Drag-and-drop data analytics

June 27 2019, by Rob Matheson

For years, researchers from MIT and Brown University have been developing an interactive system that lets users drag-and-drop and manipulate data on any touchscreen, including smartphones and interactive whiteboards. Now, they've included a tool that instantly and automatically generates machine-learning models to run prediction tasks on that data. Credit: Melanie Gonick

In the *Iron Man* movies, Tony Stark uses a holographic computer to project 3-D data into thin air, manipulate them with his hands, and find fixes to his superhero troubles. In the same vein, researchers from MIT and Brown University have now developed a system for interactive data analytics that runs on touchscreens and lets everyone—not just genius, billionaire, playboy philanthropists—tackle real-world issues.

For years, the researchers have been developing an interactive data-science system called Northstar, which runs in the cloud but has an interface that supports any touchscreen device, including smartphones and large interactive whiteboards. Users feed the system datasets, and manipulate, combine, and extract features on a user-friendly interface, using their fingers or a digital pen, to uncover trends and patterns.

In a paper being presented at the ACM SIGMOD conference, the researchers detail a new component of Northstar, called VDS for "virtual data scientist," that instantly generates machine-learning models to run prediction tasks on their datasets. Doctors, for instance, can use the system to help predict which patients are more likely to have certain diseases, while business owners might want to forecast sales. If using an interactive whiteboard, everyone can also collaborate in real-time.

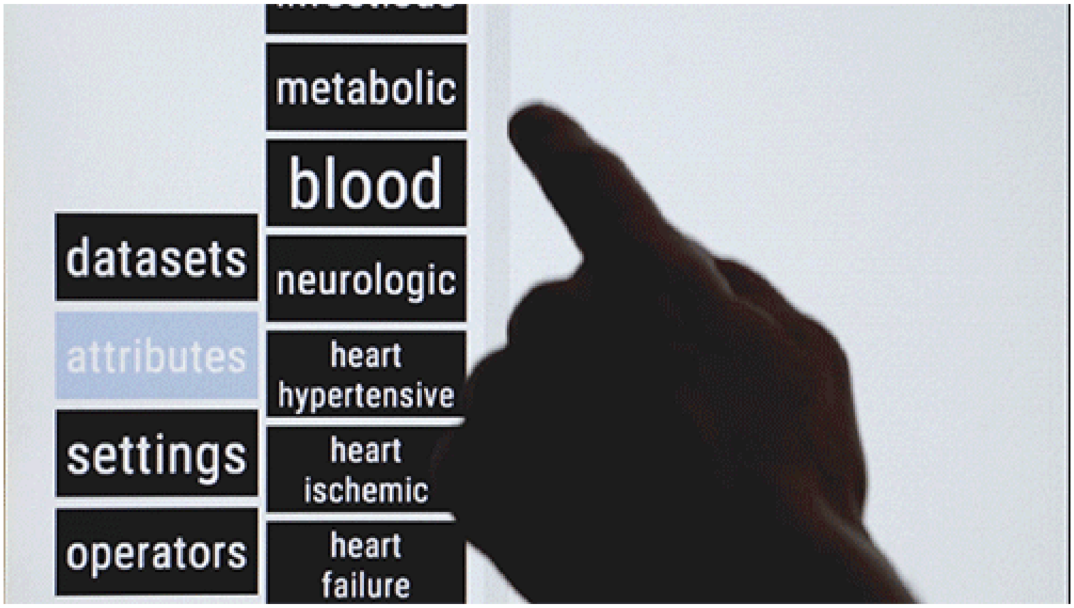
The aim is to democratize data science by making it easy to do complex analytics, quickly and accurately.

"Even a coffee shop owner who doesn't know data science should be able to predict their sales over the next few weeks to figure out how much coffee to buy," says co-author and long-time Northstar project lead Tim Kraska, an associate professor of electrical engineering and computer science in at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and founding co-director of the new Data System and AI Lab (DSAIL). "In companies that have data scientists, there's a lot of back and forth between data scientists and nonexperts, so we can also bring them into one room to do analytics together."

VDS is based on an increasingly popular technique in artificial intelligence called automated machine-learning (AutoML), which lets people with limited data-science know-how train AI models to make predictions based on their datasets. Currently, the tool leads the DARPA

D3M Automatic Machine Learning competition, which every six months decides on the best-performing AutoML tool.

Joining Kraska on the paper are: first author Zeyuan Shang, a graduate student, and Emanuel Zgraggen, a postdoc and main contributor of Northstar, both of EECS, CSAIL, and DSAIL; Benedetto Buratti, Yeounoh Chung, Philipp Eichmann, and Eli Upfal, all of Brown; and Carsten Binnig who recently moved from Brown to the Technical University of Darmstadt in Germany.



Credit: Melanie Gonick

An "unbounded canvas" for analytics

The new work builds on years of collaboration on Northstar between researchers at MIT and Brown. Over four years, the researchers have published numerous papers detailing components of Northstar, including the interactive interface, operations on multiple platforms, accelerating results, and studies on user behavior.

Northstar starts as a blank, white interface. Users upload datasets into the system, which appear in a "datasets" box on the left. Any data labels will automatically populate a separate "attributes" box below. There's also an "operators" box that contains various algorithms, as well as the new AutoML tool. All data are stored and analyzed in the cloud.

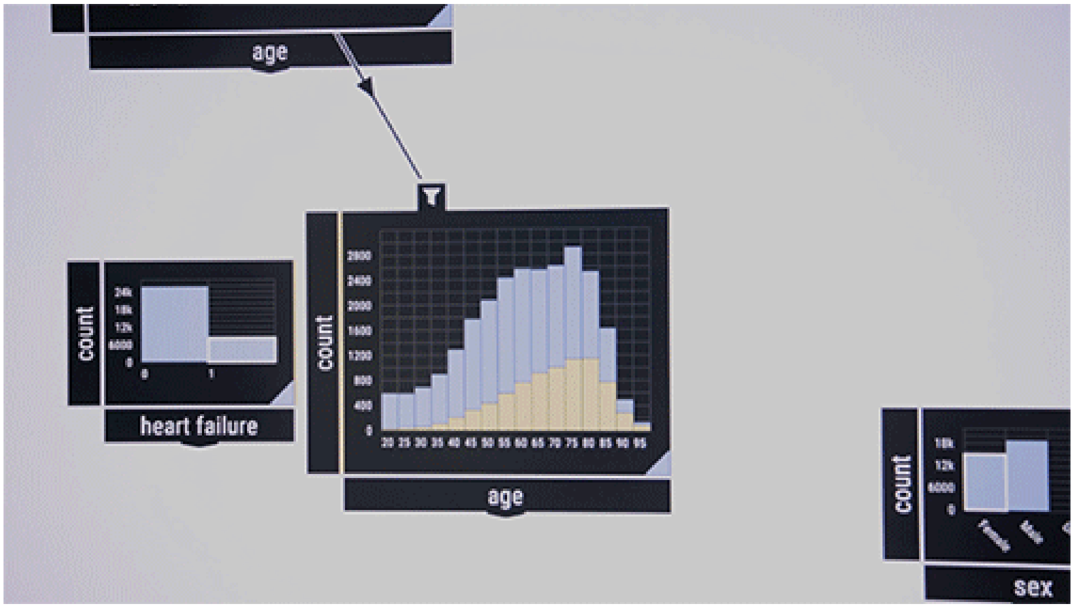
The researchers like to demonstrate the system on a public dataset that contains information on intensive care unit patients. Consider medical researchers who want to examine co-occurrences of certain diseases in certain age groups. They drag and drop into the middle of the interface a pattern-checking algorithm, which at first appears as a blank box. As input, they move into the box disease features labeled, say, "blood," "infectious," and "metabolic." Percentages of those diseases in the dataset appear in the box. Then, they drag the "age" feature into the interface, which displays a bar chart of the patient's age distribution. Drawing a line between the two boxes links them together. By circling age ranges, the algorithm immediately computes the co-occurrence of the three diseases among the age range.

"It's like a big, unbounded canvas where you can lay out how you want everything," says Zgraggen, who is the key inventor of Northstar's interactive interface. "Then, you can link things together to create more complex questions about your data."

Approximating AutoML

With VDS, users can now also run predictive analytics on that data by getting models custom-fit to their tasks, such as data prediction, image classification, or analyzing complex graph structures.

Using the above example, say the medical researchers want to predict which patients may have blood disease based on all features in the dataset. They drag and drop "AutoML" from the list of algorithms. It'll first produce a blank box, but with a "target" tab, under which they'd drop the "blood" feature. The system will automatically find best-performing machine-learning pipelines, presented as tabs with constantly updated accuracy percentages. Users can stop the process at any time, refine the search, and examine each model's errors rates, structure, computations, and other things.



Credit: Melanie Gonick

According to the researchers, VDS is the fastest interactive AutoML tool to date, thanks, in part, to their custom "estimation engine." The engine sits between the interface and the cloud storage. The engine leverages automatically creates several representative samples of a dataset that can be progressively processed to produce high-quality results in seconds.

"Together with my co-authors I spent two years designing VDS to mimic how a data scientist thinks," Shang says, meaning it instantly identifies which models and preprocessing steps it should or shouldn't run on certain tasks, based on various encoded rules. It first chooses from a large list of those possible machine-learning pipelines and runs simulations on the sample set. In doing so, it remembers results and refines its selection. After delivering fast approximated results, the system refines the results in the back end. But the final numbers are usually very close to the first approximation.

"For using a predictor, you don't want to wait four hours to get your first results back. You want to already see what's going on and, if you detect a mistake, you can immediately correct it. That's normally not possible in any other system," Kraska says. The researchers' previous user study, in fact, "show that the moment you delay giving users results, they start to lose engagement with the system."

The researchers evaluated the tool on 300 real-world datasets. Compared to other state-of-the-art AutoML systems, VDS' approximations were as accurate, but were generated within seconds, which is much faster than other tools, which operate in minutes to hours.

Next, the researchers are looking to add a feature that alerts users to potential data bias or errors. For instance, to protect patient privacy, sometimes researchers will label medical datasets with patients aged 0 (if they do not know the age) and 200 (if a patient is over 95 years old). But novices may not recognize such errors, which could completely throw off their analytics.

"If you're a new user, you may get results and think they're great," Kraska says. "But we can warn people that there, in fact, may be some outliers in the dataset that may indicate a problem."

This story is republished courtesy of MIT News *(web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.*

Provided by Massachusetts Institute of Technology

Citation: Drag-and-drop data analytics (2019, June 27) retrieved 3 May 2024 from https://techxplore.com/news/2019-06-drag-and-drop-analytics.html