Workplace Technology (A Special Report) --- How Do You Make a Data Scientist? Not Easily: Having a facility with numbers helps; but that's only a small part of the equation

Gage, Deborah . Wall Street Journal , Eastern edition; New York, N.Y. [New York, N.Y]13 Mar 2017: R.3.

ABSTRACT

The test he designed at Alpine, for example, includes a set of New York Police Department data on motor-vehicle collisions in New York City that can be subdivided in several ways -- by number and types of vehicles and drivers, number and types of injuries and deaths, contributing causes and several types of locations. Alpine Senior Data Scientist Anshuman Mishra, who is researching how a financial-services company can detect money laundering and fraud, is a former derivatives trader, while Ms.\n

FULL TEXT

The path to becoming a data scientist is not a clear one. And that's by design.

Consider the data-science team at Alpine Data, a San Francisco software startup that helps companies analyze their data to make predictions about their businesses. It includes a former marketing manager, a former physicist, a former operations researcher and a former business consultant. Helping the team as well is a former mathematician who was hired as a software engineer.

"We strongly believe that having people from different backgrounds collaborating around a problem is more important than selecting some fancy algorithms," says Alpine co-founder Steven Hillion.

In other words, despite its name, data science isn't just about being skilled with numbers. Rather, an effective data scientist also has an ability to see how particular subsets of data may be more useful than others, and what conclusions can be drawn from them.

The term data science didn't even emerge until about 2008, when it was becoming clear that the volume of data being accumulated was beyond the capacity of humans to analyze or comprehend without a machine's help. The ability to analyze billions of rows of data with hundreds of thousands of variables opened new frontiers in environmental science, medicine, politics, history and dozens of other fields.

But as the oceans of data have grown, so has the need for people who can understand statistics and machine learning, work with complex data sets and software, and explain it all to customers.

Mr. Hillion, who has a Ph.D. in mathematics, says he saw the need for data scientists at his previous company,

Greenplum, now part of EMC Corp., and had to develop techniques for creating them because there weren't enough people who could do the job. He uses the same methods at Alpine.

One test, he says, is whether a job candidate, given a choice of data sets, can pick out and work with the most interesting one. The test he designed at Alpine, for example, includes a set of New York Police Department data on motor-vehicle collisions in New York City that can be subdivided in several ways -- by number and types of vehicles and drivers, number and types of injuries and deaths, contributing causes and several types of locations. Alpine Lead Data Scientist T.J. Bay, who made the data part of the test, says it stood out to him because of the number of interesting fields that could be used to help visualize and predict accidents.

It's a particularly good test for a data scientist, Mr. Hillion says, because "it was literally, in a tech sense, multidimensional. You can break it down by geography, time, vehicle type, accident type, driver characteristics and so on. And there's no one aspect of it that is obviously a path you should go down." Also, he says, the results are something everybody's interested in -- how to avoid accidents.

Every Friday, team members explain their projects and give one another feedback. Given all the skills a data scientist needs, Mr. Hillion says, "you can't have all that in one person."

Once Alpine participated in a challenge to analyze several years of U.S. traffic accident data to better understand trends in and causes of serious accidents. Lead Data Scientist Emilie de Longueau pored over the data and isolated variables to analyze, visualize and ultimately predict the severity of injuries and accidents.

But Mr. Hillion thought her visualizations were "a little dull," he says, so he asked the other members of the team to look again at the data. While one engineer focused on Ms. de Longueau's visualizations, another asked her to explain in detail why she chose certain data and how she made her predictions.

He suggested using a new algorithm he'd devised. When the analysis was complete, two product managers -- whose focus is to translate technical into business concepts -- then used it to create an easy-to-use Web application. The app gives users a forecasting tool for estimating rates and severity of traffic accidents based on variables such as rates of drunken driving or speeding.

The finished product, Mr. Hillion says, "created a way to take the machine learning [that was first applied to the data] and make it usable by the average person."

To achieve such insights, Mr. Hillion says he hires people who can design algorithms, people who can write code to make the algorithms work on different computer systems, and people who can apply those algorithms to customers' data and then explain what they've done. That last set of skills Mr. Hillion refers to as "the human layer."

Another essential skill: knowledge of an industry. Alpine Senior Data Scientist Anshuman Mishra, who is researching how a financial-services company can detect money laundering and fraud, is a former derivatives trader, while Ms. de Longueau, who has a master's in engineering in operations research, is working on supply-chain and workforce management.

Big companies with deep pockets -- and lots of data -- often have their own data-science experts. Jeff McMillan, the chief analytics officer at Morgan Stanley, oversees about 45 people as part of a multiyear project to get more data more quickly to the firm's financial analysts and ultimately to Morgan Stanley's customers so they can make

better investing decisions.

Mr. McMillan supervises separate teams of statisticians and data-visualization experts, along with a team of data scientists who work on predicting in real time the next best actions for customers.

"We're trying to deliver advice to customers in real time," he says. "Where's the portfolio relative to the goals . . . and when was the last time you spoke to the customer? Nobody wants the portfolio that everybody else has."

He also supervises artificial-intelligence experts who are building expert systems that could "know the answer to every financial-services question," and a group of analytics managers who connect the quantitative and the business people so questions and information can be transferred faster.

"We're all going to be data scientists, just to different degrees," Mr. McMillan says. "Really what I'm focused on is connecting the science with the practice."

Helping aspiring data scientists forge their own career paths, more universities are offering programs in data science or analytics.

The University of California, Berkeley, is in its second year of a program to make data-science classes available to all undergraduates. So far, about 1,200 students from 60 majors have enrolled.

So-called connector courses are available that help them apply data-science techniques to specific areas, such as environmental engineering. An ethics class is taught as well, so students can think about "the boundaries that could be crossed if data is not used responsibly," says Cathryn Carson, an associate professor of history.

Berkeley is trying to bring students in other fields, and from underrepresented and underprivileged groups, into data science because they bring diverse perspectives, Dr. Carson says. An anthropology major, for instance, "will think deeply about the social contexts and human contexts that gave rise to the data," she says. "What kinds of questions were prompted to generate this data? Were those good questions or biased questions?"

Interest from public-health students at Berkeley drove a project to study data on child mortality in different countries. Some social-psychology students, meanwhile, want to study how humans react to mobile data collected about their health.

Dr. Carson says of such students, "They're also appreciating the social good that can be done by working with examples of human welfare, rather than just data about Twitter."

Ms. Gage is a writer in San Jose, Calif. She can be reached at reports@wsj.com.

Credit: By Deborah Gage

DETAILS

Subject:	Series &special reports; Algorithms; High technology; Artificial intelligence
Company / organization:	Name: EMC Corp; NAICS: 511210, 334112, 334118
Publication title:	Wall Street Journal, Eastern edition; New York, N.Y.
Pages:	R.3
Publication year:	2017
Publication date:	Mar 13, 2017
Publisher:	Dow Jones &Company Inc
Place of publication:	New York, N.Y.
Country of publication:	United States, New York, N.Y.
Publication subject:	Business And Economics--Banking And Finance
ISSN:	00999660
Source type:	Newspapers
Language of publication:	English
Document type:	News
ProQuest document ID:	1876330143
Document URL:	https://csuglobal.idm.oclc.org/login?url=https://search.proquest.com/docview/1876330143?accountid=38569
Copyright:	(c) 2017 Dow Jones &Company, Inc. Reproduced with permission of copyright owner. Further reproduction or distribution is prohibited without permission.
Last updated:	2017-11-24
Database:	ABI/INFORM Collection

LINKS

Linking Service