BY CHRIS PALMER

A Boost for Biomedical Data Science Training: BD2K Grants Push the Field

Training tomorrow's scientists to be as comfortable developing algorithms as they are developing assays is a vital part of the National Institutes of Health (NIH) Big Data to Knowledge (BD2K) program, which was launched in 2013 to develop new data science concepts and create specific analytic tools to extract the maximum value from biomedical big data.

In May 2015, BD2K awarded institutional training grants to three universities. As one might expect, each grant provides support for training a specific number of graduate students. But the programs do more than boost the number of people in the field. They are each trying to find the sweet spot at the interface between biology and computation where students gain appropriate skills in a rapidly growing field without being overwhelmed.

Does the future lie in training people to work in teams? Will students best learn by dividing their time between computer science and biomedical labs? What courses are truly essential for giving students the confidence and skills to better understand and manipulate big data? Are industry internships valuable?

"Across the community, there is very little agreement about the core competencies of data science, much less biomedical data science," says **Michelle Dunn, PhD**, senior advisor for Data Science Training, Diversity, and Outreach at the NIH. This produces a fair amount of confusion about how new training programs should differ from existing ones. "The central theme that ties the [grantees'] programs together," Dunn says, "is developing methods for big data—that is, teaching the fundamental skills needed to develop methods and tools to analyze large or complex data in a statistically sound way at scale."

The three current NIH grantees (there will be more) are each taking a different approach to providing this training. And universities across the country are watching them in hopes of discovering which strategies are most effective.

Training for Teamwork

Under the BD2K grant to the University of North Carolina–Chapel Hill (UNC–CH), the focus is on teaching interdisciplinary teams of students to work together.

"Historically, data analysis could be done by one person who knew enough about the various discoveries, but now I think the challenges are so much more difficult," says biostatistician **Michael Kosorok, PhD**, distinguished professor of biostatistics at UNC–CH and co-director of the UNC–CH BD2K grant. "Now, really, it's a team science endeavor."

So, Kosorok says, the overall vision of the UNC's BD2K Training Program, an effort involving nearly 50 faculty members from 11 departments, is for trainees to learn "how to work in multidisciplinary teams and develop the strengths to solve some of the difficult, open-ended research problems related to big data."

At the start, the UNC program will fund six students from diverse academic departments for three years. But Kosorok hopes the program will attract 14 additional students who will receive funding from their home departments and will join the program at different stages of their doctoral training. It's an "If you build it, they will come" model of program design.

The trainees—recruited from domains such as molecular biology and genetics, as well as computer science, statistics, biostatistics, informatics and mathematics—will come together to share their diverse backgrounds in three five-week-long modules (a total of three credit hours). Each module is designed around a big-data question, to be explored in each of four domains: biostatistics, math, computer science, and biomedical science. Trainees will follow up each module with a semester-long team-centered lab course focused on a single project, the goal of which is the submission of a research paper or conference proceeding. For example, the team might integrate RNA and DNA sequencing data to identify genetic markers of tumor aggressiveness and model the transport of materials within cells. In addition to the training modules and lab courses,

continued on page 4

POSSIBLE RECIPE FOR A BIOMEDICAL DATA SCIENCE PROGRAM:

Start with a biostatistics program

Add computational topics such as optimization and algorithms

OR

Start with a biomedical informatics, bioinformatics or computational biology program*

Add advanced statistical concepts such as machine learning and modeling techniques for complex data

Next:

Mix with exposure to multiple data and disease types

Blend in modern data visualization and data management technologies

Combine with interdisciplinary mentorship

Stir with collaborative teamwork

Bake for about four years and voilà!, you've produced a biomedical data scientist.

* Starting from scratch with a biomedical sciences program is also possible.

mobilize news

Boyd's Convex Optimization MOOC, on the Stanford OpenEdX platform, is for more advanced and mathematically-oriented students who want to get into the optimization game. It includes about 20 hours of lecture and some challenging problem sets with an applied focus. "You'll learn just enough math, which by the way is not a small amount, to be able to do convex optimization in practical settings," Boyd says in the online intro to the course.

Professor Stephen Boyd teaches a MOOC on convex optimization.

While none of these MOOCs has a biomedical focus, their applicability is quite wide, Hastie says. "The kinds of methods we teach are used in biomedical computations all the time." At the Mobilize Center, for example, statistical learning is used to analyze data from clinical databases to predict the outcomes of surgeries. And Leskovec is helping the Center mine massive datasets from mobile sensors to better understand patterns in physical activity. □

big data highlight

trainees will discuss progress on their various projects in an ongoing seminar course as a way of further solidifying their collaborative skills.

Mentors and Real Clinical Data

At the University of California, Los Angeles (UCLA), the students funded by the BD2K training grant may have less diverse skill sets than those in the UNC program—most will be Bioinformatics Program students in the second and third years of study who seek specific training related to working with massive biomedical datasets—but the program is nearly as interdisciplinary, with approximately 30 faculty mentors from eight departments participating. Students will complete coursework in data analysis as well as in breaking down various aspects of clinical science such as medical ontologies and electronic records. But the focus of the UCLA program is mentorship and real data. Trainees must work with two mentors—one with big data expertise and the other with a clinical medicine background, says **Matteo Pellegrini**, **PhD**, professor of biology at the University of California, Los Angeles (UCLA) and principal investigator on the UCLA grant. The hope is that by immersing themselves in both fields, trainees will get an understanding of how clinical genomic data is collected and how it is interpreted.

The UCLA program also emphasizes getting trainees' feet wet with real, massive-scale biomedical data, such as sequencing, proteomic and clinical data. Trainees will compete against each other in big data challenges in which they will develop machine-learning algorithms to predict disease outcomes or risk based on big data resources unique to UCLA, including data sets related to bipolar disorder, depression, autism and breast cancer.

Adding a Big Data Track to a Biomedical Informatics Program

At Columbia University, the Biomedical Informatics Department is creating a new track called "Biomedicine and Health Data Science" thanks to its BD2K training grant. Whereas doctoral students in the overall biomedical informatics program study a wide swath of biomedical informatics topics, the new track reflects the increased prevalence of observational health data, says, **Noémie Elhadad**, **PhD**, associate professor of biomedical informatics and director of Columbia's BD2K grant. Trainees will focus on developing high-throughput methods specific to healthcare, utilizing massive amounts of biomedical knowledge and health-related data coming from the biomedical literature, the Internet, self-reported health data, and electronic health records.

One crucial aspect of the new track will be training students to seamlessly integrate a variety of evolving data types into a full picture of individual patient health as well as public health–related issues. Lab tests, diagnostic codes, and continuously generated data from wearable sensors all need to be woven into a single framework. In addition, says Elhadad, natural language processing will be important for capturing various "free text" formats such as clinician notes, online health community discussion forums, tweets and other social media pertinent to an individual's health.

Big Data Equals Big Opportunities

In addition to earning a certificate or degree designation as big data experts upon graduation, the trainees in each of the three training programs will have opportunities to attend high performance computing and big data workshops or work at summer internships in industry or academia—all great resumé builders. These experiences are expected to give trainees a distinct advantage over their peers. "The grant will make our trainees very competitive for positions in both industry and academia," says Pellegrini.

Kosorok agrees. "Our students will be quite valuable on the job market," he says. "For nearly all of my recent students, expertise with big data has been a big part of their being hired." □