Traduire Désactiver pour : Anglais

Options ▼



thetp://www.schallengib.gc.data-now-integrate-now-integrat





Home (http://www.datanami.com/) About (https://www.datanami.com/about/)

Whitepapers (http://www.datanami.com/whitepaper/) Events (http://www.datanami.com/events/)

FEATURES

Subscribe (https://www.datanami.com/subscribe/)

HOME (HTTP://WWW.DATANAMI.COM/)

• <u>• </u> Sélectionner une langue | ▼ Search this site Follow Datanami: (http://www.facebook.com/pages /Datanami/124760547631010) (http://www.twitter.com/datanami) in (http://www.linkedin.com/groups/Big-Data-News-Network-4166980) (http://www.datanami.com/feed/) TECHNOLOGIES VENDORS

JOB BANK (HTTP://WWW.DATANAMI.COM/JOB-BANK/)

Leading Solution

### July 5, 2017 What's Challenging in Big Data Now: Integration and Privacy

SECTORS

NTERPRISETECH

JAPAN

(Panchenko Vladimir/Shutterstock)

It's been said many times before, but it's worth stating again: big data presents large opportunities fo improving business and society, but it also involves sizable computing challenges, as well as moral challenges. A panel of renowned professors in the field expounded on the obstacles blocking big data's path forward during a recent meeting of the Association of Computing Machinery (ACM). Privacy and integration issues led the way

APPLICATIONS

In celebration of the 50th anniversary of the A.M.

Turing Award, which is sometimes called the "Nobel Prize of computing," the ACM convened a gathering of some of the brightest minds in computing, including Michael Stonebraker, an adjunct (http://www.datanami.com/) professor at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and the 2014 winner of the Turing Award.

> When asked what challenges big data represents in a super-connected world with 25 billion datagenerating devices, Stonebraker commented that two of the three "Vs" have basically beer

> "If you have a volume problem and you're interested solely in running SQL-style business intelligence on a lot of data, in the data warehouse market, there are at least a few dozen production petascale warehouses that do exactly this day in and day out," said Stonebraker, who created the Vertica MPP database now sold by HPE. "[T]he volume problem is basically solved and shouldn't get much harder in the future.

want to process a million messages a second, current stream processing engines can do this quite easily," he said. "I'm not aware of anybody that wants to go faster than that....I don't consider the velocity issue to be all that difficult."

Similarly, the data velocity problem is under control. "If you

But the third "V" the one that pertains to variety, is a potential deal-breaker, according to Stonebraker, who labeled it the 800-lb. gorilla in the corner of the room.

"As near as I can tell, [data variety] is what is causing problems for nearly every major enterprise on the planet," he said. "I think what is going to kill everybody isn't necessarily the number of connected devices, but the variety of independently-constructed data sources that enterprises are going to want to put together. Whether you're talking about healthcare, manufacturing, or financial services, all of these independently structured databases are going to be a killer.'

David Blei, a professor at Columbia University and a winner of the ACM-Infosys 2013 Foundation Award, said there are great opportunities to benefit from big data, but also some unmet challenges.

"If you take the example of genes and diseases, it's an important computer science and statistics problem that's unsolved," Blei said. "Data scientists are looking to answer how we take data that we observe from the world and use it to identify causal connections between two variables.

Dealing with the uncertainty and biases that can arise from basing conclusions around relations is something that all big data practitioners must tackle. It's also something that Daphne Koller, an adjunct professor of computer science at Stanford University and an ACM-Infosys 2007 Foundation Award, brought up during the panel

Professor David Blei

"Bias will always be a challenge, and there isn't a single, magic solution," Koller said. "The bigger question is, How do we disentangle correlation from causation?...I'll turn to healthcare for an example: the gold standard in the medical state is that of randomized case control. In the case of web data, it's called A/B testing—basically tech industry jargon for the same type of control.

Michael Stonebraker is a Turing

Award winner

Although A/B testing is not perfect as a randomized case control it's "about as good a tool as we've been able to develop for addressing some of the confounders," she continued. "Unfortunately, this type of control is not feasible in all ca

Where A/B testing is not feasible, practitioners must take pains to ensure that "processes [are] carefully scrutinized to check for different confounders and to look for any and all correlations that give rise to the phenomenon being viewed," she continued, "It's a process that requires a lot of thought and a lot of care and cannot

be overstated in its importance.

While there are many challenges with the data science and technical aspects of big data, there are also moral questions, as well as questions of privacy. Vipin Kumar, the regents professor and William Norris chair in Large Scale Computing at University of Minnesota, says these challenges come to a head in healthcare.

## This Just In Most Read July 10, 2017

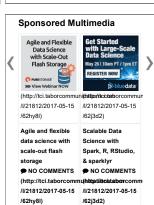
Information Builders Introduces Design Thinking for BI and Analytics (https://www.datanami.com /this-just-in/information-builders-introducesdesign-thinking-bi-analytics/) Pactera Wins Risk Data and Analytics Technology Implementation of the Year Award (https://www.datanami.com/this-just-in/pacterawins-risk-data-analytics-technologyimplementation-year-award/)
Qognify, Bosch Expand Cooperation (https://www.datanami.com/this-just-in/qognify-bosch-expand-cooperation/)

# Sponsored Whitepapers

How to Drive Big Data Projects to Maturity (http://tci.taborcommunications.com/l/21812 /2017-06-22/637lms)

Overcoming Obstacles to Data Lake Success (http://tci.taborcommunications.com/l/21812 /2017-06-12/636acl)

View the Whitepaper Library (https://www.datanami.com/whitepaper/)





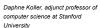
1 sur 2



concerns about possible misuse of these kinds of information, such as discrimination in hiring or in purchasing health insurance, if this information is not handled properly. The healthcare community is on the front lines in this area, but, given the complexity of issues involved, progress in addressing these concerns is very slow."

Big data has the potential to help society in massive ways, but it will require solving thorny privacy questions that go right to the heart of the matter, Stonebraker says

"Privacy is a really good big data question," he says. "Imagine this simple example: you show up at your doctor's office and have an x-ray done and you want the doctor to run a query that shows who else has x-rays that look like yours, what was their diagnosis and what was the morbidity of the patients. That requires



integrating essentially the country's entire online medical databases and presumably would extend to multiple countries as well

Vipin Kumar, Regents Professor and William Norris Chair in Large

Scale Computing at University of

"While that is a daunting data integration challenge, because every hospital chain stores its data with different formats, different encodings for common terms, etc., the social value gained from solving it is just huge," Stonebraker continues. But that also creates an incredibly difficult privacy problem one that I believe is not a technical issue. Because by and large, if you're looking for an interesting medical query, you're not looking for common events: you're looking for rare events and at least to my knowledge, there aren't any technical solutions that will allow access to rare events without indirectly disclosing who in fact the events belong to."

We must find the right balance between making data public and respecting people's privacy. It's a question of "how public is too public," said Professor Blei.

'To me, we get the most bang for our buck if we make everything public; however, that's going to have some serious security and morality issues with it, so we shouldn't do that," the ACM Fellow said. "But making everything completely private and not benefiting from that data also doesn't seem like a great option. I think this is a difficult, thorny issue. It lives at the intersection of policy, philosophy, morality, computer

science, data science, and machine learning."

There are no simple answers to questions like these. But considering the societal good that will come from solving them, it's a challenge that must be undertaken.

How to Survive Tech's Chaotic Transformation (https://www.datanami.com/2017/06/30/survive-

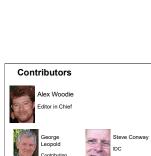
Big-Data Backlash: Medical Database Raises Privacy Concerns (https://www.datanami.com /2014/04/17/big-data\_backlash\_medical\_database\_raises\_privacy\_concerns/)



reddit this! (//www.reddit.com/submit?url=https://www.datanami.com/2017/07/05/whats-

challenging-big-data-now-integration-privacy/)

Tags: big data (https://www.datanami.com/tag/big-data/), integration (https://www.datanami.com/ /tag/integration/), machine learning (https://www.datanami.com/tag/machine-learning/), privacy (https://www.datanami.com/tag/privacy/)





© 2017 Datanami. All Rights Reserved. A Tabor Communications Publication

Contact (https://www.datanami.com/about/contact/) Back to Top About Datanami (https://www.datanami.com/about/)

2 sur 2