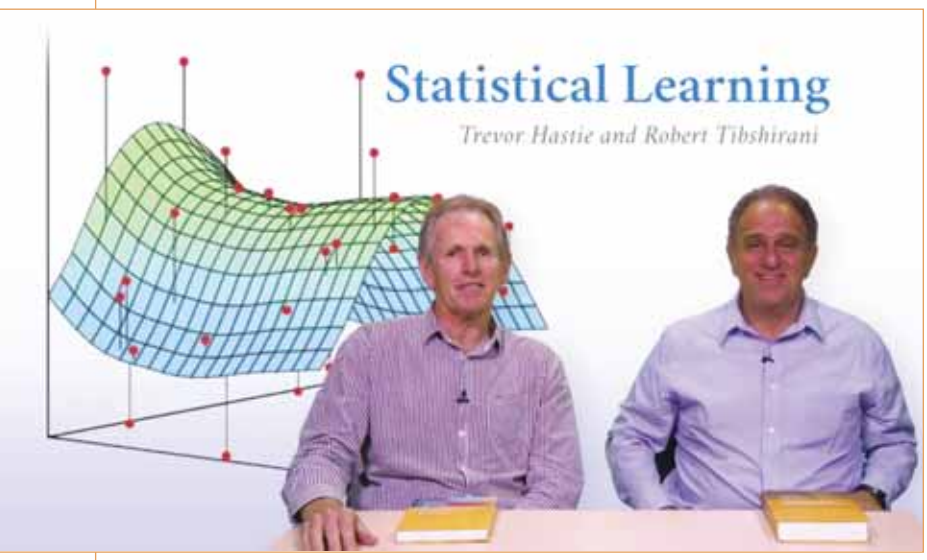
BY KATHARINE MILLER

Free Doses of Data Science

Want to dip a toe in data science? Why not take a MOOC (massively open online course) from someone who literally wrote the book on the topic at hand?

Several MOOCs offered by Stanford professors who are part of the Mobilize Center fit the bill. **Trevor Hastie, PhD**, professor of statistics, co-wrote *Introduction to Statistical Learning*; **Jure Leskovec, PhD**, assistant professor of computer science, co-wrote *Mining Massive Datasets*; and **Stephen Boyd, PhD**, professor of electrical engineering, co-wrote *Convex Optimization*. And each of them teaches a MOOC by the same name.



Statistics professors Trevor Hastie and Rob Tibshirani co-teach a MOOC on statistical learning.

In Hastie's case, the book inspired the MOOC. "We had a book that was at the right level for a MOOC so we decided we'd do it." He and **Robert Tibshirani, PhD**, co-author of the book and co-teacher of the MOOC, also made a deal with the publisher: The book became free online just six months after publication. It's an extra draw for students—not only is the course free, but the text is as well. The same is true for the Mining Massive Datasets MOOC.

The statistical learning MOOC, offered on Stanford's OpenEdX platform, has proven popular with people looking to broaden their horizons. "They get a free dose of what the field is like, especially now that data science is so popular," Hastie says. "And they can decide whether to make a career move."

Hastie's MOOC follows the structure of the *Introduction to Statistical Learning* text. Typically, it's appropriate for people who did a little bit of statistics in college, he says. "It gets them into more modern-day applied statistical modeling and how to implement with software." The MOOC has been taught twice, with nearly 40,000 people signing up each time, 20,000 showing up on day one, and about 3,000 to 4,000 completing each course. This is typical of MOOCs, Hastie says: "There's a kind of exponential decay [in the number of students]." But the MOOC still reaches more people than is possible in a traditional in-person class.

Leskovec's MOOC, which is offered through Coursera, introduces fundamental algorithms and techniques for dealing with very big data as well as how to apply these techniques efficiently at large scales. The course covers algorithms for extracting models and information from large datasets, including locality-sensitive hashing, clustering, decisions trees, and dimensionality reduction. It also introduces students to MapReduce, a software framework for easily writing applications that process vast amounts of data. Offered on Coursera, the MOOC had over 54,000 people visit the course, of which over 9,800 submitted at least one exercise.

continued on page 4

Jure Leskovec, assistant professor of computer science at Stanford, co-teaches a MOOC on mining massive data sets.

DETAILS

MOOCs:

Statistical Learning:
https://statlearning.class.stanford.edu/

Mining Massive Datasets:
https://www.coursera.org/course/mmds

Convex Optimization:
https://www.class-central.com/mooc/1577/stanford-openedx-cvx101-convex-optimization

The Mobilize Center web site provides a list of other training resources, including videos from the 2015 Big Data in Medicine conference at Stanford. Go to http://mobilize.stanford.edu/training/

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. □

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." □