The New York Times | http://nyti.ms/1mZywng

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Technology revolutions come in measured, sometimes foot-dragging steps. The lab science and marketing enthusiasm tend to underestimate the bottlenecks to progress that must be overcome with hard work and practical engineering.

The field known as "big data" offers a contemporary case study. The catchphrase stands for the modern abundance of digital data from many sources — the web, sensors, smartphones and corporate databases — that can be mined with clever software for discoveries and insights. Its promise is smarter, data-driven decision-making in every field. That is why data scientist is the economy's hot new job.

Yet far too much handcrafted work — what data scientists call "data wrangling," "data munging" and "data janitor work" — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

"Data wrangling is a huge — and surprisingly so — part of the job," said Monica Rogati, vice president for data science at Jawbone, whose sensor-filled wristband and software track activity, sleep and food consumption, and suggest dietary and health tips based on the numbers. "It's something that is not appreciated by data civilians. At times, it feels

like everything we do."

Several start-ups are trying to break through these big data bottlenecks by developing software to automate the gathering, cleaning and organizing of disparate data, which is plentiful but messy. The modern Wild West of data needs to be tamed somewhat so it can be recognized and exploited by a computer program.

"It's an absolute myth that you can send an algorithm over raw data and have insights pop up," said Jeffrey Heer, a professor of computer science at the University of Washington and a co-founder of Trifacta, a start-up based in San Francisco.

Timothy Weaver, the chief information officer of Del Monte Foods, calls the predicament of data wrangling big data's "iceberg" issue, meaning attention is focused on the result that is seen rather than all the unseen toil beneath. But it is a problem born of opportunity. Increasingly, there are many more sources of data to tap that can deliver clues about a company's business, Mr. Weaver said.

In the food industry, he explained, the data available today could include production volumes, location data on shipments, weather reports, retailers' daily sales and social network comments, parsed for signals of shifts in sentiment and demand.

The result, Mr. Weaver said, is being able to see each stage of a business in greater detail than in the past, to tailor product plans and trim inventory. "The more visibility you have, the more intelligent decisions you can make," he said.

But if the value comes from combining different data sets, so does the headache. Data from sensors, documents, the web and conventional databases all come in different formats. Before a software algorithm can go looking for answers, the data must be cleaned up and converted into a unified form that the algorithm can understand.

Data formats are one challenge, but so is the ambiguity of human language. Iodine, a new health start-up, gives consumers information on drug side effects and interactions. Its lists, graphics and text descriptions

are the result of combining the data from clinical research, government reports and online surveys of people's experience with specific drugs.

But the Food and Drug Administration, National Institutes of Health and pharmaceutical companies often apply slightly different terms to describe the same side effect. For example, "drowsiness," "somnolence" and "sleepiness" are all used. A human would know they mean the same thing, but a software algorithm has to be programmed to make that interpretation. That kind of painstaking work must be repeated, time and again, on data projects.

Data experts try to automate as many steps in the process as possible. "But practically, because of the diversity of data, you spend a lot of your time being a data janitor, before you can get to the cool, sexy things that got you into the field in the first place," said Matt Mohebbi, a data scientist and co-founder of Iodine.

The big data challenge today fits a familiar pattern in computing. A new technology emerges and initially it is mastered by an elite few. But with time, ingenuity and investment, the tools get better, the economics improve, business practices adapt and the technology eventually gets diffused and democratized into the mainstream.

In software, for example, the early programmers were a priesthood who understood the inner workings of the machine. But the door to programming was steadily opened to more people over the years with higher-level languages from Fortran to Java, and even simpler tools like spreadsheets.

Spreadsheets made financial math and simple modeling accessible to millions of nonexperts in business. John Akred, chief technology officer at Silicon Valley Data Science, a consulting firm, sees something similar in the modern data world, as the software tools improve.

"We are witnessing the beginning of that revolution now, of making these data problems addressable by a far larger audience," Mr. Akred said.

ClearStory Data, a start-up in Palo Alto, Calif., makes software that recognizes many data sources, pulls them together and presents the results

visually as charts, graphics or data-filled maps. Its goal is to reach a wider market of business users beyond data masters.

Six to eight data sources typically go into each visual presentation. One for a retailer might include scanned point-of-sale data, weather reports, web traffic, competitors' pricing data, the number of visits to the merchant's smartphone app and video tracking of parking lot traffic, said Sharmila Shahani-Mulligan, chief executive of ClearStory.

"You can't do this manually," Ms. Shahani-Mulligan said. "You're never going to find enough data scientists and analysts."

Trifacta makes a tool for data professionals. Its software employs machine-learning technology to find, present and suggest types of data that might be useful for a data scientist to see and explore, depending on the task at hand.

"We want to lift the burden from the user, reduce the time spent on data preparation and learn from the user," said Joseph M. Hellerstein, chief strategy officer of Trifacta, who is also a computer science professor at the University of California, Berkeley.

Paxata, a start-up in Redwood City, Calif., is focused squarely on automating data preparation — finding, cleaning and blending data so that it is ready to be analyzed. The data refined by Paxata can then be fed into a variety of analysis or visualization software tools, chosen by the data scientist or business analyst, said Prakash Nanduri, chief executive of Paxata.

"We're trying to liberate people from data-wrangling," Mr. Nanduri said. "We want to free up their time and save them from going blind."

Data scientists emphasize that there will always be some hands-on work in data preparation, and there should be. Data science, they say, is a step-by-step process of experimentation.

"You prepared your data for a certain purpose, but then you learn something new, and the purpose changes," said Cathy O'Neil, a data scientist at the Columbia University Graduate School of Journalism, and co-author, with Rachel Schutt, of "Doing Data Science" (O'Reilly Media,

2013).

Plenty of progress is still to be made in easing the analysis of data. "We really need better tools so we can spend less time on data wrangling and get to the sexy stuff," said Michael Cavaretta, a data scientist at Ford Motor, which has used big data analysis to trim inventory levels and guide changes in car design.

Mr. Cavaretta is familiar with the work of ClearStory, Trifacta, Paxata and other start-ups in the field. "I'd encourage these start-ups to keep at it," he said. "It's a good problem, and a big one."

Correction: August 19, 2014

An article on Monday about the development of software to automate the gathering, cleaning and organizing of disparate data misstated the year when "Doing Data Science" was published. It first came out in 2013, not 2014.

A version of this article appears in print on August 18, 2014, on page B4 of the New York edition with the headline: For Data Scientists, 'Janitor Work' Is Hurdle to Insights.
