BY KATHARINE MILLER

WOMEN IN DATA SCIENCE CONFERENCE

What happens when hundreds of talented female data scientists gather in the same place?

In November 2015, the Mobilize Center co-hosted the first Women in Data Science (WiDS) Conference along with Walmart Labs, Stanford University's Institute for Computational & Mathematical Engineering (ICME), and several other Stanford entities, including the department of statistics, the engineering department's computer forum, and the Office of the President.

More than 400 people attended the one-day conference, which was aimed at inspiring, educating, and supporting women in data science—from those just starting out to those who are established leaders across industry, academia, government, and nongovernmental organizations.

And they were inspired.

"When you are surrounded by successful and talented women in a room full of support, encouragement and inspiration, your dreams and goals burst into a reality that pushes you to conquer it all," tweeted attendee Diana Riveros Mello during the conference.

Margot Gerritsen, PhD, director of ICME and a Mobilize Center faculty member, organized the WiDS conference because she recognizes the tremendous talent among women. "I see this every day when I am teaching," she says. "And it would be a real shame if that group of wonderful scientists is underutilized." Gerritsen wants more women to join the field. "It's important for society as a whole…to have a very diverse, inclusive team of people working on data science problems."

Data science involves extracting relevant information from voluminous, heterogeneous, and often messy data streams, and using that information to help inform decisions across all arenas, including research, government and business. "It's everywhere now," Gerritsen says.

The impressive roster of all-female WiDS conference speakers exemplified the field's breadth. About one-third of the speakers came from academia and two-thirds from industry, and the conference covered a diverse set of data science applications, from monitoring individuals with Parkinson's disease, to cancer genomics, cyber security and online marketplaces.

"It's important for society as a whole… to have a very diverse, inclusive team of people working on data science problems," Gerritsen says.

"Just seeing the array of possibilities makes me think, 'Yeah, I can do great things too,'" says **Shenglan Qiao**, a Stanford PhD candidate in physics who attended the conference.

In addition, panels on careers and entrepreneurship offered an opportunity for successful data scientists to reflect on their lives and offer advice to younger

continued on page 6

DETAILS

Video recordings of the November 2015 *Women in Data Science Conference* are available online under the 2015 Conference menu at widsconference.org. The next WiDS conference will be held on February 3, 2017.

continued from page 2: WiDS Conference

women. Most often, they encouraged taking risks and being flexible.

"Don't let your fear about your own abilities or fear of being an imposter have any bearing on the kinds of decisions you make," said **Jennifer Chayes, PhD**, distinguished scientist and managing director of Microsoft Research New England in Cambridge, Massachusetts. "Take that part of your brain and say thank you for sharing and just put it aside. If I'd listened to that part of my brain, I would have led a very boring life."

Interest in the conference was high: It sold out in less than 20 days with little promotion, and more than 6,000 individuals tuned in to the live-stream.

The vast majority of attendees hope to attend the next conference, which is scheduled for February 2017.

Gerritsen advises women who are interested in computational math or other scientific fields: "Jump in. It's a fabulous field with lots of opportunity." □

continued from page 3: Accelerating Search

to cover all the data.

During the fine search, the best-matching cluster and all neighboring clusters have to be searched. Fractal dimension describes the number of neighbors each cluster has. Lower fractal dimension means fewer neighbors to search. Fortunately, biological data have a fairly low fractal dimension because evolution tends to trace out relatively linear paths (see figure). "Clusters largely tend to extend along the branches of the tree rather than in all directions," Berger explains.

Berger's team showed that their compression and search framework is effective for any data that exhibit low metric entropy and low fractal dimension. Thus, potential applications extend way beyond sequence search. In their *Cell Systems* paper, Berger's team demonstrates orders of magnitude speed-ups for searching databases of chemical compounds, metagenomes, and protein structures.

PubChem is a comprehensive database of 60 million small molecules that can be used for tasks such as repositioning drugs. Until now, it was infeasible to perform even a one-molecule search of all of PubChem on a typical desktop computer. So Berger's team clustered the chemical compounds in PubChem based on the geometric similarity of chemical motifs, and then applied their two-step search process to these data. Compared with the commonly used search tool SMSD (Small Molecule Subgraph Detector), they were able to achieve a 150-fold speed up with 92 percent accuracy.

The team's framework can be wrapped around common search tools, such as SMSD for small molecule search, BLAST for DNA sequence search, and PSI-BLAST for protein sequence search. "The cool thing about all our tools is that they plug right into existing pipelines."

Berger's team has made their tools openly available at: http://cast.csail.mit.edu/, and other groups have begun building upon these tools, she says. "This whole area of compressive algorithms is really taking off because it's absolutely necessary." □

continued from page 5: SimTK Relaunched

visitors. This feature is already encouraging visitors to explore the other projects on SimTK: In the first year after it was implemented, total monthly project visits more than doubled (from 31,000 to 63,000) and 42 percent of project visits were made through the recommendation system. Ku hopes such features will also motivate visitors to host their own projects on the site.

Reproducibility in the Cloud

One major challenge of physics-based simulation is reproducibility—ensuring that researchers using the same data and software can get the same results.

To address this problem, Ku and Erdemir are working on offering members a cloud-based way to reproduce published results on SimTK. "We're hoping to further lower the barrier to entry for modeling and simulation," Erdemir says.

The feature enables users to launch a simulation by simply selecting a server, a model, and a specific software version from dropdown menus. When the results are available either for download or for browsing online, the user receives a notification. As a test case, Erdemir has created a template for running such simulations of OpenKnee. The interface allows users to run a simulation, perhaps apply a different load to the knee, and run a new simulation. Ku and Erdemir hope the cloud-based option will be up and running before the end of the year.

Plug and Play Capability

In the coming year, SimTK will also include the ability to plug-and-play with other online applications. For example, several SimTK projects use GitHub as a way to collaborate on their source code. They might also use another site to track bugs and then they use SimTK to share the software. But SimTK could be the hub that provides ways to pipe information to and from these multiple places, Ku says. Some of the developers of the site's largest projects on SimTK, such as OpenSim, OpenMM and SimVascular, are eager for this improvement. "Users and contributors alike will have one place to go to get quick updates; communicate; know where the project is headed," Ku says.

SimTK: Past, Present and Future

SimTK was novel in 2005 when it started out, Ku says, but 10 years on, "technology and our users' needs have changed, so SimTK is changing, too," she says. Ku hopes the features now being added to SimTK will put it back at the forefront—well ahead of whatever other new new things might come along—and keep it relevant for the community of researchers that flock to its pages. □

SNEAK PEEK

For a sneak peek of the new SimTK site, visit https://simtkalpha.stanford.edu.