phones, so activities that had previously been invisible are now recorded. In the past, when you watched television programs, nobody saved what you watched, so they couldn't do analytics to recommend to you what to watch next. We also didn't know where people were physically. This is amazingly useful information that just wasn't recorded before smartphones and web-linked cameras and so forth.

Things are also increasingly instrumented. With the Internet of Things, we're seeing a much wider variety of the devices we use getting instrumented, from elevators to thermostats. This newly available data means that there are opportunities for a whole new range of analytical applications.



Foster Provost, author of *Data Science for Business*, discusses trends and challenges in big data and what R&D managers need to know to manage it.

JE: When you talk about computation, Moore's Law has played a big role, but are we also seeing a shift toward more parallel computation and different kinds of data structures that allow faster processing of data?

FP: Both. In the '80s when I got my master's degree in parallel computing, that was the hot thing, but very few people had a Connection Machine or a hypercube system or a network of workstations that really could work together. And, to be honest, few really were ready for that capability for data analytics. When we were working on fraud detection at NYNEX (now Verizon), those data sets were bigger than almost everything that people these days are calling big data; they were larger data sets with higher velocity, more variety, and they were more dynamic. And what did we use? We used a massive SQL system, highly optimized. It was a big Oracle system.

But once you start gathering every single click on every web page that everybody makes, you're probably going to need something more powerful than that. And maybe more importantly, you don't necessarily know at the outset

what you're going to do with the data, so you don't know how to structure it appropriately. It was really when Google started to focus on search engine performance that they essentially rediscovered and reinvented the idea of MapReduce. Then Yahoo made Hadoop and then everybody started realizing that, with these massive amounts of data, you need to do a lot of things that are largely filtering and aggregating. You can do these things amazingly fast with these big-data architectures. And that was just the beginning.

I think there is a related development that may be even more important than the availability of data, faster computing power, and the maturing of analytic algorithms. I think that it is one of the main reasons analytics is hot right now. That is that the decisions made with the analytics can now be much more easily implemented once they are developed. Prior to this web era, the data was in one system; it had to be extracted and processed by the analytical staff. Even if you had discovered a way to significantly improve the processes, you had to figure out how to deploy it. Often, you would have to build a system specifically to do that. That is a big hurdle for most organizations.

Now, everything is actually happening within one system. The data are there; the analytics can be embedded there and the decisions can be made there and then. Walmart, which is a pretty data-savvy company, may have been able to make recommendations to people, but what could they do with the information? Put it on a receipt, after the people had already checked out? Compare that with Amazon: if they make a decision or recommendation, they can actually implement it immediately and the effort is almost trivial. It is the ability to implement the analytic insights and decisions that has been the key factor in driving the excitement with big data now.

JE: That's very interesting. It's a matter of the ease of integration of the results into websites and phone systems and integrated corporate systems as much as the analytics. Can you give any examples of new types of problems that just wouldn't have been feasible to solve before? What new problems are people solving because the algorithms have gotten better and the computation speeds have gotten better?

FP: People ask me a lot, "So what's the next frontier?" I have had some success in being ahead of the curve, but to be honest, if I really knew that, I'd be talking to you from my chateau in southern France. Nonetheless, I think there's a framework for readers to be able to answer this question for themselves. It follows from that last observation: forget about the idea of analytics algorithms and advances in processing power; think instead about the availability of data and the ability you have to implement the decisions your algorithms make.

We can use this as a lens on the history of the expansion of analytics, as well as a way to look forward. Some of the earliest applications of what we would call big data analytics today focused on fraud detection, especially in the

telecom and financial sectors. Why? Certainly, the costs of fraud were significant, but that was not sufficient. I think it is because the data already were available in the system and the decisions could be implemented within the system. Take credit card fraud: credit card transactions were already getting approved, so the system was already there to automate inferences about fraud status, and also to evaluate and improve the algorithms. The same was true for telecom fraud: somebody makes a call; if your algorithm says you should block the call, you block the call. You don't have to build a new system.

JE: That suggests how managers identify good candidates for data analytics: look for where there are decisions that are already being implemented in a system and see if they might be improved with analytics.

FP: Yes. In particular, look for cases where the data to be analyzed and the point of application for the result of that analysis are both in the same system. A great case study of the precocious application of data-analytic technology is online advertising. At a company like Dstillery, real-time machine learning runs every single day and builds thousands of new models, all completely automatically. No human interaction at all is required to build and test these models, and when they improve the results, the model that's in the system is changed automatically. I don't know any other area where we have this level of automation of predictive modeling.

JE: This seems almost like meta-analytics. Why is it necessary to keep refining these models?

FP: There are at least three reasons. One is that the environment changes. The best online advertising uses ultra-fine-grain data on all the different actions people take online, for example, the websites they visit. The websites you visit are very, very revealing as to what your predilection for different brands would be. You build a model that relates the sites someone visits to the brands they might buy. But the world changes. Next week, or even tomorrow, there's a different set of websites out there; this changes pretty fast, and the models have to adapt.

It's also true that, the longer the campaign runs, the more data you have; if you have a significantly larger amount of data, why wouldn't you build a new model that would be potentially more accurate? To tell the truth, when you start a campaign, you actually don't have any of the data you really want, which is data on people who will respond to an offer after having been shown an ad. That's what you really would like. At the start of a campaign, you don't have anyone actually buying anything after they've seen a particular ad you presented to them. So you build the models based on a proxy. Then, as soon as you start showing ads, you start to gather the kind of data you need and you can update the model. This is really sophisticated, state-of-the-art data science in action.

One of the most fundamental principles I teach to managers is *not* to think about the opportunity by asking, How should I mine the data I have?

JE: That's an application where there's so much dynamism that you need a self-learning, constantly changing system. For a lot of applications, you need real-time data, but the world doesn't change that much.

FP: That's absolutely right. And sometimes the world doesn't change but the phenomenon itself is changing. In fraud detection, for example. (Or, within the enterprise, modeling for compliance monitoring for detecting employee misconduct.) As soon as you put a model in place, the phenomenon starts to change in response to the model as those committing the fraud detect the detection and find new ways of committing fraud. You need to start working on your next model as soon as you put the last model in place, so you begin gathering the data to build the next model.

JE: I've talked to people who say, "We're going to build a big data lake or we're going to build a big data repository and we're going to use data mining to find the problems we should solve." It's the discovery approach to data mining. How successful has that been?

FP: I have never seen it work—I mean where there wasn't some solid idea for a good application in advance—but I would love to read a case study of its being successful.

JE: It may just be that people want to get going and they don't know which problem to solve, so they just start by getting data.

FP: I teach in our MBA program and in our Master's of Science in Data Science program, and one of the most fundamental principles I teach to managers thinking about data analytics and data science is *not* to think about the opportunity by asking, How should I mine the data I have? You might come up with an interesting project there, but I think there are going to be a lot of false positives and a lot of false negatives.

We need to think about data as an asset on which we might get some return through data analytics. But the return is going to come through solving business problems. We should start by thinking about what are the important business problems or opportunities we have, and then ask what ways there are in which our data might help.

The reason I like to think about data as an asset is that it causes us to ask questions like, Could we invest in data to

solve this problem? The most interesting applications are not coming from people looking to what data they have lying around and how they can mine it; it's people thinking, "Oh, if I only made this investment in data and analytics, then I could do something really interesting."

For example, if I wanted to target online advertisements, I might say, "If only I could gather the data on the locations where people spend their time, that could improve my targeting. How could I build a system that could get that data? Or how might I buy that data from someone who might already have it?"

The Signet Bank story in the first chapter of *Data Science for Business* is a very nice example of this. Back in the '80s, banks were working to achieve economies of scale. There was massive consolidation. Interestingly, the banks were offering credit cards to everybody on the same terms, which seems bizarre to us now. Banks were actually *afraid* to give different terms to different people because they believed that the customers would revolt if they found out somebody else was getting better terms than they were. So two consultants went around to the banks and said, "Let us work with you to build systems to figure out what are the best terms for each customer segment." All the big banks turned them down. They finally got to little Signet Bank in Virginia, which agreed to try it.

So they started working on the problem and realized that they could not actually build models to predict whether or not some one would be a good customer for a given set of terms because they had never given customers different sets of terms! What can you do? They decided to start by giving people credit cards with different terms—they made an investment in data.

If they hadn't taken a strategic perspective that this was an investment in data, as soon as they started having higher write-offs from these experiments, the stakeholders would have stopped the program. If you randomly, or even smartly, start giving people new sets of terms, you're not going to do as well at the start as the smart people who have figured out over the years what the optimal set of people was to give the traditional set of terms to. That exact thing happened with Signet Bank—they started to have higher write-offs. The consultants probably would have been fired except that the company took the strategic perspective, that they were building a data asset.

It took a couple years, but eventually, it started turning around. They were learning who the best customers were for a given set of terms. They started to skim the cream off of all the big banks because they could find out who

were the best customers and give them better terms. They started to be more profitable and ended up becoming the biggest consumer credit issuer in the country, in the world maybe, Capital One. It was because they took a data-driven approach and because they invested in data by doing experiments during which they were willing to tolerate increased write-offs. This is powerful, but it is also costly in the short term because the returns from the initial, near-random targeting aren't nearly as high as the returns from your best model so far.

Once you think about data as an asset, it makes sense. We're used to investing in all sorts of assets. Why shouldn't we treat data the same way?

JE: What have been some of the pitfalls or critical success factors that you have seen for those applying data analytics to their businesses?

FP: Thinking that data analytics is going to be magic or easy is a big pitfall. You can't hire a couple of data scientists and expect them to solve your biggest problems. No. This is an undertaking that will take time, and you have to sit down and decide you're going to commit to it. You have to put in place the right (cross-functional) team.

Another pitfall is not scoping the project well, not thinking it through. Yet another pitfall is not having good analytical people to work on the problem, which sometimes stems from not having the capability to assess them. If you assign some people to work on a problem and it fails, was it because it couldn't be done? Was it because there's some randomness in the process, and sometimes things fail? Or was it because you put the wrong people on it? How can you even tell whether you've got good analytic people?

Experience shows that the best data scientists are orders of magnitude better than average data scientists. A below-average data scientist is actually worse than none at all: not just because they may waste money on poorly defined projects but because you may wrongly conclude that a really good problem can't be solved. We're still at the stage where sophisticated data analytics has to prove itself in many companies, so an early failure can be deadly.

JE: How do you identify and attract data analytics talent, the people who are an order of magnitude better than the average?

FP: That is the number one question firms face with data analytics. It's hard for multiple reasons. One is because it takes one to know one right now. We don't have credentials you can rely on, like an engineer who graduated from MIT and then worked for GE for 10 years. There have only been data science degrees for a few years now, and only a small number of those people have gone out and gotten the kind of experience you need to be effective.

Firms have to somehow solve this chicken-and-egg problem. If I just have a bunch of people who don't really know much about data science, how am I going to identify whether a candidate is a good data science candidate? Even

We're used to investing in all sorts of assets. Why shouldn't we treat data the same way?

if I were able to identify them, why would they come and work for me? Data scientists can choose where they're going to work, so managers have to understand what it is that data scientists value. To tell you the truth, they don't value working with a bunch of people who don't understand data science. You might attract them somehow or another but they're not going to stay.

When I say that management needs to understand data science, I don't mean understand the technical details of the algorithms. I mean that they understand the fundamental principles—the idea of treating data as an asset; the difference between building a model and applying the model; some sense of the different sorts of analytics that could be applied; some understanding of the process—especially the evolution of uncertainty over the course of an undertaking. The process for doing data science is becoming well understood; it is important that managers understand it. When you get down to it, these are the kinds of things we teach in a data science 101 course for MBAs.

Most data scientists I know also want and need to work with other data scientists. It's pretty lonely to be the only data scientist, even if the managers understand what you're doing. You want to be able to talk about your new idea with someone else and learn from people with different expertise. I think of the good data scientists, and the great ones as well, as being very deep in some area and then having a shallower knowledge of a bunch of others.

JE: You're familiar with Topcoder and Kaggle? I spoke with Karim Lakhani, a professor at Harvard, about using contests to solve hard data analytics problems.¹ Can you outsource data analytics through computational competitions?

FP: I don't think so. Not the important stuff.

What's the contest good at? The contest is good at getting a bunch of smart technical data scientists to take a data set and figure out what the best algorithmic means are for coming up with, let's say, a predictive model that does well by some measure you establish. The reason I am not a fan of these contests is that you're somehow implying that the algorithm development is the valuable part of the process. I think that if you've gotten to that point, you've pretty much solved your problem.

The hard challenge is formulating the problem in the first place, figuring out what data you need, figuring out how to invest in getting the data, finding out that those data aren't actually what you thought they were, figuring out how to properly evaluate a result, and so on. The problem formulation is, to me, the key to solving important problems—not in finding that a 100-model custom ensemble does better than a linear regression.

In the end, if you don't have people who understand modeling and people who deeply understand the business problem, you're not going to set the problem up well enough for a contest in the first place.

¹Jim interviewed Karim Lakhani for the September 2016 edition of Conversations. See "Innovating With Crowds," *Research-Technology Management* 59(5): 15–21.

The process for doing data science is becoming well understood; it is important that managers understand it.

JE: Your view, then, is that you have to have good data scientists to formulate the problem, and once they've done that, they will probably be able to find a solution that's a good one.

FP: Yes. Let me expand on that. I think what you need to do for successful data science is to build cross-functional teams to work on solving the problems together. That team should include the scientists, of course, but you also need people involved who understand all the real-world stuff that's relevant to solving the problem, people who know the constraints around the solution.

A lot of times, you have solutions created that you can't implement for some reason. Why didn't the data scientists know about that? Because they were off trying to solve the problem in a vacuum, not working closely with the people who understand the domain. Of course, you face other hurdles. You have to escape from the organizational inertia and from the mindset that you can't do a whole bunch of stuff.

Finally, you have to have people involved who know the data. Often, the people who are on the business front line don't really know the details of the data. The data scientists don't know the details of these specific data, either. There is often a separate set of people who know the data.

I believe that there's another thing you need in order to have successful projects, which is executive buy-in at a pretty high level. You're going to be chasing around the organization trying to understand the problem and the data available to you, and if you don't have engaged political support, it will take a very long time to get cooperation.

JE: Let's talk a little bit more about how managers can think through the lens of data analytics. You said that you need to think about where value can be added through better decisions, and think backwards from there. You also said that you need to think in terms of the vehicle through which the decision will be implemented. Finally, you emphasized the importance of thinking of data as an asset, as something you should be willing to invest in. What other advice for managers do you have?

FP: I think that managers need to understand that data analytics is much more like R&D than it is like most projects that managers are accustomed to managing. Even (or maybe even especially) managers that come out of

The best way to manage data science is not through mega-projects but by going through cycles of increasing investment, reducing uncertainty with each cycle.

IT want to be able to answer a set of basic questions: What will be the ROI on this project? How long is it going to take to do it? How much investment am I going to have to put into it? Can you tell me, specifically, what will come out of it?

None of these questions can be answered for the most interesting data analytics projects. This can end up being a big problem because most managers have not worked on R&D projects, and even if they have, their procedures may not be well suited to high-uncertainty projects. In R&D, you attack problems that may have a 20 percent chance of success, and you manage the risk through a portfolio approach. Most managers just don't think or operate this way.

The best way to manage data science is not through mega-projects but by going through cycles of increasing investment, reducing uncertainty with each cycle. You might not know whether you have the right data; you don't know whether you can get the right data, if you don't have it; you don't know whether, if you had the data, you'd be able to build something that would predict well the result you're after. So do an inexpensive pilot study and see whether there seems to be any promise. If so, then make a larger investment—but then maybe just a larger pilot. As you know well, Jim, this is essentially what you do in R&D.

And even if the pilot study fails, you'll know so much more after going through the cycle once that you'll be able to invest better afterward—for example, decide whether to try again now that you know what you did wrong or whether the hurdles seem insurmountable.

At some point, you may be willing to say that the preliminary results, sort of *in vitro* studies, are good enough to invest in a production pilot: who knows whether the lab environment has replicated reality as well as is necessary. If this works, you may roll the model out for a small subset of the population, randomly or by geography or whatever makes sense. You have this cyclic project structure, and once managers embrace it, they like it, because they get feedback much more quickly and they have much more interaction with the team.

Once managers understand these fundamental principles of what data analytics is all about, they can be more effective. It's hard to invest in data analytics if you don't understand the fundamentals.

JE: Who are the lead users in data analytics in the industrial space, outside what we might call new economy firms, like Google and Facebook and Netflix and Amazon? What are

the challenges that industrial firms like Goodyear and Ford and Procter & Gamble face?

FP: One that we already talked about is whether there's a system within which to execute the decision the data analytics system makes. The new economy firms generally have everything in the system, so they have an advantage there. If they come up with a successful data science model, it can be implemented. This might not be so at Ford or Goodyear. Another advantage the new economy firms have is in putting together a good data science team; a lot of cutting edge analytics is taking place at these firms. The final difference is organizational culture. At Amazon or Google or Facebook, data is king. If they could possibly do it, every decision would be made analytically. They decide on how much food to put out on the second Thursday in November based on some algorithm; everything is done with data. That is not so at most industrial firms, which were spawned well before 1998, when we were beginning to think in this way.

JE: How do managers get up to speed with data science? Do they take one of those short courses at NYU? Do they go to seminars on the application of the technology? What do they do?

FP: Well, the first thing they do is to read my book.²

JE: Okay. I'll go with that.

FP: Actually, I'm only half kidding. The book was written specifically for managers trying to understand the fundamental principles of data science. It's not an airplane book, though. It has just enough of the technical stuff but doesn't go down into the weeds. I think managers should know just enough of the fundamentals to be able to start working through problem formulation. In the end, though, problem formulation is a craft: you can't actually teach it. You need to go through the process, make mistakes, cycle back, and try again. A manager serious about learning this needs to try to work through the process on several potential applications.

Something that I've seen work is to do this learning with a group of people in your company, with a skilled facilitator, if you can find one. This cohort develops a common understanding of the principles and a common language, and they can work through problems together. It creates a sort of community that can be very helpful. And then begin the experimentation cycle.

JE: So our time is just about up. What would you say are the most important takeaways for managers?

FP: Of all this, I think the most important takeaways are probably five: 1) start by thinking of problems and opportunities in the business, and then move to what data might be

²*Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking* by Foster Provost and Tom Fawcett (O'Reilly Media, 2013).

useful—and possibly might require an investment to acquire—for example, in labor; 2) focus where decisions can actually be implemented easily, and remember that new opportunities often arise when new systems are put in place; 3) learn the fundamentals of data science, without

which at best you'll frustrate your analytics staff and at worse you'll make bad investments; 4) keep in mind that data analytics projects are like R&D projects in terms of their uncertainty profiles, so manage them that way, and 5) create an environment where analytical employees can thrive.

ATTN Job Seekers **Find a Career in** **the R&D Industry**

Free and confidential resume posting
Upload up to 5 career-related documents
Automatic email notifications
Access to job seeker resources
Save up to 100 jobs

CAREERS.IRIWEB.ORG



Get your resume noticed by the people in your field who matter the most. Whether you're looking for a new job or ready to take the next step in your career, the IRI Career Center will help you find the opportunity you've been looking for.

Visit the IRI bookstore!

Access the best IRI has to offer in electronic or print-on-demand formats. Browse *RTM* reprint collections, white papers, and special offerings, all selected and compiled to offer a range of perspectives on central issues in the management of technological innovation.

<http://www.iriweb.org/bookstore>