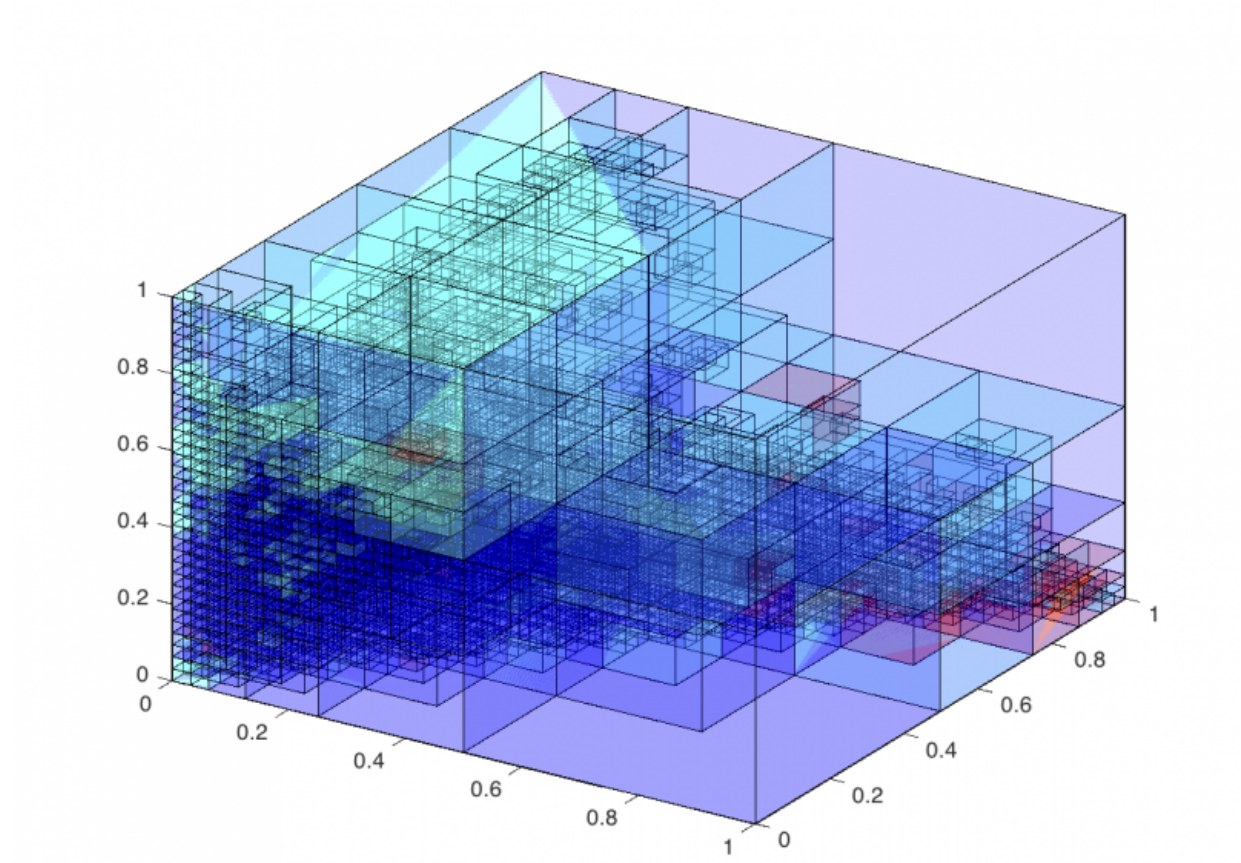
Mathematical and computational methods to analyze de-centralized information

November 17 2015, by Marlene Cimons



Graphical example of a probabilistic model used to classify stars from the Kepler Dataset. Different colors represent different types of stars projected into a 3-dimensional feature space. Credit: Trilce Estrada, Dept. of Computer Science, University of New Mexico

Scientific advances typically produce massive amounts of data, which is, of course, a good thing. But when many of these datasets are at multiple locations, instead of all in one place, it becomes difficult and costly for researchers to extract meaningful information from them.

So the question becomes: "How do we learn from these datasets if they cannot be shared or placed in a central location?" says Trilce Estrada-Piedra.

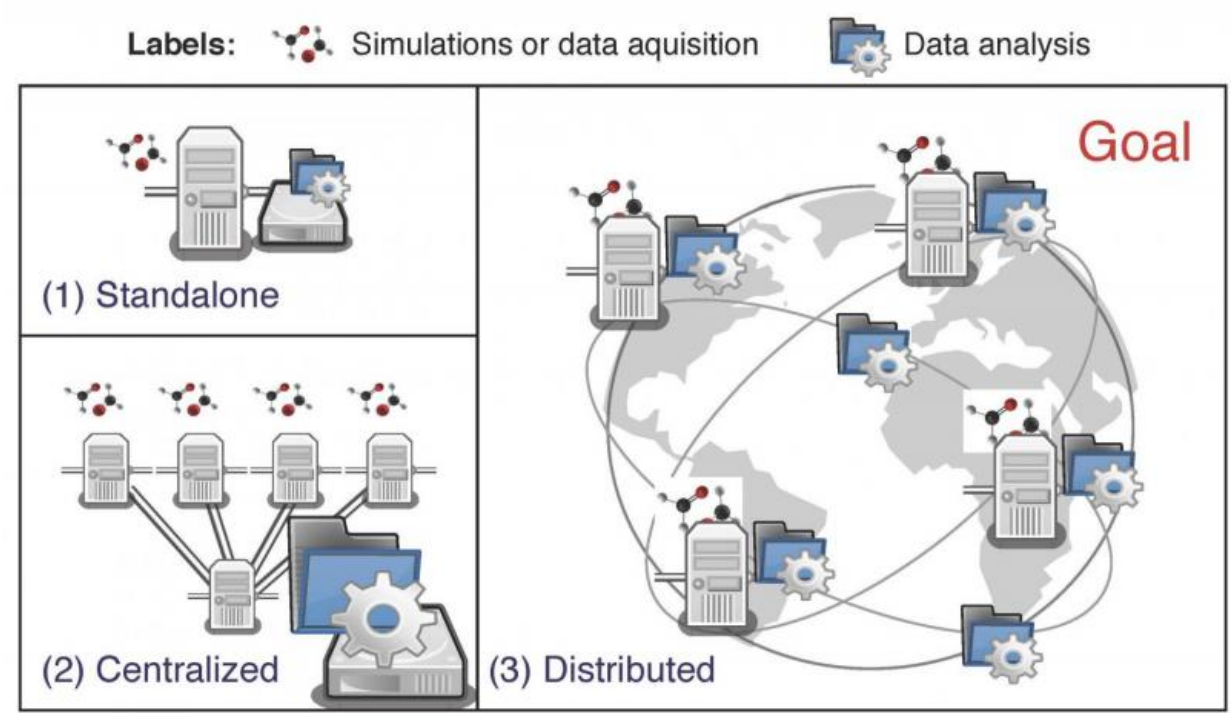
Estrada-Piedra, assistant professor of computer sciences at the University of New Mexico (UNM) is working to find the solution. She designs software that will enable researchers to collaborate with one another, using decentralized data, without jeopardizing privacy or raising infrastructure concerns.

"Our contributions will help speed research in a variety of sciences like health informatics, astronomy, high energy physics, climate simulations and drug design," Estrada-Piedra says. "It will be relevant for problems where data is spread out in many different locations."

The aim of the National Science Foundation (NSF)–funded scientist's project is to build mathematical models from each of the "local" data banks—those at each distributed site. These models will capture data patterns, rather than specific data points.

"Researchers then can share only the models, instead of sharing the actual data," she says, citing a medical database as an example. "The original data, for example, would have the patient's name, age, gender, and particular metrics like blood pressure, heart rate, etc. and that one patient would be a data point. But the models will project his or her information and extract knowledge from the data. It would just be math. The idea is to build these local models that don't have personal information, and then share the models without compromising privacy."

Estrada-Piedra is designing algorithms for data projections and middleware: software that acts as a bridge between an operating system or database and applications, especially on a network. This will allow distributed data to be analyzed effectively.



The goal of this project is to enable distributed learning in domains where the standalone or centralized processing of results are infeasible. Distributed learning, allowing scientists to learn from datasets that cannot be shared or placed in a central location, is important for scientific discovery. Credit: Trilce Estrada, Dept. of Computer Science, University of New Mexico

Finding an effective way to learn from distributed data is essential for scientific discovery, she says.

"This is a way of enabling science, meaning that researchers will more

easily be able to analyze larger datasets, especially those that, for some reason, cannot be centralized. Just the moving of the data may be prohibitive. So, if you want to get results faster, this will be a way of doing it."

Once complete, these products will be made available through a GitHub repository.

In May 2015, Estrada-Piedra was part of a group—which included researchers from the University of Delaware, the San Diego Supercomputing Center and Argonne National Laboratory—that took first place in an international computing challenge for their project, "Accurate Scoring of Drug Conformations at the Extreme Scale."

"Right now we are testing the way in which the data projections and predictions work, and will be moving into the `distributed' aspect next year," she says.

Estrada-Piedra is conducting her research under an NSF Faculty Early Career Development (CAREER) award, which she received earlier this year. CAREERs supports junior faculty who exemplify the role of teacher-scholars through outstanding research, excellent education and the integration of education and research. She is receiving $412,969 over five years.

As part of the award's education component, Estrada-Piedra is using the concept of crowdsourcing for distributed problems, which will allow teachers and researchers to work on a distributed problem, similar to the game FoldIt, where "students see a molecule and they have to try to fold it into a protein," she says.

It will use the same middleware she's using for her research, named Andromeda. This application will revolve around basic chemistry and

physics concepts, providing an interactive platform for science teachers in high schools to convey concepts more effectively.

"We will upload a code so that students can use it as a game to produce science, and to enable gaming capabilities for students to play with distributed predictions," she says.

Estrada-Piedra will also use Andromeda as an integral component of her UNM course in Big Data, teaching undergraduate and graduate students techniques for distributed data analysis.

Estrada-Piedra co-founded UNM's Data Sciences Lab, a collaborative effort to impact curriculum, recruit students, and increase the visibility of data science. The lab will enable the early recruitment of undergraduate students into the data sciences, and will use summer research experiences as a pipeline to encourage undergraduates to attend graduate school.

More information: For more information, see gcl.cis.udel.edu/publications/ … CGRID_SCALE_talk.pdf

Provided by National Science Foundation