Artificial Intelligence Outperforms Human Data Scientists

By Jeremy Hsu

Posted 20 Oct 2015 | 16:00 GMT



Illustration: MIT

Artificial intelligence may be poised to ease the shortage of data scientists who build models that explain and predict patterns in the ocean of “Big Data” representing today’s world. **An MIT startup’s computer software has proved capable of building better predictive models than the majority of human researchers** it competed against in several recent data science contests.

Until now, well-paid [data scientists](http://spectrum.ieee.org/podcast/at-work/tech-careers/is-data-science-your-next-career) have relied on their human intuition to create and test computer models that can explain and predict patterns in data. But MIT’s **“Data Science Machine”** software represents a fully automated process capable of building such **predictive computer models** by identifying relevant features in raw data. Such a tool could make human data scientists even more effective by allowing them to build and test such predictive models in far less time. But it might also help more individuals and companies harness the power of Big Data without the **aid of trained data scientists**.

“I think the biggest potential is for increasing the pool of people who are capable of doing data science,” Max Kanter, a data scientist at MIT’s Computer Science and AI Lab and co-creator of the Data Science Machine software, told *IEEE Spectrum*. “If you look at the growth **in demand for people with data science abilities**, it’s far outpacing the number of people who have those skills.”

The Data Science Machine can automatically create accurate predictive models based on raw datasets within two to 12 hours; a team of human data scientists may require months. A [paper on the Data Science Machine](http://www.jmaxkanter.com/static/papers/DSAA_DSM_2015.pdf) will be presented this week at the [IEEE International Conference on Data Science and Advanced Analytics](http://dsaa2015.lip6.fr/) being held in Paris from 19–21 Oct.

Trained data scientists, who typically draw [salaries above $100,000 on average](http://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0%2C14.htm), remain a coveted but scarce resource for companies as diverse as Facebook and Walmart. In 2011, the [McKinsey Global Institute](http://www.mckinsey.com/features/big_data) estimated that the United States alone might face a shortage of 140,000 to 190,000 people with the analytical skills necessary for data science. A 2012 issue of the *Harvard Business Review* declared data scientist as the [sexiest job of the 21st century](http://spectrum.ieee.org/tech-talk/computing/it/is-data-scientist-the-sexiest-job-of-our-time).

The reason for such high demand for data scientists comes from Big Data’s revolutionary promise of tapping into vast collections of data—whether it’s the online behaviour of social media users, the movements of financial markets worth trillions of dollars, or the billions of celestial objects spotted by telescopes—to explain and predict patterns in the huge datasets. Such models could help companies predict the future behaviour of individual customers or aid astronomers in automatically identifying an object in the starry night-time sky.

But how do you transform a sea of raw data into information that can help businesses or researchers identify and predict patterns? Human data scientists usually have to spend weeks or months working on their predictive computer algorithms. First, they sift through the raw data to identify key variables that could help predict the behaviour of related observations over time. Then they must continuously test and refine those variables in a series of computer models that often use [**machine learning techniques**](http://spectrum.ieee.org/biomedical/diagnostics/the-algorithm-thats-hunting-ebola).

Such a time-consuming part of the data scientists’ job description inspired Kanter, an MIT grad student at the time, and Kalyan Veeramachaneni, a research scientist at MIT’s Computer Science and AI Lab who acted as Kanter’s master’s thesis advisor, to try creating a computer program that could automate the biggest bottlenecks in data science.

Previous computer software programs aimed at solving such data science problems have tended to be one dimensional, focusing on problems particular to specific industries or fields. But Kanter and Veeramachaneni wanted their Data Science Machine software to be capable of tackling any general data science problem. Veeramachaneni in particular drew on his experience of seeing similar connections among the many industry data science problems he had worked on during his time at MIT.

That experience helped inspire the first and greatest part of Data Science Machine’s triumph; the automation of the “feature engineering” process of identifying and extracting relevant variables from the sea of raw data. A second part of the new MIT software focuses on auto-tuning: figuring out the best set of parameters to generate the best predictions from the data. In this case, the software selects a subset of the most relevant variables and choses the best machine learning technique for determining the relationship between the variables and the model predictions.

Once the Data Science Machine was ready, Kanter and Veeramachaneni tested the software on datasets from three separate data science competitions: KDD Cup 2014, IJCAI, and KDD Cup 2015.

The Data Science Machine’s results beat 615 of 906 human teams competing in those competitions. It also achieved predictions that were 94 percent, 96 percent and 87 percent as accurate as the winning models submitted in each competition. That means the **artificial intelligence** behind Data Science Machine may not yet beat out the top tier of human data scientists, but it can fairly easily match the efforts of many data scientists with far less effort.

”Typically the Data Science Machine does well and beats many humans, but some humans beat it,” Kanter explained. “So it would be naive to say human data scientists don’t have any value.”

Kanter and Veeramachaneni see the Data Science Machine as an automated tool that can empower data scientists and **make them more efficient** rather than putting them out of a job. They have already begun tweaking the software to allow for **more human control** rather than simply leaving humans out of the loop.

For example, a data scientist could run Data Science Machine and use its results as a **baseline to build a better predictive model**. Or a data scientist could focus more on the feature engineering side but leave the machine learning optimization to the software.

But the potentially more disruptive side of Data Science Machine may come from empowering all the companies that don’t already have trained data scientists on payroll. Plenty of companies, both large and small, could benefit from Big Data, but don’t have the trained teams of data scientists found at prominent tech companies such as Google or Amazon. Data Science Machine’s efficient results might be good enough for an astronomer working in a university lab, or the marketing arm of traditional retail business that has yet to really build a data science team. As Kanter pointed out:

In the future world we’re headed for, where every company makes data-driven decisions, you can’t just make existing data scientists more efficient; you also have to increase the pool of people who can be data scientists. I don’t think that’s done by training everybody to be data scientists. It’s done by building new tools that let machines do the things they do best and letting humans do what they do best.

In a sense, Data Science Machine could turn **almost any company into a “tech company” by enabling it to make key business decisions and build new products based on Big Data**, Kanter says. The software could spread the impact of Big Data across sectors as diverse as e-commerce, crowdfunding, retail, education, financial services, and government. To that end, Kanter and Veeramachaneni have already begun courting clients through a [startup called FeatureLab](http://www.featurelab.co/industry/). The startup’s website greets visitors with the message: “Do more with your data, without more data scientists.”