The New York Times

April 11, 2013

Data Science: The Numbers of Our Lives

By CLAIRE CAIN MILLER

HARVARD BUSINESS REVIEW calls data science “the sexiest job in the 21st century,” and by most accounts this hot new field promises to revolutionize industries from business to government, health care to academia.

The field has been spawned by the enormous amounts of data that modern technologies create — be it the online behavior of Facebook users, tissue samples of cancer patients, purchasing habits of grocery shoppers or crime statistics of cities. Data scientists are the magicians of the Big Data era. They crunch the data, use mathematical models to analyze it and create narratives or visualizations to explain it, then suggest how to use the information to make decisions.

In the last few years, dozens of programs under a variety of names have sprung up in response to the excitement about Big Data, not to mention the six-figure salaries for some recent graduates.

In the fall, Columbia will offer new master’s and certificate programs heavy on data. The University of San Francisco will soon graduate its charter class of students with a master’s in analytics. Other institutions teaching data science include New York University, Stanford, Northwestern, George Mason, Syracuse, University of California at Irvine and Indiana University.

Rachel Schutt, a senior research scientist at Johnson Research Labs, taught “Introduction to Data Science” last semester at Columbia (its first course with “data science” in the title). She described the data scientist this way: “a hybrid computer scientist software engineer statistician.” And added: “The best tend to be really curious people, thinkers who ask good questions and are O.K. dealing with unstructured situations and trying to find structure in them.”

Eurry Kim, a 30-year-old “wannabe data scientist,” is studying at Columbia for a master’s in quantitative methods in the social sciences and plans to use her degree for government service. She discovered the possibilities while working as a corporate tax analyst at the Internal Revenue Service. She might, for example, analyze tax return data to develop algorithms that flag fraudulent filings, or cull national security databases to spot suspicious activity.

Some of her classmates are hoping to apply their skills to e-commerce, where data about users' browsing history is gold.

"This is a generation of kids that grew up with data science around them — Netflix telling them what movies they should watch, Amazon telling them what books they should read — so this is an academic interest with real-world applications," said Chris Wiggins, a professor of applied mathematics at Columbia who is involved in its new Institute for Data Sciences and Engineering. "And," he added, "they know it will make them employable."

Universities can hardly turn out data scientists fast enough. To meet demand from employers, the United States will need to increase the number of graduates with skills handling large amounts of data by as much as 60 percent, according to a report by McKinsey Global Institute. There will be almost half a million jobs in five years, and a shortage of up to 190,000 qualified data scientists, plus a need for 1.5 million executives and support staff who have an understanding of data.

North Carolina State University introduced a master's in analytics in 2007. All 84 of last year's graduates in the field had job offers, according to Michael Rappa, who conceived and directs the university's Institute for Advanced Analytics. The average salary was $89,100, and more than $100,000 for those with prior work experience.

"This has become relevant to every company," said Michael Chui, a principal at McKinsey who has studied the field. "There's a war for this type of talent."

Because data science is so new, universities are scrambling to define it and develop curriculums. As an academic field, it cuts across disciplines, with courses in statistics, analytics, computer science and math, coupled with the specialty a student wants to analyze, from patterns in marine life to historical texts.

With the sheer volume, variety and speed of data today, as well as developing technologies, programs are more than a repackaging of existing courses. "Data science is emerging as an academic discipline, defined not by a mere amalgamation of interdisciplinary fields but as a body of knowledge, a set of professional practices, a professional organization and a set of ethical responsibilities," said Christopher Starr, chairman of the computer science department at the College of Charleston, one of a few institutions offering data science at the undergraduate level.

Most master's degree programs in data science require basic programming skills. They start with what Ms. Schutt describes as the "boring" part — scraping and cleaning raw data and "getting it into a nice table where you can actually analyze it." Many use data sets provided by businesses or government, and pass back their results. Some host competitions to see which student can come up with the best solution to a company's problem.

University of San Francisco students have used data from General Electric to predict how much energy windmills could create. At North Carolina State, with data from the Postal Service, students have analyzed response rates to junk mail to find ways to improve its effectiveness.

Studying a Web user's data has privacy implications. Using data to decide someone's eligibility for a line of credit or health insurance, or even recommending who they friend on Facebook, can affect their lives. "We're building these models that have impact on human life," Ms. Schutt said. "How can we do that carefully?" Ethics classes address these questions.

Finally, students have to learn to communicate their findings, visually and orally, and they need business know-how, perhaps to develop new products.

"That's one of the challenges," said Terence Parr, program director of the analytics and computer science programs at the University of San Francisco. "To be successful, you need to have a wide range of skills that doesn't fit in one department."

The question, said Bill Howe, who teaches data science at the University of Washington, is whether it is even possible to instill in a single person all the skills needed, from statistics to predictive modeling to business strategy. The university's offerings range from a free online course on Coursera to a nine-month certificate program to a Ph.D. track in Big Data.

"It remains to be seen," he said, "but we're still of the mind that a curriculum that aims to train data scientists is feasible." He added: "What employers want is someone who can do it all."

Claire Cain Miller is a technology reporter for The Times.