COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

GO

Data Science Institute

ABOUT CENTERS ACADEMICS RESEARCH ENTREPRENEURSHIP INDUSTRY

New Working Group Takes On Massive Computing Needs of Big Data

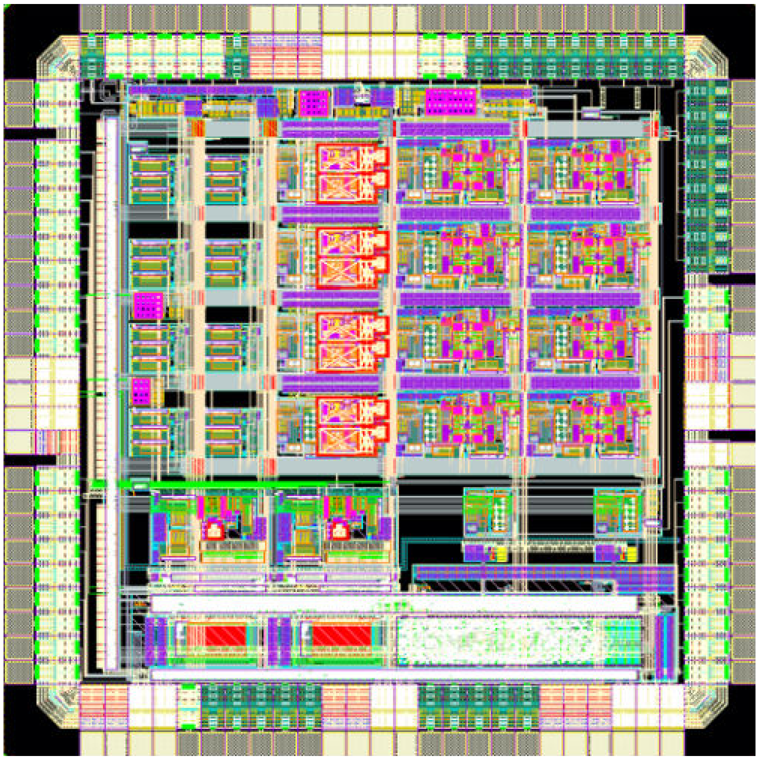
SHARE (http://www.addthis.com/bookmark.php?v=250&pub=xa-4a9be9465d42784c)

In the big data era, the modern computer is showing signs of age. The sheer number of observations now streaming from land, sea, air and space has outpaced the ability of most computers to process it. As the United States races to develop an “exascale” machine up to the task, a group of engineers and scientists at Columbia have teamed up to pursue solutions of their own.

The Data Science Institute’s newest working group —Frontiers in Computing Systems (http://datascience.columbia.edu/frontiers-in-computing-systems) —will try to address some of the bottlenecks facing scientists working with massive data sets at Columbia and beyond. From astronomy and neuroscience, to civil engineering and genomics, major obstacles stand in the way of processing, analyzing and storing all this data.

“We don’t have two years to process the data,” said Ryan Abernathey (https://rabernat.github.io/) , a physical oceanographer at Columbia’s Lamont-Doherty Earth Observatory. “We’d like to do it in two minutes.”

Launched by computer science professor Steven Nowick (http://datascience.columbia.edu/steven-m-nowick) , the initiative will combine researchers designing and analyzing extreme-scale computing systems for Big Data, and those working with massive data sets to solve ambitious problems in the physical sciences, medicine and engineering.



Engineers are experimenting with chip design to boost computer performance. In the above layout of a chip developed at Columbia, analog and digital circuits are combined in a novel architecture to solve differential equations with extreme speed and energy efficiency. (Simha Sethumadhavan, Mingoo Seok and Yannis Tsividis/Columbia Engineering)

Scientists on the application side include Mark Cane (http://www.ldeo.columbia.edu/user/mcane) , a climate scientist at Lamont-Doherty who helped build the first model to predict an El Niño cycle, and Rafael Yuste (http://www.columbia.edu/cu/biology/faculty/yuste/index.html) , a neuroscientist whose ideas on mapping the brain helped inspire (http://www.sciencemag.org/news/2013/04/white-house-embraces-brain-initiative-questions-linger) President Obama’s BRAIN Initiative. In all, 30 Columbia researchers are involved.

Frontiers in Computing Systems will complement the Institute’s s centers on automated data gathering, Sense, Collect and Move Data (http://datascience.columbia.edu/sense-collect-and-move-data) , and algorithms and machine learning theory, Foundations of Data Science (http://datascience.columbia.edu/foundations-of-data-science) . Approved by the Institute’s board in June, *Frontiers* received letters of support from corporate and government leaders in high-performance computing and data analytics, including IBM, Intel and NASA’s Jet Propulsion Laboratory.

“This is a timely and interesting initiative, which promises to attack the underlying ‘systems’ aspects of Big Data, which in our view is absolutely essential,” wrote Dharmendra Modha (http://researcher.watson.ibm.com/researcher/view.php?person=us-dmodha) , IBM’s chief scientist for brain-inspired computing who led the development of IBM’s TrueNorth chip.

The initiative comes as the U.S. tries to reclaim its lead in high-performance computing. This year, China emerged (http://www.nytimes.com/2016/06/21/technology/china-tops-list-of-fastest-computers-again.html?_r=0) as the world’s top supercomputing power, with 167 computers on a global list of top 500 machines. The race to develop a next-generation exascale machine, one thousand times faster than today’s leading petascale machines, has led to a surge of U.S. government support.

Last summer, President Obama signed an executive order creating a national strategic computing initiative (https://www.whitehouse.gov/blog/2015/07/29/advancing-us-leadership-high-performance-computing) to coordinate supercomputing research among federal agencies and advance broad societal needs. The U.S. Department of Energy has also created its own exascale initiative (http://www.industry-academia.org/download/20130913-SEAB-DOE-Exascale-Initiative.pdf) .

Modeling natural processes requires massive computer power. At the center of the distant Perseus cluster of galaxies sits a supermassive black hole driving the cyclic heating and cooling of gases. In the above simulation, hot gases emitting X-ray light (left) are juxtaposed against the cooling phase (right). It took a supercomputer 200 hours to produce this simulation. (Greg Bryan and Yuan Li/Department of Astronomy)

Energy and speed remain the key obstacles, whether building centralized supercomputers and data centers, or networks of smaller systems connected by wireless.

To perform at least a billion billion operations per second, these exascale systems need to radically cut energy use. To get there, engineers are rethinking traditional von Neumann architecture, which separates data storage from processing, as well as developing entirely new computing paradigms.

Several promising directions have emerged. Under one, computers mimic the brain with circuits designed to store and process information the way neurons and synapses do. This lets chips run massive numbers of tasks in parallel while conserving energy. Using this approach, IBM's TrueNorth chip requires less than 100 milliwatts to run more than 5 billion transistors compared to modern processors which need more energy to power far fewer transistors.

A second approach is designing customized high performance computer architectures for specialized tasks. From hardware to software, the Anton supercomputers developed by D.E. Shaw Research, a research institute in New York, are designed to simulate (http://www.nature.com/news/2010/101014/full/news.2010.541.html) protein dynamics, important in drug-discovery and understanding basic cell processes.

At Columbia, researchers are exploring both directions and more.

Nowick, the working group's founder, is tackling the daunting communication challenge facing next-generation chips. He and his lab are developing "networks-on-chip," to organize the complex data flows from hundreds of processors and memory stores running at wildly different rates. By eliminating the central organizing clock of traditional chips, their designs allow components to be easily assembled, Lego-like, and upgraded to provide more processing power and memory.

This data visualization maps the statistical relationship between birth month and disease incidence in the electronic records of 1.7 million New York City patients. (Nick Tatonetti/Columbia University Medical Center)

In related work, computer scientist Luca Carloni (http://datascience.columbia.edu/luca-carloni) and his lab are developing tools to design and program "system-on-chip" computing platforms embedded in everything from smartphones to cars to data centers. With hardware customized for specific applications, these single-chip platforms can execute tasks far more efficiently than software running on conventional processors. Other group members are developing brain-computer interfaces, chips for medical devices, and customized computers to solve differential equations for scientific applications.

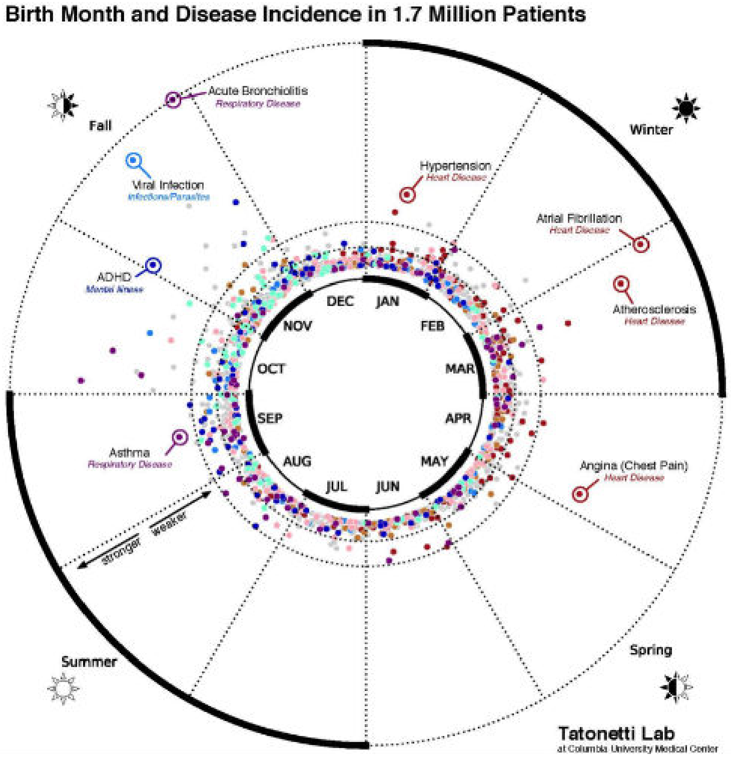
The working group is also focused on adapting software and databases to big data applications. Computer scientist Roxana Geambasu (http://datascience.columbia.edu/roxana-geambasu) is designing tools that allow developers to write and optimize programs for large-scale machine learning problems. Computer scientist Eugene Wu (http://datascience.columbia.edu/eugene-wu) is creating interactive database systems to help users visualize and analyze their results.

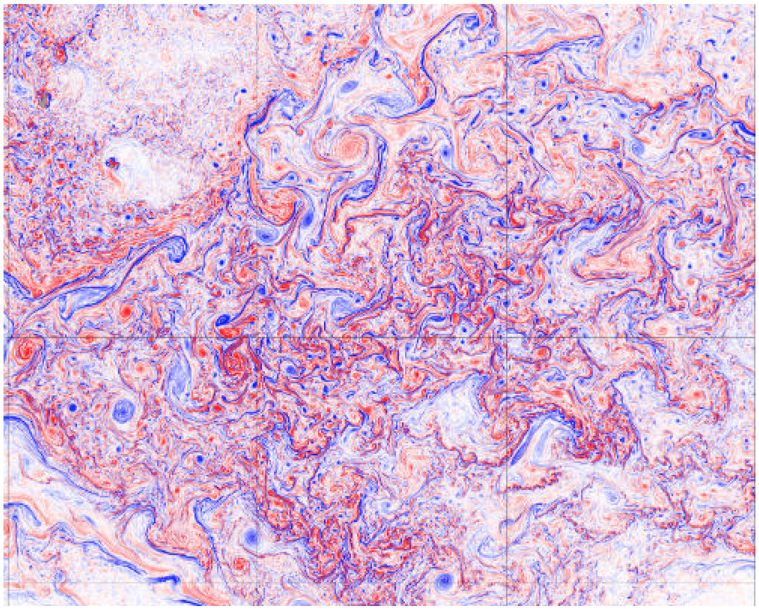
Neuroscientists in the group will also contribute. Their insights into brain structure and function could lead to important advances in computer software and hardware. Conversely, the neuroscientists hope to gain new ideas for modeling the brain from interactions with computer systems researchers.

“When you get engineers and scientists together, exciting ideas emerge,” said Nowick, chairman of the working group. “Many of the breakthroughs in extreme-scale computing are expected to happen as systems designers come to understand the unique needs of researchers grappling with big data sets.”

Those researchers span many fields. In materials science, the problems involve solving equations from quantum mechanics to explain why a material behaves the way it does. Behind these epic math problems are extremely practical applications.

Applied physicist [Chris Marianetti (http://apam.columbia.edu/chris-marianetti)](http://apam.columbia.edu/chris-marianetti), vice chair of the new working group, is trying to understand how lithium atoms pass in and out of a material, one of the secrets to designing a longer-lasting lithium battery. But computational limits have stymied him on this problem and others. “We can burn through as much computing time as we’re given,” he said. “Materials scientists are notorious. You don’t want to be sharing computer time with us.”





Ocean waves interact with large-scale currents in this snapshot taken from a new NASA simulation. Oceanographers eager to explore it are hampered by inadequate computing power. (Dimitris Menemenlis/NASA’s JPL and Chris Hill/MIT)

In medicine, the rise of electronic health records is helping researchers to study disease at unprecedented scale, but here, too, computational challenges remain. [Nicholas Tatonetti (http://datascience.columbia.edu/nicholas-p-tatonetti)](http://datascience.columbia.edu/nicholas-p-tatonetti), a biomedical researcher at [Columbia University Medical Center (http://www.cumc.columbia.edu/)](http://www.cumc.columbia.edu/), has mined data from more than a million patients to uncover dangerous side effects when combinations of drugs are taken together. However, the ability to scale this work depends on further breakthroughs in parallel computing and database design.

For now, Columbia researchers crunch their data on federal supercomputers or the university’s high performance computing system. Such massively parallel systems are optimized for modeling weather and climate, for example, but are not well suited for complex big data problems.

At Lamont-Doherty, Abernathey has explored NASA’s new groundbreaking [simulation (http://maps.actualscience.net/MITgcm_llc_maps/llc_4320/)](http://maps.actualscience.net/MITgcm_llc_maps/llc_4320/) of ocean waves interacting with large-scale currents, run on NASA’s Pleiades supercomputer. He would like to analyze the petabytes of data the simulation has generated, but lacks the processing power to do so.

“These models now output so much data it’s impossible to understand what they’re doing,” he said.

Columbia’s [Lamont-Doherty Earth Observatory (http://www.ldeo.columbia.edu)](http://www.ldeo.columbia.edu) and [Earth Institute (http://www.earthinstitute.columbia.edu/sections/view/9)](http://www.earthinstitute.columbia.edu/sections/view/9), and the [NASA Goddard Institute for Space Studies (http://www.giss.nasa.gov/)](http://www.giss.nasa.gov/) are among the institutions represented in the new working group.

The group aims to attract funding from the U.S. government and industry. In addition to IBM, NASA’s JPL and Intel, the group received letters of support from senior scientists and managers at Sandia National Laboratories, Microsoft Research, NVIDIA and D.E. Shaw Research. If all goes as planned, the working group will become a full-fledged Institute center in the next year, said Nowick.

— Kim Martineau

SHARE [(http://www.addthis.com/bookmark.php?v=250&pub=xa-4a9be9465d42784c)](http://www.addthis.com/bookmark.php?v=250&pub=xa-4a9be9465d42784c)