Big medical data

Massachusetts Institute of Technology

With the recent launch of MIT's Institute for Medical Engineering and Science *[1],* MIT News *examines research with the potential to reshape medicine and health care through new scientific knowledge, novel treatments and products, better management of medical data, and improvements in health-care delivery.*

At the end of 2012, the National Public Radio show "Fresh Air" featured a segment in which its linguistics commentator argued that "big data" should be the word of the year. The term refers not only to the deluge of data produced by the proliferation of Internet-connected, sensor-studded portable devices but also to innovative techniques for analyzing that data; and big data has received a good deal of credit [2] for Barack Obama's victory in the last presidential election.

Certainly, the term was in heavy use around MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), which in 2012 launched a new big-data initiative [3] called bigdata@CSAIL [4]. Several of the researchers affiliated with bigdata@CSAIL are developing new techniques for processing medical data, to make it more accessible to both physicians and patients and to find correlations that could improve diagnosis or choice of therapies.

Peter Szolovits, a professor in the Department of Electrical Engineering and Computer Science (EECS) and the Harvard-MIT Division of Health Sciences and Technology (HST), directs the Clinical Decision Making group at CSAIL, which is researching a whole host of methods for bringing artificial intelligence to bear on medical care. The group participates in a large initiative, sponsored by the National Institutes of Health, to create a database system that would link genomic data and clinical data so that physicians could more easily test hypotheses about connections between genetic variations and particular diseases.

The group is also investigating ways to automatically extract useful data from doctors' free-form clinical notes. Recently, the group presented a promising new approach [5] to the problem of word-sense disambiguation, or inferring from context which of a word's several meanings is intended. (The word "discharge," for instance, shows up frequently in physicians' notes, but with radically different meanings.) The same line of research has spun off several papers on anonymizing medical data — automatically stripping it of identifying information to protect patients' privacy.

Tracking disease

John Guttag, the Dugald C. Jackson Professor in EECS and another member of bigdata@CSAIL, directs CSAIL's Data-Driven Medicine group. Among other things, the group is investigating techniques for detecting and predicting hospital-borne infections. In several papers last year, Jenna Wiens, a graduate student in the group, used machine-learning techniques to comb through dozens of variables [6]

— some static, such as age and complaint upon admission, and some dynamic, such as vital signs and lab results — to find patients that suggested elevated risk of infection with the nasty intestinal bug *Clostridium difficile*.

The one member of bigdata@CSAIL who is not already a CSAIL researcher is Sandy Pentland of the MIT Media Lab. Pentland's group mines data from portable sensors — whether special-purpose devices or cellphones — to find data pertinent to a whole host of questions [7], from how to improve productivity at large companies to the likelihood that two people who just met will start dating. But the same techniques are also useful for epidemiological research. At last year's International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Pentland and his students won the best-paper award for a study tracking the spread of flu through the social networks of a group of MIT students.

Chats, graphs

Also based at the Media Lab is the New Media Medicine Group, headed by Frank Moss, professor of the practice of media arts and sciences. The group's Collective Discovery project, which involves Moss and his graduate students John Moore and Ian Eslick, seeks to provide tools to enable members of online discussion boards — a source of rich but erratic and unstructured information — gather and organize medically relevant data about their own experiences with particular diseases and courses of treatment.

But while a number of research groups at both CSAIL and the Media Lab specifically focus on medical applications, much of the theoretical work at CSAIL on machine-learning and statistical inference will inevitably have medical applications. The Stochastic Systems Group, for instance, which is led by Alan Willsky, the Edwin S. Webster Professor of Electrical Engineering and Computer Science and head of the Laboratory for Information and Decision Systems, concentrates on signal processing, image processing and machine learning, often using mathematical constructs known as graphs [8]. But while the group hasn't explicitly focused on medical applications, "the applicability of graphical to medical analysis has been recognized for some time," Willsky says.

A graph is a data structure that consists of nodes — which are usually depicted as circles — and edges — which are usually depicted as lines. Generally, Willsky says, in his group's work, "the edges between nodes encode statistical relationships." So if the nodes of a graph represented environmental, physiological and genetic factors observed in a population, techniques developed by Willsky's group [9] could, in principle, help researchers evaluate the statistical correlations between those factors and, say, incidence of asthma.

Full circle

A good example of the convergence of computer science and medicine in the age of big data is David Reshef, who has both bachelor's and master's degrees in electrical engineering and computer science from MIT. For his master's thesis, however, Reshef chose as an advisor Pardis Sabeti, an assistant professor of biology at

Harvard and a member of MIT and Harvard's joint Broad Institute. Reshef's plan was to develop algorithms for analyzing epidemiological datasets, to extract information about the conditions that contribute most to disease outbreaks.

That work came to fruition in late 2011, when Reshef and his brother, Yakir — both of whom are now MD-PhD students in HST — were lead authors on a paper [10] in *Science* [11], "Detecting Novel Associations in Large Data Sets." In some ways, that paper brings the crosstalk between computer science and medicine full circle: although born of research on epidemiological data, the algorithms the Reshefs developed — together with Sabeti, Michael Mitzenmacher of Harvard, and other colleagues — are in fact generalizable to all types of data.

Links:
[1] http://web.mit.edu/newsoffice/2012/mit-launching-imes-0202.html
[2] http://www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters/
[3] http://web.mit.edu/newsoffice/2012/big-data-csail-intel-center-0531.html
[4] http://bigdata.csail.mit.edu/
[5] http://web.mit.edu/newsoffice/2012/digital-medical-records-offer-insights-1031.html
[6] https://docs.google.com/viewer?a=v&q=cache:XHswAZEgTi0J:research.microsoft.com/en-us/um/people/horvitz/icml2012_CDiff.pdf+&hl=en&gl=us&pid=bl&srcid=ADGEESiqKS-X9FOJwheqf2Z5h6_eTeNtQB_BmmYWZpiWuUqCERZnFlu1QT0vRy4kVAjmM3ldfPaDAJKouVQZiApHwWqgT_CL8lsc0J_oxCLFBbnIUJ9ABwf1q34Ov0c7q5mWBCNRi628&sig=AHIEtbREGS4uTvQw-SfxDkGA-1Pat4N-tA
[7] http://www.technologyreview.com/article/421386/social-studies/
[8] http://web.mit.edu/newsoffice/2012/explained-graphs-computer-science-1217.html
[9] http://web.mit.edu/newsoffice/2010/sizing-samples-0825.html
[10] http://www.sciencemag.org/content/334/6062/1518
[11] http://www.sciencemag.org/