Special Audio Transcript

Headline:	University of California Spearheading New Technologies To Mine Big Data in Health Care
Reported/Produced by:	Deirdre Kennedy
Publication:	*iHealthBeat*
Date Posted:	May 9, 2012

The University of California is spearheading new computer technologies to help researchers utilize "Big Data" in medicine and biotechnology. This is a special report for *iHealthBeat*, a daily news service of the California HealthCare Foundation. I'm Deirdre Kennedy.

As part of its "Big Data" initiative announced in March, the National Science Foundation awarded $10 million to UC-Berkeley's new AMPlab. The name stands for algorithms, machines and people. The lab will create an open-source software stack to collect, organize and make sense of vast amounts of data. That information has just been made available by a number of federal agencies, including the Department of Defense, CDC and NIH.

Berkeley computer science professor Dave Patterson is one of eight faculty members and 40 Ph.D. students working on the AMPlab project. He says the problem with Big Data is that “it’s unstructured, uncurated, and can't be neatly stored in rows and columns.”

(Patterson): "We're used to dealing with lots of data and inventing algorithms that can process this data well. There's a particular area of computer science that we think will be highly appropriate called machine learning. And machine learning is based upon statistics. You've seen some of the benefits of machine learning if you've used Google for example."

One of the lab's projects will be designing a platform to interact with and analyze a huge cancer database from NIH.

(Patterson): "We’re working with The Cancer Genome Atlas, which is five petabytes of medical information about the 20 most common cancers, having 500 patients, both the tumor cells and their normal cells. It's only just been possible to sequence hundreds of cancer tumors. Before, it was astronomically expensive so no one could possibly afford it."

At the NSF announcement, NIH Director Francis Collins said, faster, more affordable computer processors and cloud storage have made the price of gene sequencing drop faster than anyone could have predicted.

(Collins): "The average cost of sequencing an entire genome has fallen from $400 million when the first one was completed in 2003 to less than $8,000 today, and there are two companies that have announced an intent to make it possible to sequence an entire genome in 24 hours for less than $1,000 in the near future."

Researchers hope that scientists can learn more about how tumors grow and create effective personalized treatments for patients.

The first hurdle is aggregating data from many different sources and cross-referencing them says Taylor Sittler -- a pathologist at UC-San Francisco and a member of the AMPlab team.

(Sittler): "Just putting together those 4,000 specimens and 4,000 tumor normal pairs has proved to be a major effort. It's collected by different institutions in slightly different ways. Each has [its] own system with slightly different fields. That all requires manual work to integrate. [And we're seeing this across not just this database but many others.]"

They might have inconsistencies in date and time format or how medical terms are abbreviated.

Sittler would like to see electronic health records incorporated into the datasets to better track the patient's prognosis and outcome.
For now, EHRs are not available without patient consent.

But Atul Butte, chief of the Division of Systems Medicine in the Department of Pediatrics at Stanford School of Medicine, says that EHRs are the next frontier of Big Data.

(Butte): "Everyone's talking about improving health care quality with it to try to prevent harm perhaps in real time, try to prevent mistakes. But yet no one is in a position to put that out into the public yet, right? The closer it gets to an actual human patient there is a lot of anxiety."

Butte says it's possible to mine EHRs without violating patient privacy. Stanford did just that. The university released a widely publicized study earlier this year that found that women reported higher pain levels than men. Researchers gleaned that information from thousands of patient records by just combing one piece of data -- how they rated their pain when nurses asked them.

(Butte): "It was the largest study ever for pain. That data was just sitting there in the repository waiting for someone to do something with it. It didn't even have to be publicly available. Data is on its way to getting more and more public. I actually think the new challenge is what do we want to ask of that data?"

This has been a special report for *iHealthBeat*, a daily news service from the California HealthCare Foundation. If you have feedback or other issues you'd like to have addressed, please email us at IHB@chcf.org. I'm Deirdre Kennedy. Thanks for listening.