



“SPOCK
IN THE
BOARDROOM.”

WHERE BIG DATA
MEETS BIG VALUE



*Interview by
Alison Buckholtz*

*A discussion with DJ Patil,
Chief Data Scientist, U.S. Office of
Science and Technology Policy*



You have to be clever with data—you have to use it to tell a story.

You and your colleague, Jeff Hammerbacher, coined the term “data scientist.” It’s a job title that could only exist in a time known for Big Data. But what does a data scientist actually do?

▲ There was no desire by Jeff or me to make up a whole new field. Back in 2011, Jeff was running the data team at Facebook, and I was at LinkedIn. Despite what people think about big companies being competitive, we actually got to meet pretty frequently and trade notes. We shared the idea that the things we were building were hard, and collaboration could help us meet our goals. One thing we started talking about was how to come up with common labeling for our team members, all of whom were engaged in some area of data science. We had research scientists, statisticians, designers, front end engineers—but we needed one term that described all of them. “Analyst” sounds too Wall Street. “Statistician” makes the economists mad and vice versa. “Research scientist” sounds too academic. The term that seemed to fit best was “data scientist” because it referred to those who use both data and science to create something


new. We used a data-driven approach to test it via job postings on LinkedIn, and all the right people applied to the job postings that used “data scientist.”

The term quickly became part of the lexicon, despite its imprecision. Did that surprise you?

▲ It surprised me very much. I never expected it. The reason I think the term has taken off the way it has is because it’s ambiguous. No one knows what it means, so it gives you permission to be what you want to be. It’s empowering to have a title that allows you to control your own destiny.

Why does the role of data scientist matter for institutions and companies?

▲ As with everything in life, you can look at this through the lens of Star Trek. The first person that the Captain turns to when there is a problem is Spock. And Spock always responds to these problems the same way: he answers, in one form or another, with the word “curious.” This is the viewer’s cue that he’s going to start



figuring out how to solve the problem. For a company or institution, the goal of having a Spock on the bridge—or in the boardroom—is to solve problems. In today's world, that's the role of the data scientist. The data scientist is the person who understands things.

Has the field of data science spawned other career tracks that didn't exist before Big Data?

▲ You have to be clever with data—you have to use it to tell a story. And that's what the new field of data journalism is all about. Sites like FiveThirtyEight are using data as their weapon of choice to get to the heart of whatever story they are telling.

Is Big Data viewed differently in the public sector versus the private sector?

▲ Most people think that Big Data originated in the private sector. But among those of us

who came into this recent wave of Big Data over the past 10 to 15 years, almost all of us got our start in the federal government, particularly in national security and bioinformatics. We had these gigantic data sets and wanted to know how to use that data most effectively—whether that's preventing terrorism or trying to understand the genome. So the push started in the public sector. That's true for me as well: I worked on weather data, and that data was produced by the National Weather Service.

A data-driven organization, whether public or private, acquires data in a very sophisticated manner. Then they process it—and in doing so, they turn it into value. That value ultimately translates into an outcome. They want to use the data to build a product for the user. They don't show us data—they just get us to the goal that we want. Whether that final product is private or public, the line gets blurry. For example, a self-driven car from a for-profit company could stream data from the National Weather Service to warn passengers when there are dangerous conditions ahead.

For a company or institution, the goal of having a Spock on the bridge—or in the boardroom—is to solve problems. In today's world, that's the role of the data scientist.

”



If a government is trying to decide where to build a new bridge, data can show where the congestion is greatest, and computer simulations integrate these statistics as they model these questions. Data is a very powerful force multiplier when trying to answer these sorts of infrastructure questions.

In building large-scale infrastructure, there can be great uncertainty. Does it make sense to start collecting data even if the particulars you want to measure are not very clear?

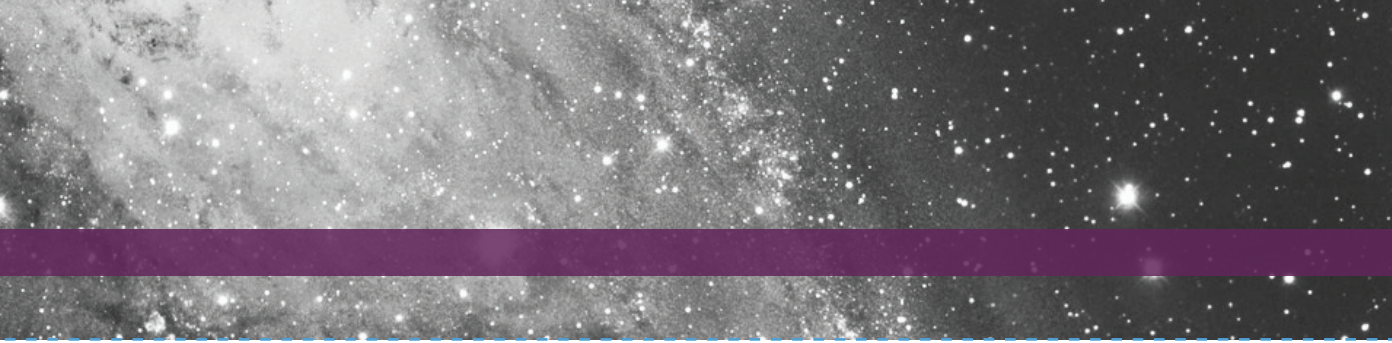
▲ When we start asking the question, “What do we need to do with the data?”—no matter what the situation—that determines what sort of data we need and how often we need to collect it. In an unpredictable situation, we measure, with increased fidelity, and we see if we can get a better understanding with targeted observations.

A lot of time, whether it’s infrastructure or not, we don’t know the question we are asking. That’s why we try to generate a large pot of data. This allows us to see the data set and identify the question. Once we know the question, we can build an operating hypothesis from

those data sets. We can look at data and from there determine what are the right questions to ask. Then the iterative process begins. Data science uses the scientific method, which has never been possible before because we haven’t had access to data of this quality.

How can data make it possible for governments to connect policy and practice when it comes to infrastructure?

▲ Let’s look at the challenge of climate change. The White House climate data initiative gathers together all the data sets that are related to climate change so if an institution is thinking about funding infrastructure, they can see in advance what the climate impact will be on that infrastructure. Previously, there was nowhere to get that data. So we can see a fast effect there. Another example is if a government is trying to decide where to build a new bridge. Data



can show where the congestion is greatest, and computer simulations integrate these statistics as they model these questions. Data is a very powerful force multiplier when trying to answer these sorts of infrastructure questions.

How can the use of data strengthen outcomes for emerging economies, especially when governments may not have the resources to generate data themselves?

▲ When you look at the UN’s Sustainable Development Goals, it’s striking that data is central to creating strategies to reach those goals, whether it’s ending poverty, enabling education, or any of the others. Let’s talk about providing efficiency in education by allowing students to learn in a non-linear fashion. Customized learning experiences like MOOCs, that may be non-linear, are one option. That’s where you use data to better track students’ learning outcomes and produce proposals for personalized educational environments.

Another important area in which we have not utilized data enough is in agriculture. The

United States Department of Agriculture has adopted “precision agriculture,” and what it’s enabled has been really incredible. It allows monitoring of soil with satellites, which in turn determines how to increase crop yield—maximizing crops while ensuring the soil stays healthy. This is a game changer for many environments. If you are using data you’re entering each year, increasing your yield by three to five percent, then you double your capacity over a number of years. That has direct implications for emerging economies that have the benefit of using data to think about the problem. And they can use someone else’s open data, they don’t have to generate their own.

Speaking of open data, how is the field of Big Data evolving?

▲ On an institutional level, in 10 to 20 years, data will become increasingly open, so you won’t have to work so hard to get the information you need. Data will be ubiquitous in our lives, adding value, and data will be safe. On a personal level, data will allow us to be proactive in our lives and in terms of our health, for better decision-making all around.

