Forbes / Tech

BETA

JAN 7, 2016 @ 01:47 AM 19,987

Big Data Uncovered: What Does A Data Scientist Really Do?

Bernard Marr, CONTRIBUTOR

FULL BIO

Opinions expressed by Forbes Contributors are their own.

The world of Big Data and data science can often seem complex or even arcane from the outside looking in. In business, a lot of people by now probably understand the basics of what Big Data analysis involves – collecting the ever growing amount of data we are generating, and using it to come up with meaningful insights. But what does this actually involve on a day to day level for the professionals who get their hands dirty with the nuts and bolts?

To have a look under the hood of a job that some describe as the 'Sexiest Job Of The 21st Century' I spoke to leading data scientist Dr Steve Hanks to get an get an overview of what the work of a data scientist actually involves, and what sort of person is likely to be successful in the field.

Dr Hanks gained a PhD in computer science at Yale University, has spent 15 years as a professor of computer science and has worked at companies including Amazon, Yahoo YHOO +0%! and Microsoft MSFT +0.01%. Today he is chief data scientist at Whitepages.com where he is responsible for overseeing the Contact Graph – a database containing contact information for over 200 million people. The database is searched around two billion times every month and is the company's primary business asset.

This database has driven Whitepage's business since it was launched in 1997 and more recently it has diversified into app development. Caller ID, its replacement mobile user interface, queries the main Whitepages database to give more complete information on who is calling, and to help cut nuisance and spam calls. It also generates another revenue stream by providing its data to other companies to use in fraud prevention.

Key Capabilities of a data scientist

The term “data scientist” can cover many roles across many industries and organizations from academia to finance or Government. Hanks leads a team of 12 to 15 members responsible for all of the analytics at Whitepages, and their skillsets and duties vary. However, he tells me, there are three key capabilities which every data scientist has to understand.

1. **You have to understand that data has meaning**

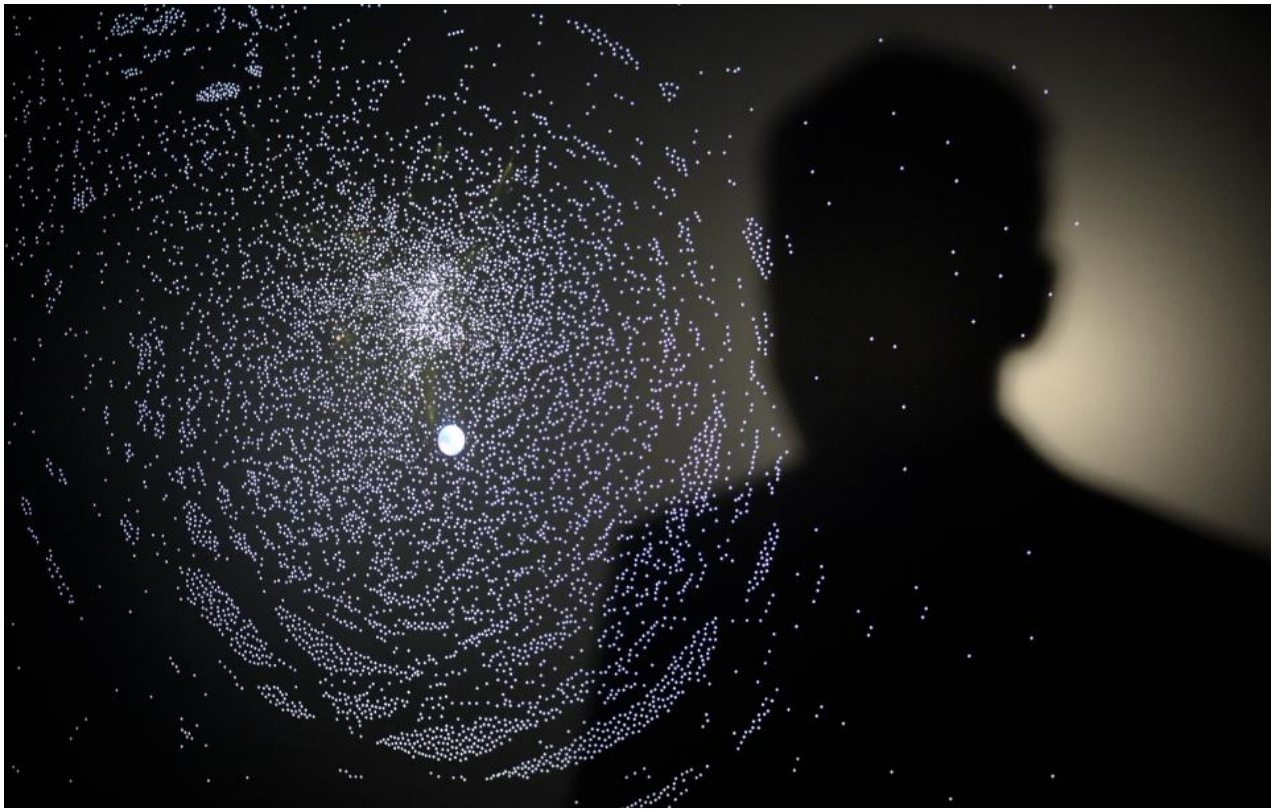
Hanks makes the point that we often overlook the fact that data means something and that it is important to understand that meaning. We have to look beyond the numbers and understand what they stand for if we are to gain any valid insights from it. Hanks points out “It doesn’t have anything to do with algorithms or engineering or anything like that. Understanding data is really an art, and it’s really important.”

2. **You have to understand the problem that you need to solve, and how the data relates to that**

Here is where you open your tool-kit to find the right analytics approaches and algorithms to work with your data. Hank talks about machine learning – which is very popular right now, but makes the point that there are hundreds of techniques to use data to solve problems – operations research, decision theory, game theory, control theory – which have all been around for a very long time. Hank says “Once you understand the data and you understand the problem you’re trying to solve, that’s when you can match the algorithm and get a meaningful solution.”

3. **You have to understand the engineering**

The third capability is about understanding and delivering the infrastructure required to perform any analysis. In Hank’s words “It doesn’t do any good to solve the problem if you don’t have the infrastructure in place to deliver the solution effectively, accurately and at the right time and place.”



Information scientist presents a screen with 12 million pixel used for data mining. AFP PHOTO / CHRISTOF STACHE (Photo credit should read CHRISTOF STACHE/AFP/Getty Images)

Being a good data scientist is really about paying attention to all three of those capabilities. You have to pay attention to the data and what it means, understand the problems and know about matching algorithms to those problems, and you have to understand the engineering to come up with solutions.

At the same time it doesn't mean there's no room for specialization. Hanks makes the point that it is virtually impossible to be an expert in all three of those areas, not to mention all the sub-divisions of each of them. It is okay to specialize in one of these areas as long as you have an appreciation of all of them. Hanks tells me: "Even if you're primarily an algorithm person or primarily an engineer. If you don't understand the problem you're solving and what your data is, you're going to make bad decisions."