

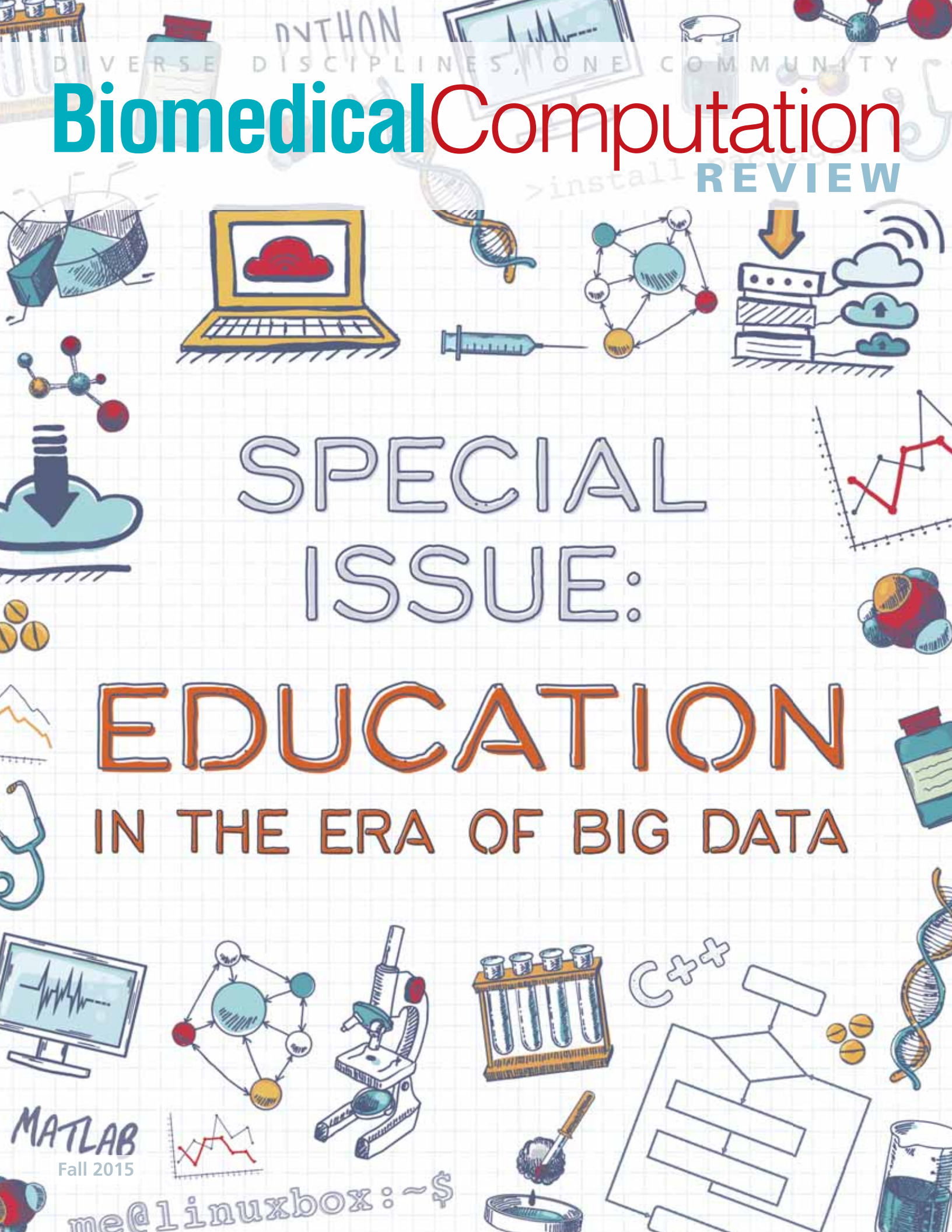
DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation

REVIEW

SPECIAL
ISSUE:

EDUCATION
IN THE ERA OF BIG DATA



MATLAB
Fall 2015

me@linuxbox:~\$

C++

12 The Ever-Expanding and Heterogeneous Landscape of Biomedical Education

BY KRISTIN SAINANI, PhD

20 Career Paths: A Seller's Market for Biomedical Data Science Jobs

BY ALEXANDER GELFAND

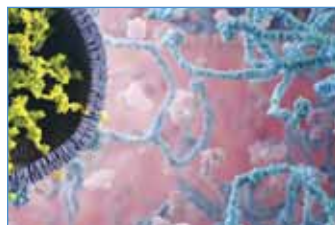


DEPARTMENTS

- 1** GUEST EDITORIAL | TRAINING FOR THE FUTURE BY MICHELLE DUNN, PhD
- 2** MOBILIZE NEWS | FREE DOSES OF DATA SCIENCE BY KATHARINE MILLER
- 3** BIG DATA HIGHLIGHT | A BOOST FOR BIOMEDICAL DATA SCIENCE TRAINING: BD2K GRANTS PUSH THE FIELD BY CHRIS PALMER
- 5** EARLY BLAST OFF: BRINGING BIOINFORMATICS TO SECONDARY SCHOOLS BY ESTHER LANDHUIS
- 7** SKILLS UPGRADES: BD2K BUILDS TRAINING RESOURCES BY KATHARINE MILLER
- 28** UNDER THE HOOD | LEARNING ENGINEERING: LEVERAGING SCIENCE AND TECHNOLOGY FOR EFFECTIVE INSTRUCTION BY NORMAN BIER
- 30** SEEING SCIENCE | ANIMATION & INSPIRATION BY KATHARINE MILLER

Cover Art: Created by Rachel Jones of Wink Design Studio using sketches © Macrovector | Dreamstime.com, © Aldanna | Dreamstime.com, and © Antishock | Dreamstime.com.

Page 12 Art: Digital landscape from RuleByArt.com. **Page 20 Art:** Created by Rachel Jones of Wink Design Studio.



Fall 2015

Volume 11, Issue 3
ISSN 1557-3192

Co-Executive Editors

Scott Delp, PhD
Russ Altman, MD, PhD

Associate Editor Joy Ku, PhD

Managing Editor Katharine Miller

Science Writers

Alexander Gelfand, Esther Landhuis,
Katharine Miller, Chris Palmer,
Kristin Sainani, PhD

Community Contributors

Michelle Dunn, PhD
Norman Bier

Layout and Design

Wink Design Studio

Printing

Advanced Printing

Editorial Advisory Board

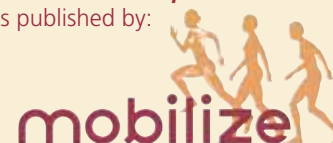
Ivet Bahar, PhD, Jeremy Berg, PhD,
Gregory F. Cooper, MD, PhD,
Mark W. Craven, PhD, Jiawei Han, PhD,
Isaac S. Kohane, MD, PhD,
Santosh Kumar, PhD, Merry Lindsey, PhD,
Avi Ma'ayan, PhD, Mark A. Musen, MD, PhD,
Saurabh Sinha, PhD, Jun Song, PhD,
Andrew Su, PhD, Paul M. Thompson, PhD,
Arthur W. Toga, PhD, Karol Watson, MD

**For general inquiries, subscriptions,
or letters to the editor,
visit our website at www.bcr.org**

Office

Biomedical Computation Review
Stanford University
318 Campus Drive
Clark Center Room S271
Stanford, CA 94305-5444

Biomedical Computation Review
is published by:



mobilize
The Mobilize Center, an NIH
Big Data to Knowledge (BD2K)
Center of Excellence
mobilize.stanford.edu

Publication is supported by NIH Big Data to Knowledge (BD2K) Research Grant U54EB020405. Information on the BD2K program can be found at <http://datascience.nih.gov/bd2k>. The NIH program and science officers for the Mobilize Center are:

Grace Peng, National Institute of Biomedical Imaging and Bioengineering,

Theresa Cruz, National Institute of Child Health and Human Development, and

Daofen Chen, National Institute of Neurological Disorders and Stroke

BY MICHELLE DUNN, PhD, SENIOR ADVISOR FOR DATA SCIENCE TRAINING, DIVERSITY, AND OUTREACH AT THE NATIONAL INSTITUTES OF HEALTH



Training for the Future

This issue of *Biomedical Computational Review* (BCR) is about transforming the biomedical workforce into a biomedical big-data workforce. It focuses on one of the most pressing problems facing biomedical science today: evolution of the workforce in response to the flood of data. Because the rate of that evolution is affected by the availability of appropriate training and education, this issue of BCR is a timely and welcome addition to the continuing conversation on the topic.

The biomedical Big Data workforce is diverse—it includes biomedical scientists who are users of Big Data, data scientists who develop methods, data engineers who build tools, as well as librarians, who organize and manage data. Although each of these requires a different mix of skills, core areas are common to all of these groups and include three components: 1) an understanding of the processes represented by the data (biomedical knowledge), 2) the facility to handle and manage the data (computational skills), and 3) the knowledge to conduct and interpret analyses and draw conclusions (statistical skills).

These three components—computer science, statis-

tics, and a type of biomedical science—are each whole fields of study in their own right. Acquiring expertise in one area is a long process; acquiring expertise in multiple areas is something very few people will achieve. A more realistic goal for this age of Big Data is to have many individuals with some knowledge in all three areas along with expertise in at least one area.

Bayes risk, often approximations must be made to reduce the computational cost; approximations done in a skillful, deliberate, and measured way, with attention to diagnostics, can maintain interpretability and reliability of results. Tool developers bring skills that turn ideas and prototypes into hardened products through creative algorithm design based on a thorough understanding of the computational framework being used. Although method developers are often tool developers (and vice versa), separating the description of the roles illuminates the tradeoff between time and accuracy that often exists. The competing demands between computational cost and confidence in results need to be weighed by the whole team, including biomedical scientists, method developers, and tool developers.

Intra-team communication is essential for a group of researchers to function well. Communication is aided by having enough overlapping expertise to be able to trans-

Ideally, all scientists would know (and trainees would be taught) how to discover and conduct standard analyses of their data and, more importantly, how to determine when standard analyses are not appropriate.

Ideally, all scientists would know (and trainees would be taught) how to discover and conduct standard analyses of their data and, more importantly, how to determine when standard analyses are not appropriate. When standard analyses break down and new challenges are discovered, collaborations with method and tool developers—a group that has been collectively referred to as biomedical data scientists—are needed.

Method developers often recognize that doing a precise principled analysis is computationally infeasible. Although a model-based analysis may have foundations in principles such as maximizing likelihood or minimizing

late from one field to another, as each field may have its own specialized language. Many departments are already making a conscious effort to build overlap between fields by, for example, adding more statistics training to bioinformatics programs and more computational training to biostatistics programs.

A goal of the NIH Big Data to Knowledge Initiative is to foster the development of training and education opportunities that enable trainees and scientists to gain the skills needed to contribute most effectively to biomedical Big Data teams. Awards issued in the past year—for Big Data training programs, courses, open educational resources, and career development—represent early efforts toward transforming the biomedical workforce into a biomedical Big Data workforce. Achieving this goal will bring challenges, some of which are illuminated in this issue of BCR. With challenges come opportunities, and the NIH is keen to seize those opportunities, and ultimately, to turn data into discovery into health. □

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.

In Hastie's case, the book inspired the MOOC. "We had a book that was at the right level for a MOOC so we de-

cidated 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.

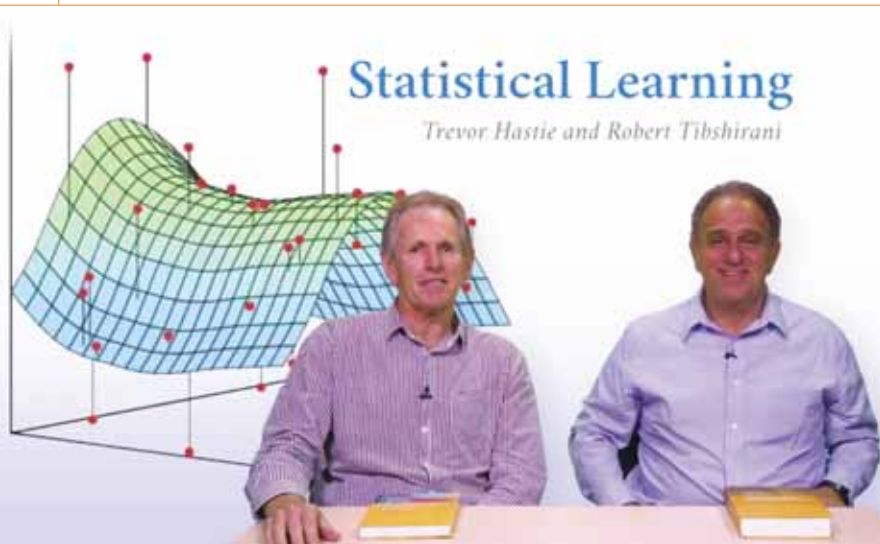
The statistical learning MOOC, offered on Stanford's OpenEdX platform, has proven popular with people looking

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



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

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.

In Hastie's case, the book inspired the MOOC. "We had a book that was at the right level for a MOOC so we de-

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/>



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

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 Out-

reach 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.

Boyd's Convex Optimization MOOC, on the Stanford OpenEdX platform, is for more advanced and mathematically-oriented students who want to get into the optimiza-



Professor Stephen Boyd teaches a MOOC on convex optimization.

tion 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.

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—but 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 health-care, 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." □

EARLY BLAST OFF: Bringing Bioinformatics to Secondary Schools

By Esther Landhuis

Two decades ago at a genomics workshop for educators, a high school biology teacher isolated DNA from a snippet of his hair and got it sequenced. He then used a computer algorithm to compare his DNA sequence to that of Neanderthals, ancient primates that roamed Earth some 200,000 years ago.

that program Form fiddled with 20 years ago to explore the Neanderthal genome is now a researcher's mainstay. Called Basic Local Alignment Search Tool (BLAST), the algorithm scans multiple DNA or protein sequences for similar regions. It is used to determine evolutionary relationships and gain insight into genetic diseases.

"It's about using bioinformatics tools to make biological discoveries," says **Fran Lewitter, PhD**, founding director of the bioinformatics and research computing department at the Whitehead Institute in Cambridge, Massachusetts.

In the classroom, Form shows his biology students how to use BLAST and other

For middle and high school students, it's not about learning the technical know-how. "It's about using bioinformatics tools to make biological discoveries," says Fran Lewitter, PhD.

It was not only an auspicious journey into human history but also "my introduction to bioinformatics," says **David Form, PhD**, who teaches at Nashoba Regional High School in Bolton, Massachusetts. His next thought: "This would be exciting for students."

When Form started teaching, genome sequencing and bioinformatics were still new-fangled. But biomedical data has since exploded in volume and complexity, and

Building BLAST and other computational resources requires programming skills and advanced mathematics, mostly taught at the undergraduate and graduate levels. So why is it important to introduce bioinformatics in secondary schools?

Facilitating Biological Discovery

For middle and high school students, it's not about learning the technical know-how.

computational tools so they can explore human diseases. "They learn which genes and proteins are involved, find a suitable lab model to study the condition, and go from there," Form says. "It's an active way for students to learn some very sophisticated things."

Though Form was self-motivated to develop his bioinformatics curriculum—buying books to bone up on bioinformatics after that first workshop 20 years ago—biology teachers today have a greater imperative to train their students to use computational tools. In 2013, the Advanced Placement (AP) biology exam was revised to include bioinformatics. Test takers are given BLAST data to interpret evolutionary relationships depicted in phylogenetic trees.

Training Workshops

Various institutions—including Cold Spring Harbor Laboratory, Whitehead Institute, Harvard University, University of Utah and Marine Biological Laboratory at Woods Hole Oceanographic Insti-

David Form speaks on how to teach high school students bioinformatics at an educator's workshop at the Whitehead Institute. Photo credit: Ceal Capistrano/Whitehead Institute.



tute—now offer crash courses in modern genetics and bioinformatics to equip teachers to introduce these concepts in their secondary school classrooms. The International Society for Computational Biology, a pro-



Joanne Fox (center) works with high school students in the teaching laboratory/educational facilities at the Michael Smith Laboratories. Courtesy of Joanne Fox.

fessional society headquartered at the University of California, San Diego, and the European Molecular Biology Laboratory in Heidelberg, Germany, also offer teacher training in bioinformatics.

High school teachers who attended week-long summer bioinformatics workshops in 2008 to 2012 at Franklin & Marshall College in Lancaster, Pennsylvania, not only left with valuable wet lab and computer skills. They also earned professional development credit, made \$25 per hour and received an additional \$250 toward classroom supplies. “We value teachers’ time and wanted to make the experience economically attractive for them,” says **Ellie Rice, PhD**, senior adjunct professor of biology, who directed the teacher outreach program. During the week, attendees created bioinformatics-based classroom activities and put their lesson plans on the publicly available Bioinformatics Activity Bank.

Though training workshops give a good introduction to the nuts and bolts of cutting-edge biology, some teachers still come away feeling a bit daunted teaching the material to their students, Form notes. Going into a lab and doing an activity once isn’t enough, he says. “You need at least three days to explore the data with someone who knows what they’re doing.” Many workshops compress the bioinformatics portion into a single day or afternoon. Still, many high school biology teachers, especially those teaching AP or other advanced classes, do incorporate some bioinformatics into their lessons after attending even a brief workshop.

Field Trips

Other educational outreach efforts take a different approach. Rather than trying to

equip teachers to bring new concepts into the classroom, some programs invite teachers to bring their classes to a research facility on campus. There, graduate students and scientists teach the visiting students to perform lab experiments and use bioinformatics tools. “Field trips give students the view that science is accessible—something they can see themselves doing,” says **Joanne Fox, PhD**, senior instructor in the Michael Smith Laboratories at the University of British Columbia in Vancouver, Canada. “There’s power in being in an actual research space and talking with scientists as you do the activity.”

Fox helped develop a genomics outreach program that launched in 2009, bringing 1,500 to 2,000 students each year to the Michael Smith Laboratories. When her team first piloted the program, they wanted to get students onto computers browsing genome data right from the get-go. “But the students found [the computational tools] very abstract,” notes Fox. “They didn’t see the connection to their own lives.”

Now, rather than going straight to the computers, students begin by isolating their own DNA from a cheek swab. “It’s very impactful for them to see the goeoy DNA,”

Fox says. It also helps equip students to engage in discussions related to personal genomics—such as prenatal screening to test for genetic diseases in unborn babies. The students also get acquainted with bioinformatics as they learn how to transcribe, translate and align DNA sequences.

Bringing the Lab into Classrooms

Instead of bringing classes into research labs to learn lab techniques and bioinformatics, other programs do the reverse—they bring the lab into the classroom. Every year since 2006, bioinformatics@school has deployed its DNALab to some 60 high schools in the Netherlands, reaching a total of 17,000 students. “We go into classrooms and run a two-hour lesson,” says education coordinator **Judith Rotink, MS**. “We bring our own computers, even our own networks.” Funded by Radboud University Medical Center and a government grant, the program is free of charge to schools and taught by science students at Radboud University in Nijmegen.

Teachers can also download guides and lessons to lead the course themselves. In one project entitled “Murder at the Airport,” students learn that a man lies dead on the floor next to a bottle containing a liquid. The students use online databases to figure out which of four suspicious proteins in the liquid was responsible for the man’s death.

“Bioinformatics tools can make abstract concepts more understandable,” Rotink says. □

RESOURCES:

DNA Learning Center (Cold Spring Harbor):
https://www.dnalc.org/programs/teacher_training.html

Bioinformatics workshops, courses and course materials (Whitehead Institute): <http://jura.wi.mit.edu/bio/education/>

Harvard Life Sciences Outreach programs:
<http://outreach.mcb.harvard.edu/index.htm>

Genetic Science Learning Center (University of Utah):
<http://learn.genetics.utah.edu/>

Professional development workshops (Marine Biological Laboratory at Woods Hole Oceanographic Institute):
<http://discover.mbl.edu/workshop.htm>

Howard Hughes Medical Institute (HHMI):
<http://www.hhmi.org/>

A philanthropic organization that supports US biomedical research, produces online multimedia resources for educators, including a series of virtual labs that let students practice wet lab techniques such as DNA sequencing and polymerase chain reaction (PCR), as well as learn to analyze DNA sequences, in an interactive setting on their own computers.

SKILLS UPGRADES: BD2K Builds Training Resources

By Katharine Miller

The field of biomedical data science is growing so fast that it threatens to leave some researchers behind.

“Some of these big data skills were not needed 5 to 10 years ago, and many of the tools that we now use were simply not available,” says **Daniela Witten, PhD**, associate professor of biostatistics and statistics at the University of Washington. These skills and tools are not part of the standard curricula at many universities.

To make a dent in this problem, the National Institutes of Health (NIH) have funded a number of Big Data to Knowledge (BD2K) training grants designed to develop a variety of training resources, including summer workshops; massive online open courses (MOOCs); and repositories for training materials, including materials at the bleeding edge of educational philosophy. The grants, which were made a year ago, are already making a difference.

“It was a good decision by the NIH to offer a smorgasbord of learning opportunities around the general topic of biomedical big data,” says **Rommie Amaro, PhD**, assistant professor of chemistry and biochemistry at the University of California, San Diego (UCSD).

Summer Boot Camps

Some researchers who want to gain new skills seek out the intensive learning experience provided by a summer training institute. Two BD2K-funded workshops launched this summer—one at the University of Washington and the other at the Mayo Clinic in Rochester, Minnesota.

Short courses of this type appeal to a broad spectrum of people, Witten says, from PhD students and post-docs to research scientists or faculty. “None of us is an expert in everything,” she says. “I’m only teaching one of these modules for a reason; for some of the others, I’m learning along with the other students.”

This year, the Summer Institute for Statistics of Big Data, run by Witten and her colleagues, consisted of five separate 2.5-day-long courses or modules covering how to access biomedical big data; data visuali-

zation; reproducible research; and both supervised and unsupervised methods for statistical machine learning.

“Whether you are a current student or were trained 20 years ago, chances are you don’t know this stuff,” Witten says. The institute has been incredibly popular. “We quickly ran into a room capacity problem,” Witten says. After 150 people enrolled, they had to turn people away for some of the modules. “We’re clearly filling an unmet need,” she says.

The BD2K grant paid for the instructors as well as tuition and scholarships for attendees, with a maximum of three scholarships (for three modules) going to one person. “We have some people staying for all five modules,” Witten says. “That’s a big time commitment.”

Witten sees a clear benefit to the in-person classroom experience afforded by her summer institute. “Being there, talking to other students with teaching assistants walking around—it’s really a hands-on computational experience,” she says.

Mayo Clinic’s Big Data Coursework for Computational Medicine offered six modules:

- 1) data and knowledge representation standards;
- 2) information extraction and natural language processing;
- 3) visualization analytics;
- 4) data mining and predictive modeling;
- 5) privacy and ethics; and
- 6) applications in comparative effectiveness research and population health research and improvement.

“None of us is an expert in everything,” Witten says. “... I’m learning along with the other students.”

The Mayo short course, called Big Data Coursework for Computational Medicine, was also in high demand, with 80 applicants for 20 spots. “There seems to be a lot of interest in spending summer vacation in a boot camp,” says **Claudia Neuhauser, PhD**, who directs the Institute of Informat-

ics at the University of Minnesota Twin Cities. She co-leads the BD2K program with **Jyotishman Pathak, PhD**, professor of biomedical informatics at the Mayo Clinic.

By its nature, a weeklong intensive course covering six topics in six days can't go very deep. "The workshop gives them pointers and literature references and exposure. That entrée then lets them dig in further," Neuhauser says. "Many of the students are used to learning on their own."

Students work together and learn from each other. "There's almost always someone in the room who is experienced and someone who isn't," Neuhauser says. "The diversity of students means the questions are quite rich."

Neuhauser and Witten agree that both in-person workshops and MOOCs are needed to address the training gaps of biomedical data science. But live workshops allow greater interactivity. "Being in a group for a whole week—talking about things and asking questions, even ones that go off topic—allows students to get what they

biomedical data science. Like summer institutes, MOOCs in biomedical big data science serve a heterogeneous group of people who want to retool or get involved in a new area such as genomics, says **Brian Caffo, PhD**, professor of biostatistics at the Johns Hopkins Bloomberg School of Public Health.

Caffo and his team developed the first data science specialization on Coursera, the popular education platform that partners with top universities and organizations worldwide to offer free courses online. A specialization is a program of study—a bundle of courses designed to be taken serially. Caffo is now using a BD2K training grant to launch two new Coursera specializations in genomics and neuro-imaging. "Our specializations are longer than a summer institute but a tad shorter than a full-on masters' degree," Caffo says.

For the new genomics program, which started in the summer of 2015, many of the students are people who want to work in the field or already work in the field and need genomics skills for their current jobs. "We've had some people say 'our whole office is doing this.' Or 'I make all new employees do it,'" Caffo says.

One benefit of MOOCs: They are typically free, and students can choose their level of engagement. The genomics series can be completed in about six months if taken serially, Caffo says. But students can take modules simultaneously or out of order. And those in the data science specialization who choose to fork over a nominal fee (\$50 or less) to get Coursera signature track verification also get another bonus—a project-based class available only to those who pay. And the MOOC completion rate for people who make this minor investment is quite high—on the order of 90 percent, Caffo says.

With his BD2K funding, **Rafael Irizarry, PhD**, professor of biostatistics at the Harvard School of Public Health and professor of biostatistics and computational biology at the Dana Farber Cancer Institute, revised a very dense data science for genomics MOOC he launched two years ago by dividing it into eight parts. To complete the series, a student takes the first four modules and one of the last four, which are case studies using specific types of data. "But a generalist in genomics might want to take all of them because someday they might face all of those types of data," Irizarry says.

The course has proven quite successful, with the usual caveat: Many people sign up for MOOCs and don't finish them. In some cases, perhaps they watch a few lectures and learn what they needed to know. For Irizarry's

"THERE'S ALMOST ALWAYS SOMEONE IN THE ROOM WHO IS EXPERIENCED AND SOMEONE WHO ISN'T," NEUHAUSER SAYS.

"THE DIVERSITY OF STUDENTS MEANS THE QUESTIONS ARE QUITE RICH."

want out of the workshop," Neuhauser says. "And we can adjust how we teach."

In-person trainings provide another key benefit: networking opportunities. Often, says Neuhauser, bioinformatics and health informatics researchers can be the lone quantitative people in their workplaces. "So talking to others can be very important," Neuhauser says. "This kind of work doesn't have a recipe book. There's a lot at the arts level where you have to figure it out. Personal contact becomes important."

Biomedical Big Data MOOCs

Several BD2K training grants are being used to launch new MOOCs for teaching

MOOC, two or three thousand completed the first of the eight modules—a very general statistics course for the life sciences—and about 300 completed the entire series of eight. Of these, Irizarry says many are post-docs who want to be better able to do their jobs. Another subset, he says, are educators—“people tasked with teaching this kind of thing who take the class to help them prepare a class.”

Building a Better MOOC

Caffo and Irizarry are both serious about incorporating interactive learning into their MOOCs. So too is **Pavel Pevzner, PhD**, professor of computer science at the University of California, San Diego. He also received BD2K funding to launch several MOOCs. Indeed, Pevzner wants to change the nature of MOOCs from being massive and impersonal to being more like the experience of receiving one-on-one tutoring in a professor’s office. “There’s a need to address individual breakdowns in students’ learning,” Pevzner says. Large lecture courses don’t and can’t do that.

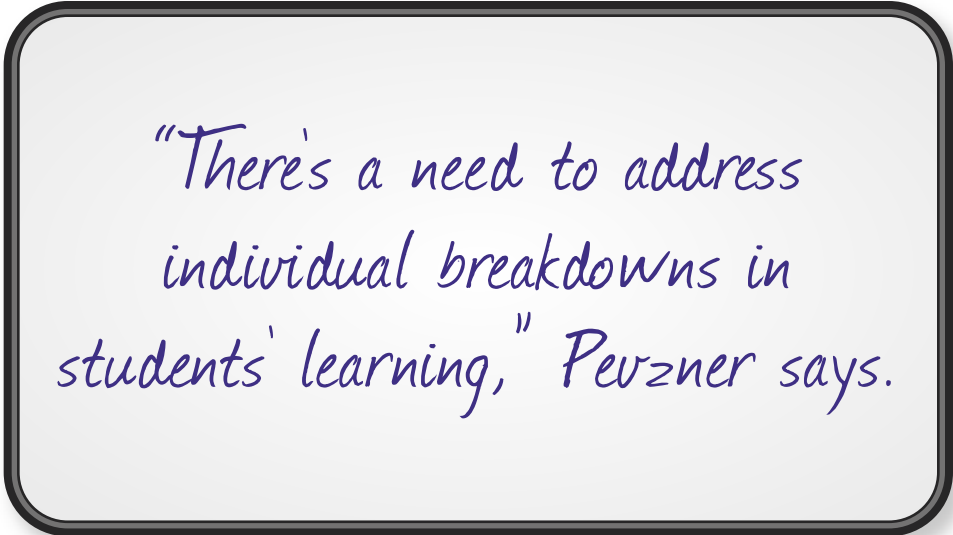
“We wanted to build a better MOOC,” Pevzner says. And he’s been at it for a while, having created a MOOC for bioinformatics algorithms several years ago. The key to a better MOOC, he says, is short lectures (under ten minutes), and intelligent tutoring systems such as one called Rosalind that Pevzner created, or another called SWIRL that Caffo uses as part of his MOOCs. Rosalind allows for automated individualized assessments of students’ work on robust, “just-in-time” assignments that are evaluated using a sophisticated software system at the exact moment that assessment would facilitate the transition to the next topic.

Similarly, SWIRL is an active learning tool for learning data analysis using the programming language R. “It prompts you to do things, and if you mess up it asks you to do it again,” Caffo says. SWIRL, which is free and open source, was developed by **Nick Carchedi** in 2013 while he was pursuing his master’s degree in biostatistics at Johns Hopkins. “It is now very mature,” Caffo says. “We’re focused on making content for it.”

Pevzner has seen professors at other universities use his bioinformatics algorithm MOOC in a flipped classroom—students watch the videos and do the lessons outside of class and come to class to discuss and work through any questions or problems they are having. This is a sign that his approach has to some extent succeeded, Pevzner says. Indeed, he believes MOOCs of the future will

turn into MAITs, Massive Adaptive Interactive Text. His paper outlining the concept of MAITs appeared in *Communications of the Association of Computing Machinery (CACM)* in September 2015.

Like Pevzner, Irizarry’s MOOC avoids multiple-choice assessments (widespread in the MOOC world) because, he says, they aren’t effective teaching tools. Instead, the BD2K-funded MOOCs he’s developing use fill-in-the-blank questions that offer the student multiple chances to get it right. “They have to download data, analyze it the way they think best, and tell us, for example, how many genes show statistically significant differences in cancer samples com-



“There’s a need to address individual breakdowns in students’ learning,” Pevzner says.

pared to controls,” Irizarry says. “There’s a correct answer (it might be a specific number, like 154). And many times they get it wrong. Then they go to discussion boards and talk about it.” With repeated effort to solve a problem, and the support of the people on the board, students often get the question right. And if they don’t, the correct approach is revealed, along with a follow-up question to ensure students really understand the material.

“The discussion board can get pretty crazy,” Irizarry says. With thousands of students, a single question can generate several hundred posts. And while that might sound like a lot of work for the professors, there are usually students in the class who answer other students’ questions before the professors do. Once they’ve proven their reliability, Irizarry can tag these individuals as community TAs, alleviating the discussion board burden.

Michelle Dunn, PhD, senior advisor for data science training, diversity, and outreach in the Office of the Associate Direc-

tor for Data Science at NIH, is enthusiastic about MOOCs such as Caffo's data science specialization. "The fact that they can put out thousands of students per year through a mini-masters program can only help the rest of us have the quality people we need in order to get the job done," she says.

Some MOOC graduates might become the programmers who are given direction about what algorithms to program. Others with advanced biology backgrounds have used MOOCs to obtain needed data science skills. "People with PhDs are self-learners and do well with MOOCs," Dunn notes.

Courselets and Concept Inventories

Another BD2K grant recipient is applying the latest advances in educational psychology to bioinformatics education and making the results available online.

After more than 10 years teaching bioinformatics theory to computer scientists, physicists and life science students, **Christopher Lee, PhD**, professor of chemistry and biochemistry at the University of California, Los Angeles, felt discouraged. "After a quarter-long class, students were still not understanding basic things," he says. In addition about 50 percent of his students were dropping the class.

Then he learned about concept inventory studies from the field of education.

shocking results," Lee says. "And this is universal." The same phenomenon is seen in many fields. To Lee, this matched up with his frustration in teaching introductory bioinformatics.

To address the problem, Lee changed his teaching methods. He now presents a single concept and then, immediately after, poses a question designed to test understanding of that concept. Students then have a few minutes to think about how the concept applies to the question and write an answer on a web page on their laptop—just a few lines to capture their thinking. "We can then identify the underlying conceptual errors that we are seeing in all the students' answers," Lee says.

As an example, Lee says, students in his class should understand the concept of conditional and unconditional probability from prior coursework in statistics and probability. But, he says, "My experience is that their understanding is brittle and falls apart when they try to use it." Shifting to concept-based instruction has proven helpful in bringing students up to speed.

"As soon as I started doing this, it was eye-opening," Lee says. He could see what every student was thinking on every concept every single day. "I'd realize that one word has two meanings and half the class is off on a wrong tangent. It's wrong and nobody's going to repair it for them."

After three years of teaching the introductory bioinformatics course this way, the attrition rate dropped from 50 percent to about 10 percent—without any detriment in overall test scores, Lee says. "So we've taken the lower half of the class (the ones who dropped) and put them up where the top half were."

For his BD2K concept network grant, Lee is taking all that he's learned from his work with concept inventories in his introductory bioinformatics course and putting it online as courselets that any teacher or student can use. A courselet can allow someone to understand a concept really well in a single sitting. It includes a brief explanation and definition followed by exercises that are broken into pieces: question, answer, and error models—common misconceptions—as well as resolutions for various error models.

Courselets.org is still in the early stages (the user interface needs refinement and Lee needs to do some usability studies), but having it online allows others to dip a toe into Lee's methods by trying out one or two concept exercises a week, either in the classroom or as homework. Lee's team will also provide support for instructors who use the platform. "We have a lot of experience creating these concept tests,"

"We have a lot of experience creating these concept tests," Lee says. "and we are totally willing to work closely with people to do this."

About 20 years ago, researchers discovered that students of freshman physics—including bright Harvard freshmen—scored about 45 percent on a test of physics concepts before taking the class, and only about 50 to 55 percent immediately afterward. "It got peoples' attention because of these pretty

he says, “and we are totally willing to work closely with people to do this.”

Lee also plans to cross-link Courselets with Rosalind, Pevzner’s interactive learning site. Eventually, he says, “If you are working on a Rosalind problem and you feel that you are missing a concept, you can jump over to Courselets.”

Repositories and Virtual Machines

Summer short-courses, MOOCs and Courselets will serve a vast constituency, but plenty of principal investigators just want to train the students in their lab or in a class of 15 to 20 people. These folks don’t necessarily need to launch a MOOC, Amaro says, but they could benefit from a resource of plug-and-play training materials.

Amaro and her co-PI, **Ilkay Altintas, PhD**, Director for the Center of Excellence in Workflows for Data Science at the San Diego Supercomputer Center (SDSC), UCSD, received a BD2K grant to build such a resource. Called the Biomedical Big Data Training Collaborative (BBDTC), it will serve as a sort of clearinghouse for training materials related to biomedical big data. It will allow instructors and students to create playlists and add them to an educational queue, designing a personalized, flexible, online learning experience, she says. “Instructors can easily create their own modular courses based on the content we serve and what they create, and deploy it to their students in a more flexible way,” Amaro says.

The site also provides virtual toolboxes—virtual machines that will contain all that a student would need to run hands-on exercises. “Instructors can create the environment the students will be working in,” Amaro says. “And we can package these toolboxes up and ship them out with the course materials in a way that scales,” Amaro says.

Amaro’s prototype site is now up and running at biobigdata.ucsd.edu, and she is eager for the BD2K Centers of Excellence and others to put their content there. “As we get content, we’re working on developing tags for the various kinds of training that get uploaded to the BBDTC so people can sort and search and find what they are looking for,” she says.

Choices, Choices, Choices

Online education is not for everyone. “In the end, most people will agree that face-to-face training is always the best,” Amaro says. “But there are so many people

who we need to reach, it’s just not possible to train them all in-person.” Online resources allow training in a scalable way, which will be needed in order to close the gap that exists and that will continue to

“INSTRUCTORS CAN EASILY CREATE THEIR OWN MODULAR COURSES BASED ON THE CONTENT WE SERVE AND WHAT THEY CREATE, AND DEPLOY IT TO THEIR STUDENTS IN A MORE FLEXIBLE WAY,” AMARO SAYS.

exist in the trained workforce, she says.

Caffo agrees that because the demand for trained people outstrips the supply, there’s plenty of room for all different sorts of solutions for training people. “More MOOCs, more institutes, more online degrees, more in-person degrees—all of those things are going to be necessary,” he says. □

DETAILS

The Summer Institute for Statistics of Big Data:
<http://www.biostat.washington.edu/node/2295>

Big Data Coursework for Computational Medicine:
<http://bdc4cm.org/>

Coursera Genomic Data Science Specialization:
<https://www.coursera.org/specialization/genomics/41>

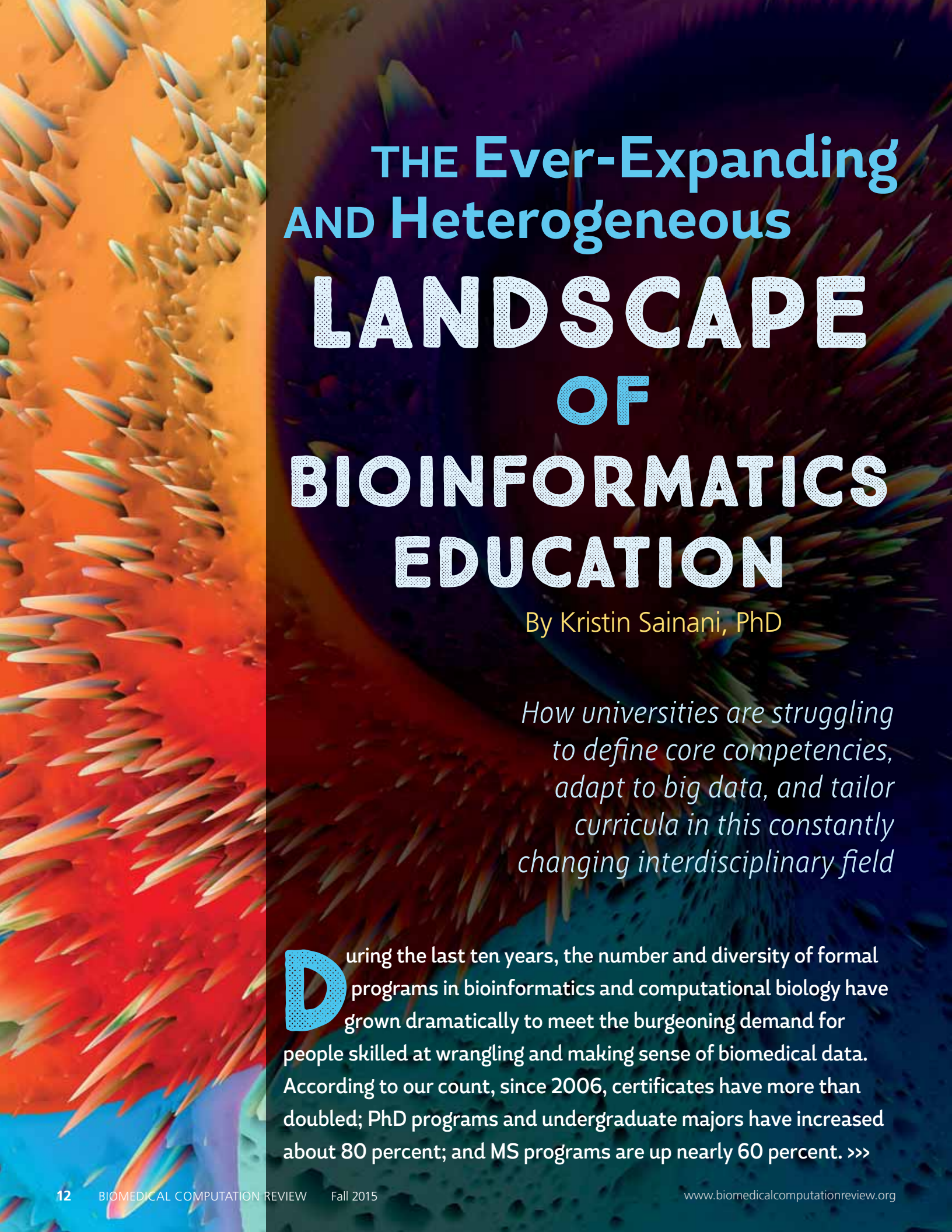
Coursera Data Science Specialization:
<https://www.coursera.org/specialization/jhudatascience/1>

SWIRL:
<http://swirlstats.com/>

Rosalind:
<http://rosalind.info/about/>

Courselets.org

Biomedical Big Data Training Collaborative:
Biobigdata.ucsd.edu



THE Ever-Expanding AND Heterogeneous LANDSCAPE OF BIOINFORMATICS EDUCATION

By Kristin Sainani, PhD

How universities are struggling to define core competencies, adapt to big data, and tailor curricula in this constantly changing interdisciplinary field

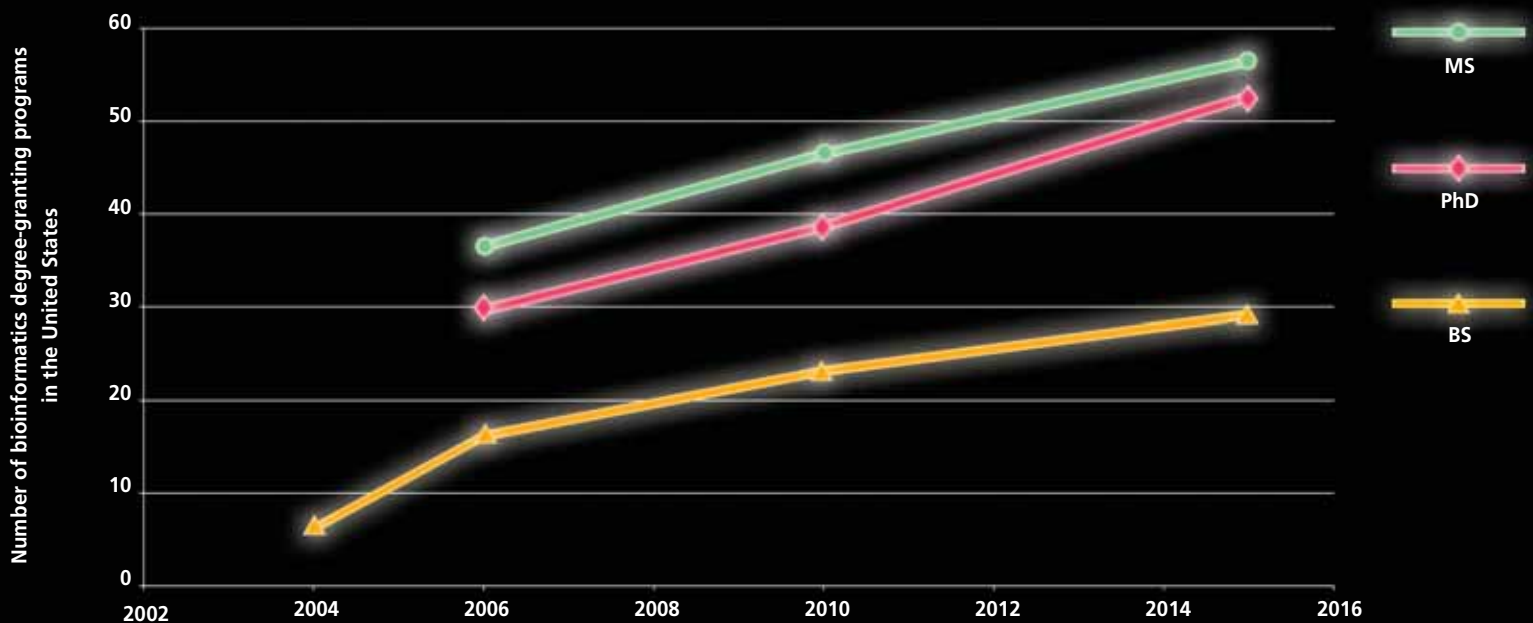
During the last ten years, the number and diversity of formal programs in bioinformatics and computational biology have grown dramatically to meet the burgeoning demand for people skilled at wrangling and making sense of biomedical data. According to our count, since 2006, certificates have more than doubled; PhD programs and undergraduate majors have increased about 80 percent; and MS programs are up nearly 60 percent. >>>



To assess the current educational landscape, *Biomedical Computation Review* talked to directors of graduate and undergraduate programs in this space. The emphases of these programs range widely as do their titles, which include not only bioinformatics, biomedical informatics and computational biology but also health data science, biomedical data science, quantitative biomedical science, and various combinations of these titles with genomics. Despite the diverse names, all of these programs encompass bioinformatics in its broadest sense (the use of computation and statistics to gather, store, analyze, interpret, and integrate data to solve biological problems).

Two key themes emerged. First, there is more than ever to learn. Students have to grapple with the explosion of new technologies and data types. The big data revolution has also upped the ante on computational skills and heightened the emphasis on statistics. “You can’t expect anyone but a superman or superwoman to get all of that knowledge out of graduate school,” says **Michelle Dunn, PhD**, a senior advisor at the National Institutes of Health (NIH), who is involved with the Big Data to Knowledge (BD2K) training initiatives. Second, programs suffer from considerable heterogeneity—both in what and whom they teach. Lacking

This graph shows the number of bioinformatics degree-granting programs in the United States over time. Counts of current programs were compiled from lists maintained by the International Society of Computational Biology (<https://www.iscb.org/iscb-degree-certificate-programs>) and [www.bioinformatics.org \(http://www.bioinformatics.org/wiki/Education_in_the_United_States\)](http://www.bioinformatics.org/wiki/Education_in_the_United_States), as well as from story interviewees. Data for 2006 were available from an archived list created by University of North Carolina (http://ils.unc.edu/informatics_programs/doc/Bioinformatics_2006.html). Data for 2010 and for 2004 (for undergraduate programs only) were available from comprehensive lists compiled by Mark Pauley of the University of Nebraska at Omaha.



formal guidelines, educators have largely pieced curricula together based on local needs and resources. Plus, educators face the daunting challenge of jointly teaching biologists and computer scientists—who come from vastly different cultures with highly variable skill sets.

Bioinformatics educators are confronting these issues on several fronts. Efforts are underway to define the core competencies of the discipline and to recommend key changes for the big data era. Educators are also sorting programs into distinct groups by trainees' goals—and tailoring curricula accordingly. “There’s so much work to be done that we need people across the spectrum,” Dunn says. Educators are also experimenting with new ways to bridge the divide that separates researchers with disparate training backgrounds; and they are leveling the playing field by infusing more interdisciplinary training at the undergraduate level. Finally, several initiatives are creating a plethora of publicly available educational resources to meet training needs at all levels.

DEFINING CORE COMPETENCIES

Bioinformatics is a young, highly interdisciplinary, and rapidly changing field—so it’s been difficult for practitioners to agree on a set of standardized curriculum guidelines.

An early attempt to delineate core competencies appeared in a 1998 *Bioinformatics*

paper by **Russ Altman, MD, PhD**, professor of bioengineering, genetics, and medicine, and the director of the Biomedical Informatics Training Program at Stanford University. Altman laid out proficiencies in five domains: biology, computer science, statistics, core bioinformatics, and ethics.

But a decade and a half later, when **Lonnie Welch, PhD**, was searching for formal curriculum guidelines sanctioned by the International Society for Computational Biology (ISCB), he was surprised to learn that there weren’t any. “I come from the computer science community, where they have a lot of standards and guidelines. And I just assumed that there were such things,” says Welch, professor of electrical engineering and computer science at Ohio University.

Welch—who directs graduate and undergraduate certificate programs in bioinformatics—offered to lead an ISCB task force to remedy this gap. The team surveyed core facility directors and combed job listings and curricula from individual universities looking for cross-cutting themes.

The committee found high variability in what programs are teaching and what students come out knowing—leading to mismatches between students’ skill sets and employers’ expectations. “One thing the core facility directors told us is that oftentimes the skills they most need are lacking in the students they hire,” Welch says. “That’s a wake-up call for us.”

Altman stresses the importance of core programs are falling short. **Casey Greene, PhD**—who has helped to shape Dartmouth’s PhD program in quantitative biomedical sciences—points to weaknesses in statistics, software engineering, and biology training. “Some training programs are failing to teach statistics that are relevant for big data,” says Greene, now an assistant professor of systems pharmacology and translational therapeutics at the University of Pennsylvania’s Perelman School of Medicine. Just learning about ANOVA (analysis of variance) and t-tests isn’t going to cut it anymore, he says. Within computation, students are well-trained in programming and algorithms but lack the engineering skills needed to build robust, reproducible, and usable tools, he says. A subset of students also need more exposure to “real biology”—the kind of wet-lab experiences that “give you an idea of how many things you can screw up.”

In this era of large-scale collaborative research, programs also need to better emphasize general skills—such as project management, creative problem solving, and communication, says **Li-San Wang, PhD**, associate professor of pathology and laboratory medicine at the University of Pennsylvania. These skills are learned by working on research projects, says Wang, who chairs the interdisciplinary PhD program in genomics and computational biology. “These are things you can’t even teach in classes.”

Altman stresses the importance of core

Skill Category	Specific Skills
General	time management, project management, management of multiple projects, independence, curiosity, self-motivation, ability to synthesize information, ability to complete projects, leadership, critical thinking, dedication, ability to communicate scientific concepts, analytical reasoning, scientific creativity, collaborative ability
Computational	programming, software engineering, system administration, algorithm design and analysis, machine learning, data mining, database design and management, scripting languages, ability to use scientific and statistical analysis software packages, open source software repositories, distributed and high-performance computing, networking, web authoring tools, web-based user interface implementation technologies, version control tools
Biology	molecular biology, genomics, genetics, cell biology, biochemistry, evolutionary theory, regulatory genomics, systems biology, next generation sequencing, proteomics/mass spectrometry, specialized knowledge in one or more domains
Statistics and Mathematics	application of statistics in the contexts of molecular biology and genomics, mastery of relevant statistical and mathematical modeling methods (including experimental design, descriptive and inferential statistics, probability theory, differential equations and parameter estimation, graph theory, epidemiological data analysis, analysis of next generation sequencing data using R and Bioconductor)
Bioinformatics	analysis of biological data; working in a production environment managing scientific data; modeling and warehousing of biological data; using and building ontologies; retrieving and manipulating data from public repositories; ability to manage, interpret, and analyze large data sets; broad knowledge of bioinformatics analysis methodologies; familiarity with functional genetic and genomic data; expertise in common bioinformatics software packages, tools, and algorithms

Summary of the skill sets of a bioinformatician, identified by surveying bioinformatics core facility directors and examining bioinformatics career opportunities. Reprinted from L Welch, F Lewitter, R Schwartz, C Brooksbank, P Radivojac, B Gaeta, MV Schneider, Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies, PLoS Comp Biol, 10.1371/journal.pcbi.1003496 (2014).

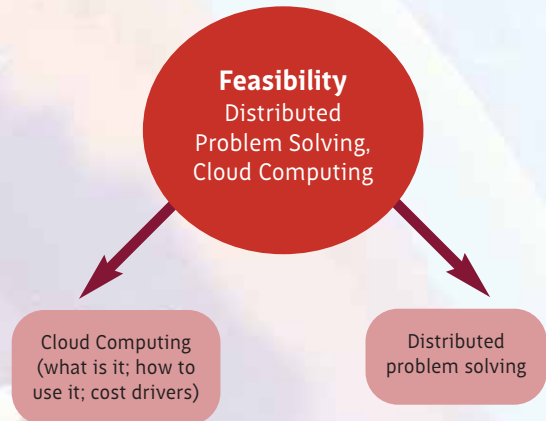
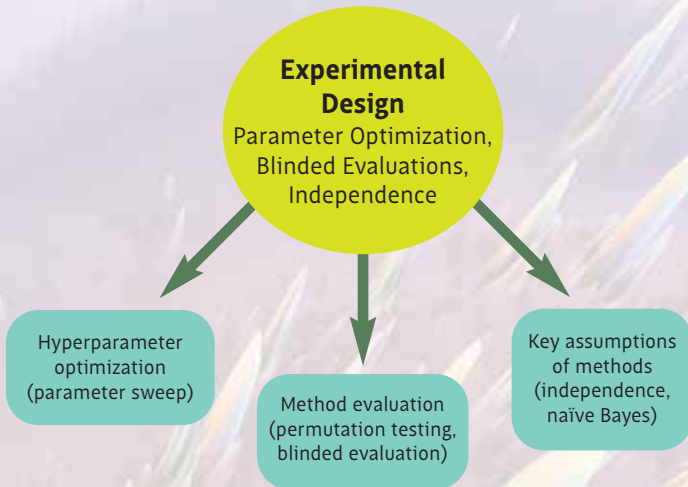
The team published curriculum recommendations in *PLoS Computational Biology* in 2014, including core competencies in five categories similar to Altman’s: general, computational, biology, statistics and math, and core bioinformatics (see the table above for specifics).

From this list, our interviewees highlighted several areas in which training pro-

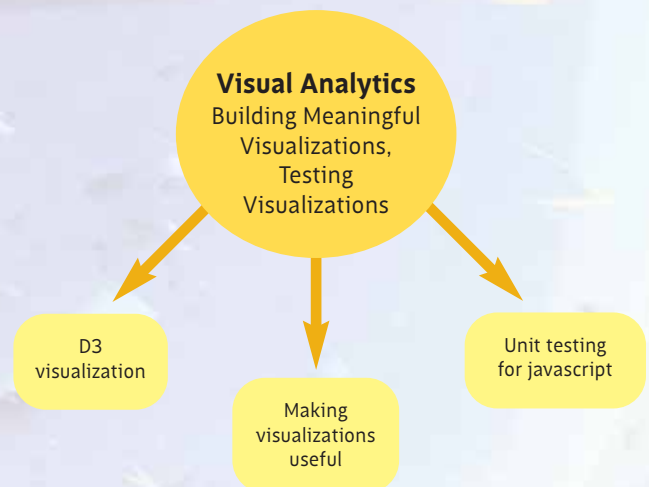
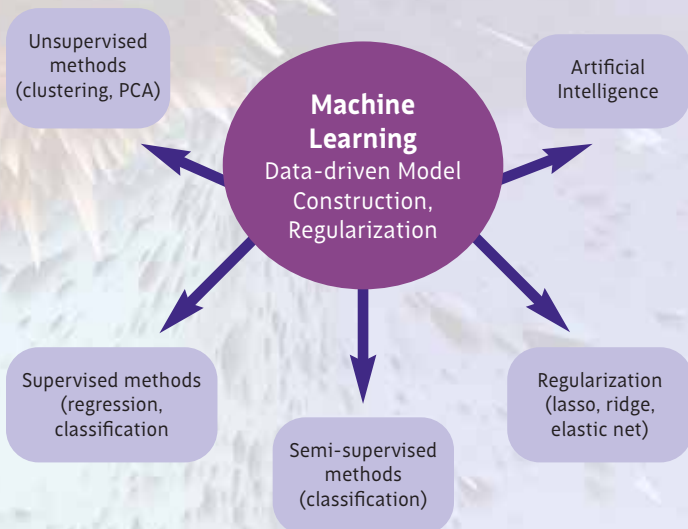
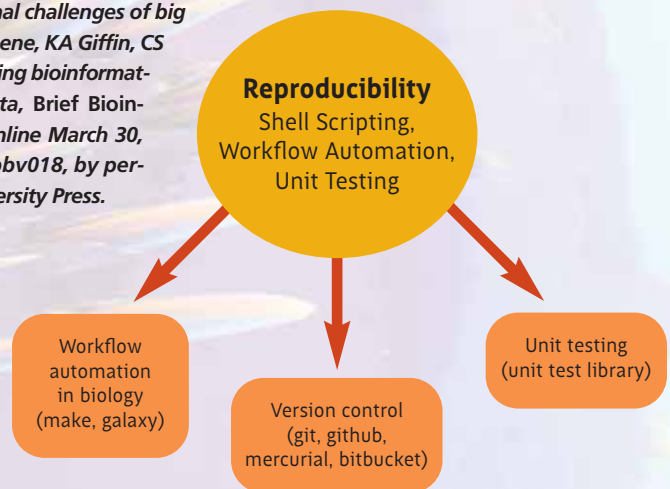
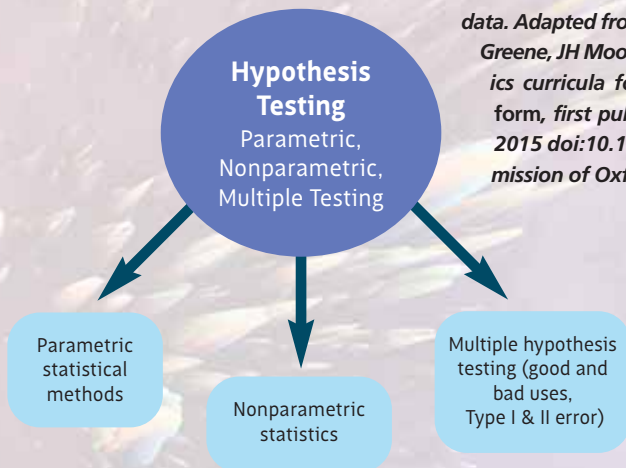
gram bioinformatics training. “Just because you know biology and computer science doesn’t mean you know biomedical or biological informatics,” he says. Students need more capstone courses “where statistics and biology come together or where computer science and biology come together.” This requires an investment in new faculty who are trained as bioinformaticians. “A lot of

STATISTICAL CHALLENGES OF BIG DATA

COMPUTATIONAL CHALLENGES OF BIG DATA



In a 2015 paper about adapting bioinformatics curricula for big data, Greene and his colleagues proposed these courses on statistical and computational challenges of big data. Adapted from AC Greene, KA Giffin, CS Greene, JH Moore, Adapting bioinformatics curricula for big data, Brief Bioinform, first published online March 30, 2015 doi:10.1093/bib/bbv018, by permission of Oxford University Press.



institutions have hired our bioinformatics graduates,” he says.

Though the wishlist of skills seems long, just having a defined list can help focus a curriculum. By emphasizing fundamental skills—rather than specific tools, problems, or data types—the guidelines also help narrow the learning space. Stressing fundamentals also “suits people up for the long haul,” Altman says. “The world is going to be changing a lot in the next 30 years. Whatever you’re looking at on the horizon, it’s just a good idea to go back to fundamentals.”

ADAPTING TO BIG DATA

The onslaught of big data is also putting additional demands on bioinformatics curricula—in particular in the realm of

students have not kept pace with the times, Greene says. “Some classes are geared toward molecular biologists, who may have one thing that they want to analyze. Or they need to know how to do an ANOVA. Don’t get me wrong: These are important skills. But the idea that these skills are going to scale is not really right.” Statistics courses need to include material on machine learning, multiple hypothesis testing, and dealing with bias and confounding in the data, he says. He and others published recommendations for adapting bioinformatics curricula for big data in a 2015 paper in *Briefings in Bioinformatics*.

Dealing with big data also requires additional computational skills to ensure robust and efficient data storage, management, and

Most statistics courses aren’t teaching this approach yet, but it’s coming, he says.

TAILORING THE CURRICULA

Some heterogeneity in training is warranted. Curricula need to be tailored to the degree (BS, MS, PhD, or certificate) and the trainees’ end goals. Welch’s task force is making these different needs explicit. In particular, they have pointed out that bioinformatics practitioners fall into three distinct groups: Bioinformatics *users* are bench biologists or physicians who use bioinformatics tools in research or patient care; bioinformatics *scientists* develop algorithms and pipelines to answer specific biomedical questions; bioinformatics *engineers* support science by building robust software and computational infrastructure.

...Bioinformatics practitioners fall into three distinct groups: Bioinformatics users are bench biologists or physicians who use bioinformatics tools in research or patient care; bioinformatics scientists develop algorithms and pipelines to answer specific biomedical questions; bioinformatics engineers support science by building robust software and computational infrastructure.

statistics. Training programs have to keep abreast of these developments or they risk producing graduates who aren’t prepared for the job market.

Once viewed as a pairing of biology and computation, bioinformatics is increasingly recognized as a three-way pursuit: biology, computation, and statistics. This shift is reflected in departmental changes at many universities. In 2015, both Dartmouth and Stanford merged divisions of biostatistics and bioinformatics into new departments of biomedical data science. In recent years, Harvard’s biostatistics department has also adopted a bioinformatics focus—with a new MS program in computational biology and quantitative genetics in full swing and an MS program in health data science in the works.

What does statistics add to the mix? Whereas computer scientists focus on finding patterns in the data, statisticians worry about sorting out real patterns from spurious ones. “Statistics provides unique expertise in making inference by accounting for errors,” says **Xihong Lin, PhD**, professor of biostatistics at Harvard. “This is especially important when one deals with massive data, as more data means more noise and a higher chance for more mistakes.”

But statistics courses for bioinformatics

analysis. Students need to know about high-performance computing and parallel computing, for example. “Let’s face it, however,” says **John Quackenbush, PhD**, professor of computational biology and bioinformatics in the department of biostatistics at Harvard, “the amount of data we’re dealing with in biomedicine is nothing compared to what the folks in Silicon Valley are amassing at Google or eBay or Facebook.” So, the computational challenges in biomedicine are not trivial, but they’re not as pressing as the statistical challenges, he says.

The era of big data will also require entirely new ways of thinking about data, says **Sean Eddy, PhD**, professor of molecular and cellular biology and of applied mathematics at Harvard. “We have to learn to interact with these massive datasets in an experimental fashion,” he says. For example, you can simulate data that come from the null hypothesis as a negative control—if the statistical tools you’re applying find a positive signal, then you know the approach is faulty. When approaching data from this empirical view, biologists actually have an advantage. “We’re trained to deal with big black boxes where we can’t see all the moving parts. We know how to do experiments to ask questions out of complicated systems,” Eddy says.

Some graduate programs are aiming to train engineers, others scientists, and still others (often at the master’s level) are more focused on training users.

“One of the reasons that there is so much friction and tension in bioinformatics education is that we’ve tried to put bioinformatics into a single box,” Welch says. “Just calling out these three categories helps us to move the conversation forward.”

Each group needs varying levels of depth and breadth across the different competencies. You can view each skill as lying on a continuum—and the depth that you need in each depends on which of the three groups you fall into, Welch says. For example, users and scientists are going to be further along the continuum of life sciences knowledge than engineers, whereas engineers will have more depth in software engineering and system administration. “So what we’re working on now is: How do you specify the points along the different axes where a person roughly should be if they want to be a certain type of bioinformatician at a certain level of career and degree.”

BRIDGING THE DIVIDE

One of the most vexing issues in bioinformatics education is the heterogeneity of

the students. Cross-trained students do exist, but they're in short supply—and they tend to get siphoned off by the most elite programs. Even when admission requires firm grounding in both computation and biology, “There’s almost nothing you can assume in common among the whole incoming student body, even at the PhD level,” says **Russell Schwartz, PhD**, professor of biological sciences and computational biology at Carnegie Mellon University and

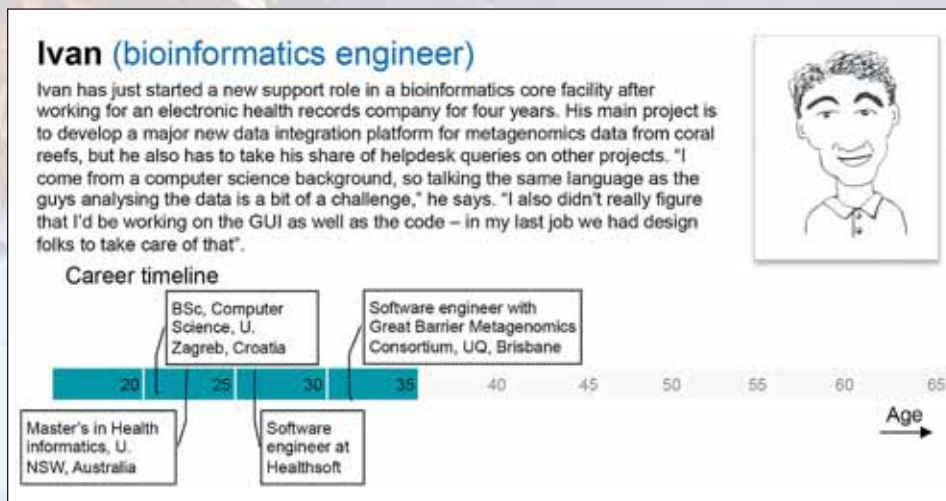
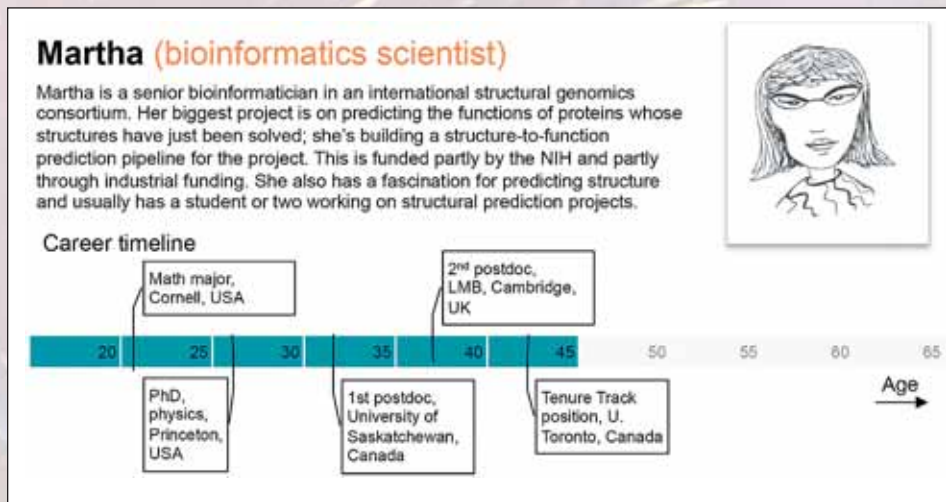
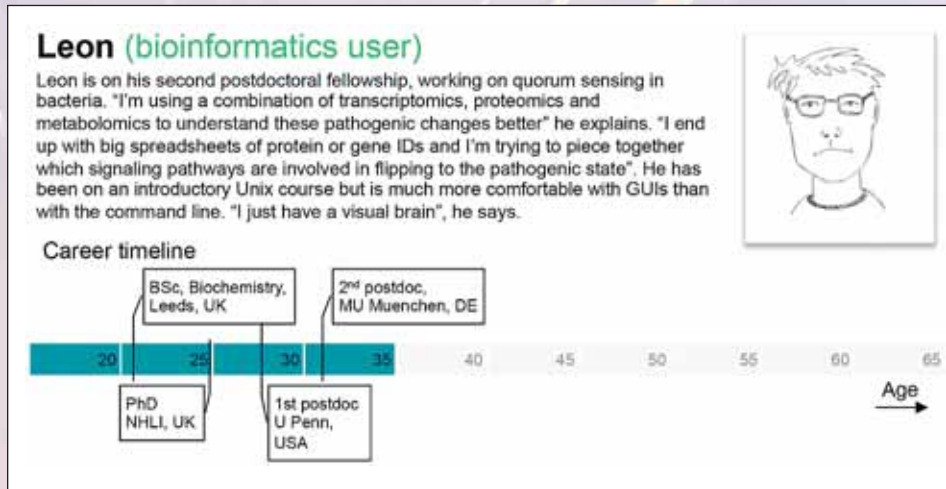
codirector of their PhD program in computational biology, which is offered jointly with the University of Pittsburgh.

For most programs, students tend to come in with strength in one area—either biology or computation—but not both. To train up the other side, many programs offer short courses or boot camps such as “Programming for Scientists” or “Crash Course in Biology for Engineers.” The idea is to break down the barriers, Eddy says. For a biologist, a key bar-

rier is writing a Perl or Python script. “So you need to hold their hand, give them example scripts, and convince them that this is actually not as hard as they might think. It’s not computer engineering; it’s just like pipetting—just do it,” he says. Once students get past their initial fears, they can go learn more on their own. Students are increasingly able to acquire new skills by taking massive open online courses (MOOCs), which offer user-friendly introductions to computer programming, statistics, and biology. (See “Skills Upgrades” on page 7 of this issue.)

According to Greene, students with biology backgrounds are often portrayed as mathematically challenged and thus harder to bring up to speed. But, he says, “We have a lot of smart people that come from molecular biology and a lot of smart people that come from computer science.” In fact, he believes that mastering the biology is the tougher job because of the field’s nuances. “You can teach someone a superficial amount of biology quickly. But it’s hard to give them enough training so that they will have an intuitive grasp of why the data that they are analyzing look weird,” he says. For example, Greene recounts how a student analyzing data noticed that one day’s worth of numbers looked strange. It took a field trip to the lab to figure out the explanation: A new individual had begun washing the glassware on that day and had likely left residual soap. “It’s amazing what a trip to the lab can reveal that data analysis won’t.”

Bridging the culture gap can be harder than bridging the skill gap. Computer scientists and biologists have different ways of thinking, and they speak different languages. This is where joint advising and joint research ventures can help. At the University of Pennsylvania, doctoral students in genomics and computational biology often have dual advisors—one from biology and one from a quantitative discipline. “This works really well because after students finish their coursework, they still



According to a *PLoS Computational Biology* paper published in 2014 by an ISCB task force, *bioinformatics curricula should be tailored for three different types of practitioners—users, scientists and engineers. To illustrate these varieties of bioinformatician, the paper created a graphic description of three fictitious characters—Leon, Martha and Ivan—which has been abbreviated here. Adapted and reprinted from L Welch, F Lewitter, R Schwartz, C Brooksbank, P Radivojac, B Gaeta, MV Schneider, Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies, PLoS Comp Biol, 10.1371/journal.pcbi.1003496 (2014).*

get exposure and advice from both sides,” Wang says. In his courses at Ohio University, Welch uses team projects to foster interaction. “When it works, you see the computer scientists helping the biologists to program and the biologists explaining the molecular biology to the computer scientists—it’s just so fun to watch.”

STARTING EARLIER

To reduce heterogeneity at the graduate level, students need better undergraduate training. “Just like someone going into a physics graduate program could be assumed to have a certain amount of undergraduate training in physics and not have to start from zero, I’d like to see us get to the same point with bioinformatics and computational biology,” Schwartz says. We need more undergraduate programs in bioinformatics/computational biology; and more training in statistics and computer programming for all science majors, he says.

The University of Nebraska at Omaha was one of the first institutions to offer a major in bioinformatics—starting in 2004. At first, they just pieced the degree together from pre-existing courses in chemistry, computer science, and biology, says **Mark Pauley, PhD**, senior research fellow at the College of Information Science & Technology and one of the developers of the major. However, over time the program has developed a series of discrete courses in bioinformatics. The program prides itself on being comprehensive, Pauley says. The major requires 24 credits in bioinformatics, 24 in computer science, 16 in biology, 17 in chemistry, and 11 in math.

Stanford has offered an undergraduate major in biomedical computation since 2003. “We used to have a lot of students trying to craft their own programs. So a bunch of faculty got together and designed an independent major,” Altman says. The major comprises four math courses (including statistics), four computer science courses, three chemistry courses, three biology courses, two engineering courses, a physics course, and a “technology and society” course.

It’s tricky to pack so much into a four-year degree and hard to find faculty to build these programs—particularly at small liberal-arts colleges. “So there still aren’t a whole lot of majors in existence, though it’s growing,” Pauley says. Other universities have compromised by creating certificate programs or minors. For example, Washington University in St. Louis offers a minor in bioinformatics that comprises three biology courses (including a lab), two computer science courses, a statistics course, and a bioinformatics algorithms course. “It’s kind of

past together, and we don’t have as many discrete courses in the field as we would like, but it helps prepare students for bioinformatics graduate work,” says **Sarah Elgin, PhD**, professor of biology.

An even deeper problem is that undergraduate science majors in general aren’t receiving adequate statistics and programming training. “I don’t think it’s tenable for people to be entering a sciences graduate program or graduating from a sciences undergraduate program and not have a solid grounding in statistics or not know how to write a basic computer program,” Schwartz says.

In particular, undergraduate programs in biology have been slow to add quantitative requirements—despite repeated calls for curricular updates. “It’s the usual story that many people who go into biology do so because they love science but are scared of equations,” says **Cath Brooksbank, PhD**, head of the training program at the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL–EBI) in the United Kingdom. “Biology is not portrayed as a quantitative science. But it is a quantitative science, and it’s becoming increasingly quantitative.” At Washington University, Elgin says they have added a statistics requirement to the biology major, but programming is still not required.

TEACHING THE BIOLOGISTS

Though biology education has been slow to adapt, the field of biology is rapidly changing. Before long, every biologist will have to be a bioinformatics user. So, educa-

gram?” he says.

Undergraduate biology majors are already jam-packed, so it’s not always feasible to add an entire bioinformatics class. Pauley and others, including Elgin and Schwartz, are designing and curating bioinformatics modules that can easily be inserted into existing biology classes. Some are available on CourseSource (<http://www.coursesource.org/>), such as a module in which students are genotyped by 23andMe and then explore their own SNPs (single nucleotide polymorphisms).

Elgin is also integrating real genomic research into the biology classroom. She directs the Genomics Education Partnership (GEP, <http://gеп.wustl.edu/>), which brings together students from over 100 colleges and universities in a “massively parallel undergraduate” effort. “You teach everyone the same methods, but each person is responsible for their own part of the action,” Elgin explains. Elgin parses megabases of raw fruit fly sequence data into small stretches that individual students correct and annotate. The resulting wealth of high quality, carefully annotated sequence data can be used to answer biological questions. “There are huge amounts of data that nobody’s ever looked at. So there are lots of opportunities for undergraduates to get in there and get involved.”

Students participate in the research all the way through publication—including reading, critiquing, and approving the final manuscript. In fact, GEP published a paper in *G3: Genes Genomes Genetics* in 2015 that listed 940 students as co-authors. Having so

Before long, every biologist will have to be a bioinformatics user. So, educators are trying to embed bioinformatics, statistics, and computer science into biology education at all levels.

tors are trying to embed bioinformatics, statistics, and computer science into biology education at all levels.

Pauley is the principal investigator of the Network for Integrating Bioinformatics into Life Sciences Education (NIBLSE), an NSF-funded project aimed at determining how much and how best to integrate bioinformatics into biology (<http://niblse.unomaha.edu>). “One of the big questions for us is what bioinformatics do biologists need to know? For example, do they need to be able to pro-

many student authors caused a stir, but Elgin believes that each student made a significant intellectual contribution that should be recognized.

The students also came away with a deeper knowledge of biology and bioinformatics, says **Anne Rosenwald, PhD**, associate professor of biology at Georgetown University, whose students participated. “Students have heard since high school that there are introns and alternative splicing, but until they have to puzzle piece together

what a gene looks like, they don't understand gene structure very well," she says. In formal assessments, GEP students improved their scores on a genomics/genetics quiz, and reported gains in understanding the nature of scientific investigation on par with students who spent a summer in a research lab, according to a 2014 paper in *CBE Life Science Education*, by Elgin, Rosenwald, and others. Even more importantly, Elgin says, the students gained awareness of the vast amounts of data available and the importance of computers in extracting new knowledge from that data. "Hopefully we are also waking up those biology students, inspiring them to sign up for more math and computer science courses," she adds.

Another way to add bioinformatics content is to link an existing biology course with an existing computer course. Such "in-concert teaching" is described in a 2014 paper in *PLoS Computational Biology* by **Anya Goodman, PhD**, associate professor of chemistry and biochemistry, and **Alexander Dekhtyar, PhD**, professor

of computer science at California Polytechnic State University, San Luis Obispo. Students attend separate lectures but collaborate on joint labs and projects. The computer science students write the programs, and the biology students specify the programming requirements and test the software—so the two groups learn how to work in a cross-disciplinary team.

One of the major stumbling blocks to bringing bioinformatics into biology is the lack of biology faculty trained in this area. "I was a full professor before I had a personal computer," Elgin notes. So, Rosenwald has created a project, GenomeSolver (<http://genomesolver.org>), aimed at training biology faculty to use basic bioinformatics tools. "If the faculty don't know this, then the students don't get the exposure to this important way of thinking about biology," Rosenwald says.

POOLING RESOURCES

With so much to cover and so many audiences to serve, bioinformatics educators

are getting together to pool resources.

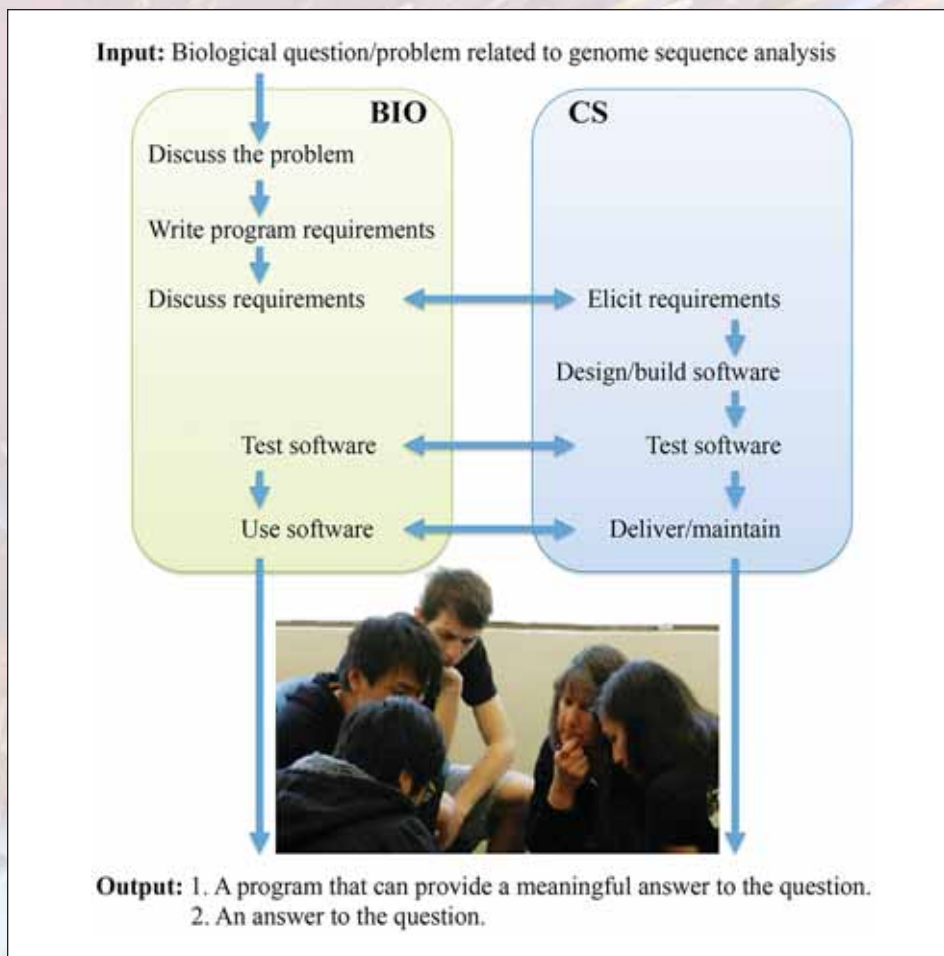
In 2012, educators founded the Global Organization for Bioinformatics, Learning, Education, and Training (GOBLET, <http://www.mygoblet.org>). The goal is to connect trainers across the globe so they can share expertise and training materials. The group is working to establish global curriculum standards and accreditations; and they have created a training portal where educators can deposit and find high quality lectures, exercises, and datasets for teaching.

Other organizations are at work building repositories of publicly available biomedical data. Though originally meant to facilitate bioinformatics research, these organizations are also playing a significant role in education. For example, EMBL–EBI maintains a comprehensive range of freely available molecular databases and the tools to share, analyze, and query data. "When I first joined the EMBL–EBI, the vast majority of users were bioinformaticians," Brooksbank says. "But our user base has grown and diversified hugely since then. Our training program needs to cater to this diversity." So EMBL–EBI offers online training as well as workshops for graduate students, postdocs, faculty members, and industry professionals.

MOOCs are also bringing bioinformatics education to a wide audience. **Pavel Pevzner, PhD**, a professor of computer science at the University of San Diego, offers six short courses in bioinformatics on the MOOC platform Coursera. These problem-driven courses can be taken with or without a programming component, opening them up to biology and bioinformatics students alike.

In a 2014 paper in *PLoS Computational Biology*, **David Searls, PhD**, an independent consultant, argues that "a sufficient number and variety of free video courses have made their way to the web that it is possible to obtain a reasonably comprehensive bioinformatics education on one's laptop." He has assembled a catalog of relevant online courses organized into virtual departments, such as math, computer science, and biology, and proposed comprehensive curricula for different groups (such as bioinformatics users versus engineers).

The availability of so many training resources takes some pressure off formal university programs. Programs don't have to teach every student everything. Rather, they need to give students a firm grounding in the fundamentals plus the tools for lifelong learning. "What we're trying to do by the end of the graduate program is to have people who are pluripotent—who can go many directions from there," says Dunn. □



By linking an existing biology course with an existing computer science course, some universities are engaging in "in-concert teaching." As diagrammed here, the biology students define the problem and discuss the program requirements. They then work with computer science students who build the software. Reprinted from **AL Goodman, A Dekhtyar, Teaching Bioinformatics in Concert**. *PLoS Comput Biol* 10(11): e1003896. doi:10.1371/journal.pcbi.1003896 (2014).

By Alexander Gelfand

Career Paths:

A Seller's Market for Biomedical Data Science Jobs

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, biomedical informatics, 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 exam-

ple, 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.

Developing the Future's Mathematical Tools: An Academic's Career Path



Daniela Witten, PhD

Associate Professor of Statistics and Biostatistics,
University of Washington

"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."

academia

The Road (Almost) Not Taken



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."

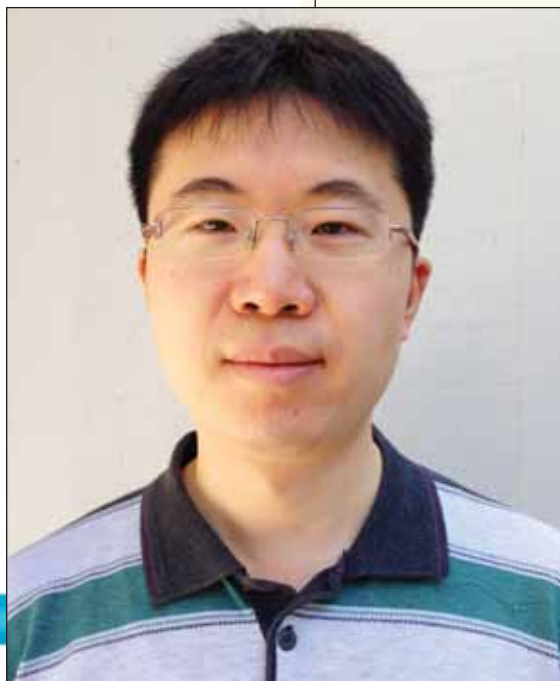
Not much has changed since then. As a doctoral candidate in Stanford's BMI program, Tatonetti developed novel statistical and computational methods that allowed him to mine the Food and Drug Administration's voluminous records on adverse drug reactions, identifying pairs of medications that caused problems 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."

academia

Developing Drug Therapies in an Industry Setting: The Appeal of Clear Goals



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).

industry

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 Kadcyra 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.”

Keeping Your Options Open: From Industry to Academia and Back Again



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 Therapeu-

tics, 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."

Keep It Exciting: Add a Dash of Startup Energy

Grace Zheng, PhD Application Scientist, 10X Genomics

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.

BY NORMAN BIER

Learning Engineering: Leveraging Science and Technology for Effective Instruction



There is currently unprecedented interest in the potential of technology to transform learning. This buzz around technology and learning is especially loud in higher education, where pundits, entrepreneurs

ous, putting them into practice requires time and effort as well as dedication to the goal of using technology to be a more effective teacher.

The accelerating pace of advances in bioinformatics demands new, more effective approaches to training and education for students and experienced practitioners alike.

and academics offer outspoken predictions that technology-enhanced learning (TEL) will productively disrupt the sector by addressing long-standing structural issues and the dual challenges of cost and attainment.

The massive online open course (MOOC) approach, which uses technology to scale lecture and teaching efforts, has been a particular focus of attention in many fields, including biomedical data science. But while this video-focused approach has been successful in expanding access to educational resources, its impact on learning is less clear. The accelerating pace of advances in bioinformatics demands new, more effective approaches to training and education for students and experienced practitioners alike. How can technology be used to meet this need?

One approach, called learning engineering, entails the use of learning research and the affordances of technology to design and deliver innovative, instrumented educational practices with demonstrated and measurable outcomes. This approach has proven quite successful in a variety of contexts over the last few decades, resulting in accelerated learning, higher outcome achievement, improved retention, and higher-order skills attainment, all across diverse and often vulnerable learner populations.* The data from these innovations are also used to refine and advance theory, fueling a virtuous cycle of research and practice.

This article highlights seven core practices that characterize how the learning engineering approach can be applied to develop technology-enhanced learning tools and courses in bioinformatics, biomedical data science and related fields. Although some of these ideas may seem obvi-

1) Use evidence-based design: You are a scientist. So let learning science inform the design of your learning environment. Despite evidence that research-based instruction supports robust learning more effectively than instruction guided by intuition, many faculty continue to design online courses using their personal sense of what works. Don't fall into that trap. Spend time with the learning engineering literature (the Global Learning Center whitepaper referenced below makes a good start) and get advice from learning engineering experts. Guidance and support in evidence-based instruction can often be found in your own institution's center for teaching excellence, but resources are also available from programs such as the Simon Initiative at Carnegie Mellon; the iAMSTEM program at the University of California; and the Kirwan Center for Academic Innovation at the University System of Maryland. A new report from MIT's Online Education Policy Initiative also provides evidence and advocacy for the learning engineering approach.

2) Begin with a model: As you design, develop an explicit model of the learning that you are supporting. What are the objectives and related skills that learners need to achieve? A useful model will describe these elements in a measurable way, and will map them to the activities and assessments that you develop. In a statistics course, for example, a simple model might first explicate the learning objective "Relate measures of center and spread to the shape of the distribution, and choose the appropriate measures in different contexts," with sub-skills that could include "Compute median" and "Identify outlier." These skills would then be mapped to learning activities and assessments, creating a model that can support your design and data-driven analysis. Courselets.org is an example of applying this kind of modeling in the bioinformatics arena (See Skills Upgrades story, page 7).

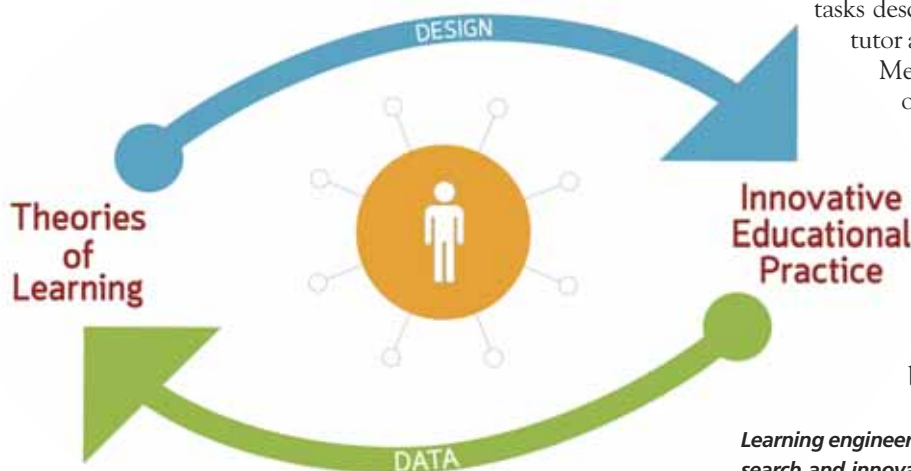
DETAILS

Norman Bier is director of the Open Learning Initiative (OLI) and the executive director of the Simon Initiative at Carnegie Mellon University.

3) Focus on learn-by-doing activities: Design active learning experiences, embed practice opportunities in the problem-solving context, and provide learners with targeted feedback that speaks to misconceptions. While this learn-by-doing approach can be expensive, the potential pay-off for your learners is tremendous. A recent study found relatively simple learn-by-doing activities required no additional time investment by the learner but were six times more effective in supporting learning than the readings and video lectures that are central to many MOOCs.* In the bioinformatics and biomedical data science fields,

those misconceptions. Also, students should be given multiple opportunities to solve open-ended exercises in order to generate data about students' errors, thereby highlighting conceptual problems and demonstrating ways the course should be improved. These activities allow students to engage meaningfully with authentic problems in ways that provide observable information about the learner's knowledge state in relation to specific learning objectives.

6) Choose high quality platforms and tools: Avoid reinventing the wheel. Locate and use existing materials that were designed using learning science research; and choose tools and platforms with care to ensure that they provide features that allow you to accomplish the tasks described above. For example, the cognitive tutor authoring toolkit (CTAT) out of Carnegie Mellon can help educators develop valuable online educational tools. In addition, extensive open educational resources (OERs) are available in many domains that are effective in their own right and can serve as a strong foundation for continued improvement and experimentation. Indeed, several are currently being created for the bioinformatics arena, including Courselets.org and Bio-bigdata.ucsd.edu, which are both de-



some MOOC designers are finding creative ways to incorporate learn-by-doing activities into online learning. (See the work by Pavel Pevzner, Brian Caffo and Rafael Irizarry described in the Skills Upgrades story, page 7).

4) Continuously research, iterate and improve instruction: As with most fields of study, the science of learning changes over time. By revising courses to address developments not previously studied, educational materials grow increasingly effective and robust over time. And again, it's scientific: Learning engineers treat the development of instructional activities as a hypothesis on how learners will best achieve a given learning outcome; capture data from learners' interactions to evaluate the hypothesis; and then take the essential step of closing the loop by using this new information to refine the theory and improve the learning activities. The Carnegie Mellon DataLab offers some tools (links below) to help teachers cycle through this process.

5) Capture rich learning data: Rich learning data fuels effective feedback to learners and educators, drives iterative improvement, and supports advances in learning research. Most online courses only track which pages students visit in which order (so-called click-stream data) rather than meaningful interactions with the material. Activities should be designed to produce useful data. For example, multiple-choice problems should include incorrect answers designed to highlight learner misconceptions (rather than serving as mere distracters), and offer an immediate means to correct

Learning engineering produces a data-driven virtuous cycle of learning research and innovative educational practice, causing demonstrably better learning outcomes for students from any background.

scribed in the Skills Upgrades story in this issue, as well as OERs under development through Oregon Health and Science University with funding from the National Institutes of Health Big Data to Knowledge program.

7) Bring together multi-disciplinary teams: Online instruction is too often developed by a single educator working alone. The educator might have disciplinary expertise and classroom experience, but producing effective TEL resources requires many other talents including expertise in design, instruction, technology, learning science and human-computer interaction. □

* The online version of this story, posted at <http://bcr.org>, includes detailed references.

RESOURCES:

Simon Initiative (cmu.edu/simon) as well as its DataLab (<http://www.cmu.edu/datalab/>) and cognitive tutor authoring tools (<http://ctat.pact.cs.cmu.edu/>)

LearnLab (www.learnlab.org) – the Pittsburgh Science of Learning Center – offers tools and research.

The Eberly Center for Teaching Excellence and Educational Innovation (www.cmu.edu/teaching) provides additional materials to support evidence-based design.

The Open Learning Initiative (oli.cmu.edu) offers online learning environments that exemplify the learning engineering approach.

A recent white paper by the Global Learning Council titled *Technology-Enhanced Learning: Best Practices and Data Sharing in Higher Education* (2015) is available from the Council's web site (www.globallearningcouncil.org).

Stanford University
318 Campus Drive
Clark Center Room S271
Stanford, CA 94305-5444

seeing science

SeeingScience

BY KATHARINE MILLER

Animation & Inspiration

The Australian government is betting that animations can help promote science and science literacy. As part of VIZBI (Visualizing Biological Data) Plus, they funded three biomedical animators, including **Chris Hammang**, Garvan Medical Research Institute, Sydney, Australia, to each create two educational animations. The resulting films (posted at (<http://vizbi.org/plus/>)) tell great stories, are visually compelling, and also get the details right.

That's the tough part, says Hammang. He spent six months

preparing the four-minute animation called *Alzheimer's Enigma* while he worked at CSIRO (Commonwealth Scientific and Industrial Research Organization). He read scientific literature and talked to researchers in order to hone in on a consensus view of how plaques form in Alzheimer's disease patients' brains. He also hunted down the molecules' component parts in the Aquaria database and then assembled them using ePMV, embedded Python molecular viewer. Finally, he used the 3-D animation software called

Blender to create the film.

In the resulting animation, molecules wriggle like little critters, and enzymes snip other proteins with a light crunching sound—qualities that these still images cannot convey. (Hint: Watch the video!)

- The animation's lead character is APP (amyloid precursor protein, shown in yellow/orange), which resides in the membrane of brain cells (fig. 1).

- To periodically recycle and replace the membrane and its embedded proteins, leggy proteins called clathrins (in blue) assemble into a lattice on the inside of the membrane (fig. 2), chunking off a vesicle with APP proteins on the inside (fig. 3).

- During recycling, several enzymes clip APP into three pieces, including a small stub (orange) that, in Alzheimer's disease, somehow escapes recycling and accumulates outside the cell. In high concentrations, these bits can glom together to form long fibers, which clump together in masses called plaques (fig. 4).

