Privacy and accuracy: How Cynthia Dwork is making data analysis better

Posted August 7, 2015 By **Allison Linn**

19

Cynthia Dwork has spent much of her career working on ways to ensure that your personal data stay private even when it is being used for scientific research.

Now, she's also applying those mathematical methods to making certain that the conclusions researchers draw from analyzing big data sets are as accurate as possible.

Dwork, a cryptographer and distinguished scientist at Microsoft Research, and several colleagues recently published a paper in *Science* magazine showing how their groundbreaking work on differential privacy also can help researchers guarantee the accuracy of their results.

We spoke with her about her work and what has inspired it.

ALLISON LINN: I want to start by talking about differential privacy. How would you explain it to a person who isn't an expert in this field?

CYNTHIA DWORK: Differential privacy is a definition of privacy that is tailored to privacy-preserving data analysis.

So, assume that you have a large data set that's full of very useful but also very sensitive information. You'd like to be able to release statistics about that data set while simultaneously preserving the privacy of everybody who's in the data set.

What differential privacy says is that, essentially, the same things are learned whether any individual opts in or opts out of the data set. So what that means is I wouldn't be harmed by things that you learn from the data set. You won't learn anything about me that you wouldn't learn had I not been included.

ALLISON LINN: Can you give me a real-life example of when a researcher might want to use one of these techniques?

CYNTHIA DWORK: So imagine that somebody asks, "How many members of the House of Representatives have sickle cell trait?" Our intuition says that getting an exact answer to that shouldn't compromise the privacy of anybody in the House of Representatives because it's a pretty big set of people and you're just getting one number back.

But now suppose you have, in addition to the answer to that question, the exact answer to the question, "How many members of the House of Representatives, other than the Speaker of the House, have the sickle cell trait?"

Now, that also, by itself, seems like an innocuous question and getting the answer doesn't seem to cause any problems because it's still a pretty big set of people that you're asking about.

But if you take these two answers together and you subtract one from the other, then you will learn the sickle cell status of the Speaker of the House.

ALLISON LINN: What drew you to this area of research?

CYNTHIA DWORK: Conversations with the philosopher Helen Nissenbaum. Nissenbaum is a philosopher who studies issues that arise in the context of new technologies, and she was doing some work on privacy in public. What is privacy in public when you have video cameras everywhere?

That got me thinking about privacy in general, and I realized that privacy is this sort of catch-all phrase that means many different things in different contexts. I wanted to bite off a piece of the privacy puzzle that I would be able to chew on, and so I thought of privacy-preserving data analysis.

ALLISON LINN: You have a new paper coming out this week in *Science* that builds on some of the ideas around differential privacy to focus on data accuracy. Can you tell me a little bit about this work?

Winners and losers in a melting Arctic
Molecular mechanisms of neurodegeneration
Hints of Weyl fermions found in solids
Science
AAAS
Concrete connection
High-strength rock explains caldera uplift and Roman concrete

Photo: © Roger Ressmeyer/CORBIS

CYNTHIA DWORK: There is a technique from the machine learning community where you take your whole data set and you split it into two parts: a training set and a holdout set. Then, you do whatever you want on the training set in order to try to come up with some hypotheses about the general population. To check the validity of your conclusion, you test whether the hypothesis holds on the holdout set.

So far, so good. But now suppose that you'd like to do more study of your training set. Now, suddenly, the questions that you're asking of your training set depend on the holdout set, and so the holdout set can no longer be looked at as fresh data that's totally independent of everything that you've done so far.

What we show is that if you only access the holdout set through a differentially private mechanism, then it is okay to reuse it over and over again.

ALLISON LINN: How does that guarantee that people aren't going to draw spurious conclusions from the data?

CYNTHIA DWORK: So let's say that this is actually a very large data set. You publish your conclusions, and now somebody else comes along and they say, "Oh, that's interesting, I want to study a few other things in that data set." They can do that, and then they can check their conclusions on this same holdout set, and this can be repeated.

We're trying to capture the fact that science is an adaptive process. The second question you asked might depend on the answer to the first question. The second study or the fifth study may depend on what was published in the first four studies.

ALLISON LINN: Can you give me an example of how your new method could be applied to helping people ensure that their data is accurate?

CYNTHIA DWORK: We're coming to a time when data sets will be very, very large, and lots of people will be studying the same data sets. I think this is going to be happening with medical data, for example, and with genomic data.

I don't think it will be feasible to always go out and recruit completely fresh samples and start all over again, so I think that this question of remaining statistically valid in the adaptive scenario where new questions and new studies depend on the outcomes of previous studies is going to become more and more important. We are proposing a tool that will help with this process.

Related:

The Reusable Holdout: Preserving validity in adaptive data analysis, by Cynthia Dwork, Microsoft Research, Vitaly Feldman, IBM Almaden Research Center, Moritz Hardt, Google Research, Toniann Pitassi, University of Toronto, Omer Reingold, Samsung Research America, and Aaron Roth, University of Pennsylvania
Differential privacy
The Algorithmic Foundations of Differential Privacy
Penn Research helps develop algorithm aimed at combating science's reproducibility problem
Preserving validity in adaptive data analysis

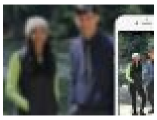
Allison Linn is a senior writer at Microsoft Research.
Follow her on Twitter.

Back to top

Featured Posts

Microsoft researchers achieve speech recognition milestone

Microsoft researcher translates defense intelligence to business intelligence



Microsoft Pix gives the iPhone camera an artificial brain

Most Popular

How Microsoft computer scientists and researchers are working to "solve" cancer

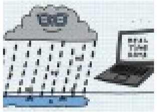
Microsoft researchers achieve speech recognition milestone

Microsoft researcher translates defense intelligence to business intelligence

Tags big data data privacy microsoft research next at microsoft privacy

Related Stories

The Next at Microsoft Podcast Ep. 4 – Using Big Data to Predict the Future

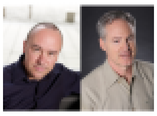


Welcome to the fourth episode of the Next at Microsoft podcast series. In this edition, AccuWeather Vice President Jon … *Read more »*

The next evolution of machine learning: Machine teaching

Microsoft researchers are at the forefront of efforts to help people without a machine learning background teach their systems … *Read more »*

The future of artificial intelligence: Myths, realities and aspirations



Only a few years ago, it would have seemed improbable to assume that a piece of technology could quickly … *Read more »*

Contact Us Terms of Use Trademarks Privacy & Cookies About our ads