Big grant for Big Data:
NSF awards $10 million to harness vast quantities of data

By Sarah Yang, Media Relations | March 29, 2012

BERKELEY — The quest to capture the massive amounts of data being produced in our world – and in so doing unveil answers to some of society's most vexing problems – has gotten a $10 million boost from a National Science Foundation award to the University of California, Berkeley.

The grant, part of NSF's "Expeditions in Computing" program, will be announced today (Thursday, March 29) at a White House-sponsored event unveiling the Obama Administration's "Big Data Research and Development Initiative."

At the event, NSF Director Subra Suresh and leaders from the National Institutes of Health, Department of Energy, Department of Defense, the Defense Advanced Research Projects Agency and the U.S. Geological Survey will discuss their agencies' respective efforts to tackle the fast-growing volumes of data in our world. Altogether, more than $200 million will be committed by the agencies.

There has been an explosive growth of data over the past decade.

The DOE-funded project is close to home. Researchers at the Lawrence Berkeley National Laboratory will lead the Scalable Data Management, Analysis, and Visualization Institute, a collaboration that brings in partners from seven universities and five other national laboratories.

The five-year NSF Expedition award to UC Berkeley – an award rarely given to individual universities – will fund the campus's new Algorithms, Machines and People (AMP) Expedition. AMP Expedition scientists expect to develop powerful new tools to help extract key information from Big Data, a term coined for the dizzying array of measurements, images, audio, video, tweets, texts and more that has grown ever larger, faster and more diverse.

"Buried within this flood of information are the keys to solving huge societal problems and answering the big questions of science," said Michael Franklin, director of the AMP Expedition team and the first holder of the new Thomas M. Siebel Chair in Computer Science in UC Berkeley's Department of Electrical Engineering and Computer Sciences. "Our goal is to develop a new generation of data analysis tools that provide a quantum leap in our ability to make sense of the world around us."

Never before has data analysis been so hot.

The Big Data challenge has attracted attention worldwide, and even the United Nations has joined the game. Its Global Pulse initiative aims to harness mounds of digital data to improve human well-being.

Where UC Berkeley stands out, said Franklin, is with its early adoption of a holistic, three-pronged approach to tackling the Big Data challenge. The AMP Expedition researchers will develop an open-source software stack called the Berkeley Data Analysis System (BDAS) that integrates the following:

- Algorithms – New, large-scale machine-learning and data analysis methods
- Machines – Systems infrastructure that allows programmers to exploit the power of cloud and cluster computing
- People – Crowdsourcing that uses human intelligence where computers fall short

"We're not going to address this challenge simply by making machines do more of what they've been doing," said Franklin. "And there are some tasks that are still performed much better by people. It's important to develop a system that can incorporate these very different types of resources and flexibly blend them on a problem-by-problem basis. We believe that will be the breakthrough."

Secrets to societal problems may lie hidden in the massive amounts of data before us.

Behind this explosive growth in data are increasingly sophisticated sensors and devices that can provide quantitative measurements at never-before-seen resolutions, as well as the rise of social media networks that can reveal real-time trends in group mood and behavior.

By teasing out useful details from the noise and clutter, patterns should emerge that could signal the early stages of an epidemic, help create detailed models for urban growth, thwart cyberattacks and detect Earth-like planets. Studies in recent weeks have shown that messages from Twitter and online discussions can be used to predict stock performance and unemployment spikes. And the power of crowdsourcing was demonstrated last fall when gamers discerned the protein structure of an AIDS-like virus through a science puzzle site called Foldit.

"Advances in technology have made it possible to capture and store increasingly massive amounts of data. The challenge now is to design better systems for organizing, extracting, and analyzing the data," said Hal Varian, chief economist at Google and a UC Berkeley professor emeritus of information. "The UC Berkeley team has both the breadth and the depth necessary to take on this challenge."

The AMP Expedition is an expansion of UC Berkeley's AMPLab, which was formed in early 2011 by UC Berkeley researchers with a wide range of data-related expertise. The founding group included eight faculty members and dozens of Ph.D. students who were backed by 18 industry titans, including Google and SAP.

In addition to Franklin, the AMP Expedition is led by co-principal investigators Michael Jordan, Scott Shenker and Ion Stoica, all professors in the Department of Electrical Engineering and Computer Sciences, and Alexandre Bayen, associate professor in the Department of Civil and Environmental Engineering. Jordan and Stoica are also co-directors of the AMPLab.

"The NSF grant recognizes UC Berkeley's leadership in the field of Big Data," said David Culler, chair of UC Berkeley's Computer Science Division. "Throughout the campus, deep data analytics is at the heart of science and engineering, from political science to astronomy to energy efficiency and climate."

The AMP expedition will work with a number of Big Data projects already underway at UC Berkeley, including Mobile Millennium, led by Bayen, that collects data from thousands of GPS-enabled mobile devices to provide real-time reports of traffic conditions on arterials as well as highways. UrbanSim, led by Paul Waddell, professor and chair of the Department of City & Regional Planning, is another Big Data project. It incorporates data from metropolitan planning organizations and other regional agencies to develop models that can assess impacts on urban development from land use regulations and transportation policies.

AMP researchers have also begun collaborations with colleagues from UC San Francisco and UC Santa Cruz to help develop applications for deciphering cancer genomics.