Just pondering the current job market for biomedical data scientists is likely to put a smile on the faces of many in the field.

“The bottom line is, compared to other disciplines, bioinformatics and computational biology are the hottest areas these days,” says Veerasamy “Ravi” Ravichandran, PhD, a program director at the National Institute of General Medical Sciences (NIGMS), which is one of the National Institutes of Health (NIH).

That heat is being supplied by many sources. On the one hand, colleges and universities across the country are either expanding existing departments dedicated to fields like biomedical informatics and quantitative biology, or building them from scratch. On the other, there is a vast and growing demand in industry for people who can wrangle biomedical big data, whether at established companies or the latest Silicon Valley startups.

The search for data-savvy researchers with backgrounds in computational biology, bioinformatics, and biostatistics is unlikely to cool down anytime soon. Cheaper and more powerful computing resources, new database systems and software tools, and novel statistical methods and machine-learning techniques hold great promise for everything from basic research to clinical applications and public health. They are also a potential treasure trove: According to a 2013 report by the McKinsey Global Institute, for example, big data analytics could generate health-care cost savings of up to $190 billion annually by 2020.

Alas, the same report also predicts that by 2018, the United States could face an overall shortage of 190,000 data scientists. Industry is responding through broad-based initiatives like the Insight Data Science Fellows Program, which pairs PhDs in
various fields with data-science mentors at a wide range of companies. The NIH, meanwhile, is developing its own pipeline for biomedical data scientists.

Ravichandran, for example, formerly managed the institutional training grants in bioinformatics and computational biology administered through the NIGMS, which funds graduate students at 13 different centers, institutes, and academic departments in nine different states.

His colleague, Valerie Florance, PhD, Director of Extramural Programs for the National Library of Medicine (NLM), a part of the NIH, coordinates the NLM’s training programs in bioinformatics, which currently support approximately 200 doctoral students and postdocs at 14 universities. She also serves on the training committee for the NIH’s Big Data to Knowledge (BD2K) initiative, which seeks to prepare the workforce needed to handle large, complex biomedical data sets. BD2K recently introduced a new training grant in what it calls “Biomedical Big Data Science” that explicitly requires trainees to work at the intersection of computer science, statistics, and biomedical science—a combination that speaks to the inherently interdisciplinary nature of biomedical data science. And it is awarding grants for the development of open educational resources, such as online courses, that will provide training in biomedical data science to graduate students and established researchers alike (See story on page 7).

Where the recipients of all this training actually wind up is the million-dollar question, says Ravichandran, since those career outcomes will have a direct bearing on how the NIH and its institutional partners can further expand and refine the supply of biomedical data scientists. But right now, it’s a question with only the vaguest of answers.

According to a 2012 report by the NIH’s own Biomedical Research Workforce Working Group, approximately 26 percent of all biomedical PhDs move into tenured or tenure-track faculty positions, while 30 percent head toward the biotech and pharmaceutical industries. But that same working group reported that it was “frustrated and sometimes stymied” by a lack of data.

Details regarding trends within the biomedical data-science community are just as fuzzy, since funding agencies have not historically tracked the career trajectories of trainees. That’s beginning to change, in part because the same working group recommended that training institutions collect information on graduate students and postdocs, and provide it to both the NIH and to prospective students. Florance, for example, has been using a software tool called CareerTrac to keep tabs on NLM training-grant recipients, about half of whom go into academia or find work with healthcare organizations. While the data in the system remains incomplete, it’s striking nonetheless: In recent years, trainees have landed jobs at companies as diverse as Pfizer and Google, and racked up titles ranging from assistant professor to chief medical officer and CEO.

And as even the handful of profiles included here illustrate, certain patterns do emerge.

Biomedical data scientists, who tend to enter graduate school with eclectic and interdisciplinary backgrounds in both biological and quantitative science, tend to head off in equally eclectic and interdisciplinary directions once they leave. Sometimes, they move straight into academia or industry; often, however, there’s a certain amount of bouncing back and forth between the two. Versatility, it would seem, is fundamental to what these people do and who they are.

That characteristic is only heightened by graduate training that is necessarily multidisciplinary: Many researchers are initially better versed in either biological or quantitative science and must fill in the blanks through a combination of coursework, individual mentoring by advisors and collaborators, and on-the-job training. “A lot of people who work in this field learn what they need on the go,” says Daniela Witten, PhD, a biostatistician at the University of Washington.

They also often serve as intermediaries between colleagues—computer scientists and cell biologists, statisticians and drug researchers—who do not speak one another’s respective languages.

And their flexibility is further reinforced by a common focus on the development of what Steven Bagley, MD, MS, Executive Director of Stanford University’s Biomedical Informatics (BMI) program, calls “generalizable methods”—i.e., ones that can be applied across many different domains.

As Florance says, “We want trainees developing methods and approaches that apply across fields. We don’t want them to learn to do just one thing, and pound that hammer forever.”

As career dilemmas go, an overabundance of options seems fairly benign. And given how strong demand is likely to be for the foreseeable future, it’s one that many biomedical data scientists are destined to confront. “It’s a seller’s market,” says Bagley. “Too many things to do, and not enough people to do them.”

Sometimes, [biomedical data scientists] move straight into academia or industry; often, however, there’s a certain amount of bouncing back and forth between the two. Versatility, it would seem, is fundamental to what these people do and who they are.
“Knowing a lot of statistics is good,” says Daniela Witten. “But knowing a little statistics is dangerous.”

By that measure, Witten herself ought to be just fine. Armed with an undergraduate degree in mathematics and biological sciences from Stanford, Witten stayed on to pursue graduate studies in statistics with the intention of focusing on computational biology. But she migrated instead toward statistical machine learning—in part, she says, because she wanted to develop a broad set of mathematical tools that would be applicable not just to the type of data that we see today, but to the type of data that we’ll be seeing for the next 30 years.

As a doctoral candidate, Witten cut her teeth on just that kind of data by developing statistical methods with senior faculty in Stanford’s School of Medicine—including Andrew Fire, PhD, the George D. Smith Professor in Molecular and Genetic Medicine, who together with Craig Mello won the 2006 Nobel Prize for the discovery of microRNA (miRNA). Like Ron Yu (Interview on page 24), Witten was attracted by the challenge of developing new statistical techniques to deal with high-dimensional data, in which the number of variables far outstrips the number of samples; and she did precisely that while helping Fire analyze high-throughput miRNA data derived from cervical cancer samples.

Witten, who has since been named to Forbes’ 30 Under 30 list three times, says that she received most of her training in two of the other central pillars of biomedical data science—namely, biology and computer science—through such collaborative projects. And now that she’s a principal investigator herself, she tries to make sure that her graduate students get the same kinds of opportunities. She’s also developing an interdisciplinary Masters program in data science at the University of Washington that will draw upon six different departments, including biostatistics and computer science. Witten estimates that perhaps a third of her own grad-school classmates went into academia, while the remaining two thirds took jobs with tech companies or in finance—a testament to just how widely applicable their skills truly are.

Witten is quick to point out that data science itself is less a single field than a broad discipline with many sub-disciplines. As a result, there’s no single path towards preparing for it; rather, it all boils down to people getting the kind of training they will need to do the kind of work they want to do. And that process never really ends. “The more you learn,” she says, “the more you realize you still have much left to learn.”
Nicholas Tatonetti, PhD
Assistant Professor of Biomedical Informatics, Columbia University

For the past several years, Nicholas Tatonetti has been busy building his lab at Columbia University Medical Center: recruiting students, mentoring postdocs and doctoral candidates, and pursuing research projects that range across bioinformatics and computational biology.

But things might have turned out very differently. Tatonetti took a couple of years off between high school and college, selling insurance and earning his real estate license. Then academia beckoned in the form of a night class in physics at a community college just outside Phoenix, Arizona. “I decided right then that I wanted to have a career as a professor,” he recalls. Subsequent exposure to genetics and computational modeling at Arizona State University, where he earned a pair of degrees in computational mathematics and molecular biosciences, sealed the deal. “From that point on,” Tatonetti says, “I was hooked.”

Not much has changed since then. As a doctoral candidate in Stanford’s BMI program, Tatonetti developed novel statistical and computational methods that allowed him to mine the Food and Drug Administration’s voluminous records on adverse drug reactions, identifying pairs of medications that caused problems when taken together. He and his students continue to work on new ways of deriving clinical insights from masses of observational data; earlier this year, they published a study in the *Journal of the American Medical Informatics Association* that trawled through 1.75 million electronic health records (EHRs) in order to demonstrate that a person’s birth month can affect his or her lifetime disease risk. In addition, they are combining information culled from EHRs with next-generation sequencing data and network biology models to both identify clinical effects like adverse drug events, and to understand the basic biology behind them. To top it all off, Tatonetti also directs the Clinical Informatics Shared Resource at Columbia’s Herbert Irving Comprehensive Cancer Center, where he develops practical bioinformatics tools to help support the work of cancer researchers.

According to Tatonetti, most of his Stanford classmates took jobs with Silicon Valley startups after graduation. “It’s a good time for health startups right now; venture capital is ready and willing,” he says. But while he’s had his own fair share of industry experience—he put himself through college as a software consultant, worked for a couple of consulting firms and startups in grad school, and continues to collaborate with a few companies here and there—that’s on the back burner for now, if only because he has so much on his plate already. “We have so much data and only so many people,” Tatonetti says of the current situation at Columbia. “There are many more exciting projects and data sources available than the lab can handle.”
Ron Yu, PhD
Senior Statistical Scientist, Genentech

Ron Yu had always been interested in applied math—enough to have double-majored in electrical engineering and mathematics at Worcester Polytechnic Institute before enrolling in Stanford University’s Scientific Computing and Computational Mathematics program (now the Institute for Computational and Mathematical Engineering, or ICME). “Then,” he says, “I got interested in biology.”

Specifically, Yu got interested in the statistical challenge posed by microarrays, a high-throughput sequencing technology that can generate expression data for thousands of genes from a single experiment. That throws a wrench into the methods of classical statistics, which break down when the number of measured variables (i.e., genes) is greater than the number of observations over which those variables are measured (i.e., samples).

So Yu took a couple of courses in bioinformatics and computational biology, read some textbooks in biology and genetics—and landed a position as a research assistant in the lab of Branimir Sikic, MD, a professor in the School of Medicine who was using cDNA microarrays to study cancer. With support from his advisor, Robert Tibshirani, PhD, a professor of statistics and health policy, Yu developed novel statistical methods to help the biologists and clinicians in Sikic’s lab analyze their data, even as they helped him understand the basic biology underlying their research.

A summer internship at Genentech gave Yu his first taste of industry. After graduation, however, he opted for a postdoctoral position at the University of California, San Diego, where he used computational methods to identify potential binding sites for transcription factors in the yeast genome. In time, though, Yu found that he missed the clear goals and benefits involved in helping to develop new therapies for large numbers of people. So he returned to Genentech, where he has worked ever since.

Currently, Yu is the study statistician for two Phase III clinical trials that seek to compare a drug called Kadcyla with the standard of care for both early and metastatic breast cancer. His formal duties include designing the randomization schemes for the trials, writing their statistical analysis plans, and analyzing the data they produce. But he also often finds himself playing the role of scientific interpreter, explaining the quantitative results of the studies to his fellow team members, who include not only clinical pharmacologists and medical doctors but also statistical programmers and project managers.

“I enjoy the work I do,” Yu says. “Because if the drug works, it will benefit thousands of patients.”
Amrita Basu, PhD
Genomics and Computational Biology Lead, Lockheed Martin

By the time Amrita Basu found herself working as a postdoctoral associate at the Broad Institute, she’d already had plenty of experience in both industry and academia.

Equipped with a dual degree in electrical engineering and computer science from Cornell University, Basu landed a job as a software developer at Oracle Corporation straight out of college. But she wanted to do work that would have more of an impact on the well-being of others; and inspired in part by a physician friend who was studying bioinformatics at Columbia University (and by the Human Genome Project, which was just coming to an end), she found it at the intersection of health and technology.

As a doctoral candidate in computational biology at Rockefeller University—part of a tri-institutional PhD program formerly run by Rockefeller, Cornell, and Memorial Sloan-Kettering Cancer Center—Basu worked under molecular biologist C. David Allis, PhD, head of the Laboratory of Chromatin Biology and Epigenetics, where she helped develop a novel software tool to predict histone and non-histone modifications in proteins. She continued to work on predictive modeling at the Broad Institute, where she led the computational component of a project designed to identify potential targets for cancer therapy.

Nonetheless, Basu still wasn’t sure what to do next. Fortunately, her co-mentors, Stuart L. Schreiber, PhD, director of the Institute’s Center for the Science of Therapeutics, and Paul A. Clemons, PhD, director of the Institute’s computational chemical biology research, offered some sound advice for anyone considering a career change: “Be open.”

And so she was.

After finishing up her postdoc, Basu moved to San Francisco and accepted a position as principal investigator in a new genomics department located in the Health and Life Sciences division of Lockheed Martin. She likens it to working for a small startup inside a big company; and so far, the transition back to industry has been a smooth one.

Basu currently leads an initiative to build a computational platform that can store, process, and analyze the millions of genomes that are collected for population-health studies in the United States and abroad. The scale of such projects means that Basu gets to work with a wide range of collaborators in government, academia, and healthcare. Best of all, she has the opportunity to empower millions of patients. “They’ll have access to their own data,” she says. “And their physicians will have it, too.”
If you want proof of how quickly biomedical data science is evolving—and how permeable the barrier between academia and industry really is—look no further than Grace Zheng.

When Zheng enrolled at the University of British Columbia in 2000, there was no formal program in either computational biology or bioinformatics. (Today there are programs in both.) So she took the handful of graduate-level courses that were available, picked up a degree in computer science and biology—and headed off to MIT, where she and three other students formed the first cohort in the brand-new Computational and Systems Biology PhD Program. "I came in well-prepared from the computational side, but that was my first time working in a wet lab," recalls Zheng, who suddenly found herself not only modeling the evolution and function of microRNAs in cancer and embryonic stem cells, but also dissecting mice to physically extract her samples.

After interning at Vertex Pharmaceuticals, Zheng took a postdoctoral position in Stanford's School of Medicine where she used computational methods and next-generation sequencing to discover how a known oncogenic transcription factor called cMyc differentially regulates the transcription of thousands of long non-coding RNAs. She also enrolled in Stanford Ignite, a certificate program in the Graduate School of Business that teaches management skills and entrepreneurship to graduate students and technical professionals—and connects them to entrepreneurs, executives, and venture capitalists. Shortly thereafter, Zheng got connected to some of the people behind 10X Genomics, a startup devoted to enhancing next-generation sequencing platforms by barcoding the fragments of genetic material that such platforms must read and reassemble. Zheng began consulting with the company while its technology was still under development, and went full-time as soon as she finished her postdoc, laboring around the clock to help get its first product out earlier this year.

So far, the situation has been ideal: Zheng gets to work closely with a diverse crew, from the biochemists and software engineers at 10X to the biomedical researchers who are the company’s customers; she’s able to write papers and present her work at conferences; and she develops cutting-edge technology that could ultimately revolutionize next-generation sequencing.

As a result, Zheng says, she’s been able to develop valuable business skills while at the same time remaining “very connected” to the world of academic research—a recipe for keeping one’s options open, as it were. “If the next job opportunity comes up in academia, who knows where I might wind up?” she says.
Bioinformatics to Make a Difference: A Personal Calling

Luke Yancy, Jr.
Data Science Consultant, NunaHealth
PhD candidate in Biomedical Informatics, Stanford University

To Luke Yancy, Jr., biomedical informatics is personal.

As an undergraduate at Morehouse College, Yancy was initially drawn to bioinformatics as a means of combining his talent for computer science with his passion for helping others. Then, as he was applying to graduate programs, his 43-year-old mother died of a massive heart attack. What’s more, within the span of a single month, three of his friends lost their own mothers—all African-American, all under the age of 50—under similar circumstances.

That string of tragedies shaped the question that has driven Yancy’s research ever since: Why do certain diseases disproportionately affect certain groups—including minorities?

Yancy pursued that question in the lab of Atul Butte, MD, PhD, whom he first met as an undergraduate through the Stanford Summer Research Program. (Butte left Stanford in April to become director of the new Institute of Computational Health Sciences at the University of California, San Francisco.) While there, his interests broadened to include serious rare diseases that are often neglected by researchers due to a lack of data—in particular, pulmonary hypertension, a rare disorder studied by Stanford clinician Vinicio de Jesus Perez, MD. Yancy’s dissertation, which he recently defended, combines patient data provided by Perez with large amounts of publicly available data to demonstrate how next-generation sequencing can be used to better understand such rare illnesses by linking them to more common ones. More generally, it also illustrates how the computational methods typically deployed against big data can be profitably used to attack small data, as well.

Yancy landed an internship at the San Francisco Bay Area startup NunaHealth, which was cofounded by BMI alumnus David Chen, PhD. That, in turn, led to a part-time position that will become full-time as soon as he graduates. Eventually, Yancy hopes to teach bioinformatics back at Morehouse. But for now, he looks forward to racking up some industry experience—again, for reasons that are as much personal as professional.

NunaHealth provides data analytics to help companies shape their own health-insurance offerings. In time, Yancy says, such data analytics should allow NunaHealth to compare the advantages of different healthcare payment schemes, and to develop better ones—something that Yancy, who was himself confronted by several thousand dollars’ worth of healthcare fees as a graduate student, is eager to do. “Eventually, we’re going to be able to suggest alternative models that will help support fair pricing for everyone,” he says.