The New York Times
Bits The Business of Technology

JANUARY 25, 2012, 2:30 PM

At Davos, Discussions of a Global Data Deluge

By NICK BILTON

Christian Hartmann/Reuters
DAVOS, Switzerland — Each year, a new theme reverberates through technology conferences around the world. A few years ago it was predictions of the coming wave of social media, location-based services and mobile, long before they became mainstream.

Today's topic de jour: data. Lots of data.

What makes the data discussion different than in previous years is that it is being discussed in high-profile nontechnology meetings too. This is clearly evident at this year's annual World Economic Forum.

Meetings here this week include: "From data to decisions: How are new approaches to data intelligence transforming decision-making?" "Data deluge and citizen science." "Incidents from digital crime to massive incidents of data theft are increasing significantly, with major political, social and economic implications." "How is big data being used to uncover individual and collective human dynamics?"

The discussions are not confined to technology attendees either. Chancellors, bankers and educators meeting at the conference are being asked to discuss what the forum calls a growing data deluge and how to manage it.

A 2012 report released by the World Economic Forum, titled, "Big Data, Big Impact: New Possibilities for International Development," outlines some of the possibilities data can bring around the globe to business and education. It also warns of its potential privacy implications.

The report says data is a new economic asset class, which touches all aspects of society, regardless of income or location.

"Big data represents one of these seismic economic shifts that happens every 10

years," said Zach Bogue, co-founder of a stealth data investment fund called Data Collective, who was attending Davos. "In some sense, this data has always existed, but until now the bandwidth, storage capability and compute power haven't existed to harness it."

Yet as there is talk of data, the discussion of privacy is not far behind.

Earlier this week in Munich, Viviane Reding, the European justice commissioner, repeatedly talked about data in respect to privacy. Ms. Reding said there were 27 laws that apply to data in Europe, most of which date back more than a decade and don't properly protect consumers today.

Ms. Reding outlined new regulations that were presented in Brussels on Wednesday and were designed to implement one sweeping data protection regulation that would apply to all of Europe.

The new regulations are part of the discussion at Davos as these new rules would drastically affect the way companies operate and collect data. For example, one component of this legislation will require companies to communicate to users why they are collecting this data and how long it is being stored on company servers.

Although the proposed data law being presented by Ms. Reding will apply to companies in Europe, it will clearly affect technology companies around the globe too, including Facebook, Twitter and Google in the United States.

At the World Economic Forum, the discussion is also focusing on one of the toughest regulation challenges in regards to data collection: How to manage a balancing act between governments overseeing data collection and its actions stifling innovation. More importantly, regulators hope to figure out a global solution to data collection.

As the World Economic Forum reports says: "Concerted action is needed by governments, development organizations and companies to ensure that this data helps the individuals and communities who create it."

February 11, 2012

The Age of Big Data

By STEVE LOHR

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

Mo Zhou was snapped up by I.B.M. last summer, as a freshly minted Yale M.B.A., to join the technology company's fast-growing ranks of data consultants. They help businesses make sense of an explosion of data — Web traffic and social network comments, as well as software and sensors that monitor shipments, suppliers and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers," says Ms. Zhou, whose job as a data analyst suits her skills.

To exploit the data flood, America will need many more like her. A report last year by the McKinsey Global Institute, the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, whether retrained or hired.

The impact of data abundance extends well beyond business. Justin Grimmer, for example, is one of the new breed of political scientists. A 28-year-old assistant professor at Stanford, he combined math with political science in his undergraduate and graduate studies, seeing "an opportunity because the discipline is becoming increasingly data-intensive." His research involves the computer-automated analysis of blog postings, Congressional speeches and press releases, and news articles, looking for insights into how political ideas spread.

The story is similar in fields as varied as science and sports, advertising and public health — a drift toward data-driven discovery and decision-making. "It's a revolution," says Gary King, director of Harvard's Institute for Quantitative Social Science. "We're really just getting under way. But the march of quantification, made possible by enormous new sources of data, will sweep through academia, business and government.

There is no area that is going to be untouched."

Welcome to the Age of Big Data. The new megarich of Silicon Valley, first at Google and now Facebook, are masters at harnessing the data of the Web — online searches, posts and messages — with Internet advertising. At the World Economic Forum last month in Davos, Switzerland, Big Data was a marquee topic. A report by the forum, "Big Data, Big Impact," declared data a new class of economic asset, like currency or gold.

Rick Smolan, creator of the "Day in the Life" photography series, is planning a project later this year, "The Human Face of Big Data," documenting the collection and uses of data. Mr. Smolan is an enthusiast, saying that Big Data has the potential to be "humanity's dashboard," an intelligent tool that can help combat poverty, crime and pollution. Privacy advocates take a dim view, warning that Big Data is Big Brother, in corporate clothing.

What is Big Data? A meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions. There is a lot more data, all the time, growing at 50 percent a year, or more than doubling every two years, estimates IDC, a technology research firm. It's not just more streams of data, but entirely new ones. For example, there are now countless digital sensors worldwide in industrial equipment, automobiles, electrical meters and shipping crates. They can measure and communicate location, movement, vibration, temperature, humidity, even chemical changes in the air.

Link these communicating sensors to computing intelligence and you see the rise of what is called the Internet of Things or the Industrial Internet. Improved access to information is also fueling the Big Data trend. For example, government data — employment figures and other information — has been steadily migrating onto the Web. In 2009, Washington opened the data doors further by starting Data.gov, a Web site that makes all kinds of government data accessible to the public.

Data is not only becoming more available but also more understandable to computers. Most of the Big Data surge is data in the wild — unruly stuff like words, images and video on the Web and those streams of sensor data. It is called unstructured data and is not typically grist for traditional databases.

But the computer tools for gleaning knowledge and insights from the Internet era's vast

trove of unstructured data are fast gaining ground. At the forefront are the rapidly advancing techniques of artificial intelligence like natural-language processing, pattern recognition and machine learning.

Those artificial-intelligence technologies can be applied in many fields. For example, Google's search and ad business and its experimental robot cars, which have navigated thousands of miles of California roads, both use a bundle of artificial-intelligence tricks. Both are daunting Big Data challenges, parsing vast quantities of data and making decisions instantaneously.

The wealth of new data, in turn, accelerates advances in computing — a virtuous circle of Big Data. Machine-learning algorithms, for example, learn on data, and the more data, the more the machines learn. Take Siri, the talking, question-answering application in iPhones, which Apple introduced last fall. Its origins go back to a Pentagon research project that was then spun off as a Silicon Valley start-up. Apple bought Siri in 2010, and kept feeding it more data. Now, with people supplying millions of questions, Siri is becoming an increasingly adept personal assistant, offering reminders, weather reports, restaurant suggestions and answers to an expanding universe of questions.

To grasp the potential impact of Big Data, look to the microscope, says Erik Brynjolfsson, an economist at Massachusetts Institute of Technology's Sloan School of Management. The microscope, invented four centuries ago, allowed people to see and measure things as never before — at the cellular level. It was a revolution in measurement.

Data measurement, Professor Brynjolfsson explains, is the modern equivalent of the microscope. Google searches, Facebook posts and Twitter messages, for example, make it possible to measure behavior and sentiment in fine detail and as it happens.

In business, economics and other fields, Professor Brynjolfsson says, decisions will increasingly be based on data and analysis rather than on experience and intuition. "We can start being a lot more scientific," he observes.

There is plenty of anecdotal evidence of the payoff from data-first thinking. The best-known is still "Moneyball," the 2003 book by Michael Lewis, chronicling how the low-budget Oakland A's massaged data and arcane baseball statistics to spot undervalued

players. Heavy data analysis had become standard not only in baseball but also in other sports, including English soccer, well before last year's movie version of "Moneyball," starring Brad Pitt.

Retailers, like Walmart and Kohl's, analyze sales, pricing and economic, demographic and weather data to tailor product selections at particular stores and determine the timing of price markdowns. Shipping companies, like U.P.S., mine data on truck delivery times and traffic patterns to fine-tune routing.

Online dating services, like Match.com, constantly sift through their Web listings of personal characteristics, reactions and communications to improve the algorithms for matching men and women on dates. Police departments across the country, led by New York's, use computerized mapping and analysis of variables like historical arrest patterns, paydays, sporting events, rainfall and holidays to try to predict likely crime "hot spots" and deploy officers there in advance.

Research by Professor Brynjolfsson and two other colleagues, published last year, suggests that data-guided management is spreading across corporate America and starting to pay off. They studied 179 large companies and found that those adopting "data-driven decision making" achieved productivity gains that were 5 percent to 6 percent higher than other factors could explain.

The predictive power of Big Data is being explored — and shows promise — in fields like public health, economic development and economic forecasting. Researchers have found a spike in Google search requests for terms like "flu symptoms" and "flu treatments" a couple of weeks before there is an increase in flu patients coming to hospital emergency rooms in a region (and emergency room reports usually lag behind visits by two weeks or so).

Global Pulse, a new initiative by the United Nations, wants to leverage Big Data for global development. The group will conduct so-called sentiment analysis of messages in social networks and text messages — using natural-language deciphering software — to help predict job losses, spending reductions or disease outbreaks in a given region. The goal is to use digital early-warning signals to guide assistance programs in advance to, for example, prevent a region from slipping back into poverty.

In economic forecasting, research has shown that trends in increasing or decreasing

volumes of housing-related search queries in Google are a more accurate predictor of house sales in the next quarter than the forecasts of real estate economists. The Federal Reserve, among others, has taken notice. In July, the National Bureau of Economic Research is holding a workshop on "Opportunities in Big Data" and its implications for the economics profession.

Big Data is already transforming the study of how social networks function. In the 1960s, Stanley Milgram of Harvard used packages as his research medium in a famous experiment in social connections. He sent packages to volunteers in the Midwest, instructing them to get the packages to strangers in Boston, but not directly; participants could mail a package only to someone they knew. The average number of times a package changed hands was remarkably few, about six. It was a classic demonstration of the "small-world phenomenon," captured in the popular phrase "six degrees of separation."

Today, social-network research involves mining huge digital data sets of collective behavior online. Among the findings: people whom you know but don't communicate with often — "weak ties," in sociology — are the best sources of tips about job openings. They travel in slightly different social worlds than close friends, so they see opportunities you and your best friends do not.

Researchers can see patterns of influence and peaks in communication on a subject — by following trending hashtags on Twitter, for example. The online fishbowl is a window into the real-time behavior of huge numbers of people. "I look for hot spots in the data, an outbreak of activity that I need to understand," says Jon Kleinberg, a professor at Cornell. "It's something you can only do with Big Data."

Big Data has its perils, to be sure. With huge data sets and fine-grained measurement, statisticians and computer scientists note, there is increased risk of "false discoveries." The trouble with seeking a meaningful needle in massive haystacks of data, says Trevor Hastie, a statistics professor at Stanford, is that "many bits of straw look like needles."

Big Data also supplies more raw material for statistical shenanigans and biased fact-finding excursions. It offers a high-tech twist on an old trick: I know the facts, now let's find 'em. That is, says Rebecca Goldin, a mathematician at George Mason University, "one of the most pernicious uses of data."

Data is tamed and understood using computer and mathematical models. These models, like metaphors in literature, are explanatory simplifications. They are useful for understanding, but they have their limits. A model might spot a correlation and draw a statistical inference that is unfair or discriminatory, based on online searches, affecting the products, bank loans and health insurance a person is offered, privacy advocates warn.

Despite the caveats, there seems to be no turning back. Data is in the driver's seat. It's there, it's useful and it's valuable, even hip.

Veteran data analysts tell of friends who were long bored by discussions of their work but now are suddenly curious. "Moneyball" helped, they say, but things have gone way beyond that. "The culture has changed," says Andrew Gelman, a statistician and political scientist at Columbia University. "There is this idea that numbers and statistics are interesting and fun. It's cool now."

Steve Lohr is a technology reporter for The New York Times.

March 24, 2012

Avalanches of Words, Sifted and Sorted

By ANNE EISENBERG

IT just keeps growing — the vast electronic archive of books, journals and scholarly literature stored on the Web. But scientists are aiming to keep up with this trove of collective knowledge by devising computer-based tools to winnow and quantify it.

David M. Blei of Princeton University is among those who are teaching computers to sift through the digital pages of books and articles and categorize the contents by subject, even when that subject isn't stated explicitly.

For decades, of course, librarians and many others have labeled books and documents with keywords. "But human categorization can only go so far," said Dr. Blei, an associate professor in computer science. "We don't have the human power to read and tag all this information."

To cope with the information explosion, Dr. Blei and other researchers write algorithms so that computers can sift through millions of works and find their common themes by sorting related words into categories. It's a field called probabilistic topic modeling.

Other research tools identify shifts in language over time that could signal important cultural, scientific or historical changes. At Harvard, Erez Lieberman Aiden and Jean-Baptiste Michel, who jointly lead a group there called the Cultural Observatory, will soon inaugurate a browser that searches for such language changes in a large online repository of scientific papers known as arXiv (pronounced like "archive").

Users will be able to type in one or two words at the site, called Bookworm-arXiv, and immediately see a graph showing the ups and downs of the phrase's use in the archive, Dr. Michel said. (A test version is at arxiv.culturomics.org.) Users can then click on the

graph and drill down to read the original papers in which the terms appear, tracing ideas back toward their roots, or to spots where scientific ideas spread from one field to another.

The new analytical techniques won't replace the close reading and interpretation of text that is the province of scholars, said Anthony T. Grafton, a history professor at Princeton and a former president of the American Historical Association.

"But these tools have enormous implications," he said, in their ability to reveal unexpected patterns and associations in the historical record. "These are tools that can pick up big changes," he said. "You can't do this by using older, conventional means of reading books and taking notes."

BOOKWORM-ARXIV will burrow its way through data stored in roughly 743,000 or so papers that have been uploaded by scientists, said Paul Ginsparg, founder of arXiv, Authors typically send their papers to arXiv as "preprints" or unpublished manuscripts before the works appear in journals. Most of the research is in physics, mathematics, computer science, statistics and the quantitative parts of biology and finance, said Dr. Ginsparg, a professor of physics and information science at Cornell.

The Bookworm-arXiv interface is the latest in a series of tools developed by the Cultural Observatory. Late in 2010, in collaboration with Google, the lab released the Google n-gram viewer, which lets people search for a phrase of up to five words in Google's database of scanned books and see the frequency of the words over time in a graph, Dr. Aiden said.

The n-gram viewer is a powerful tool, said Dr. Grafton at Princeton. For example, he said, it could trace the disappearance of the names of scientists and artists who were censored by the Nazis in Germany.

But the n-gram, however useful, has a disadvantage. It does not let users click through to the original documents, because many books included in the Google database are under copyright.

Readers who use the arXiv interface will be able to click through to the original text. "The papers are not behind a paywall," Dr. Ginsparg said.

Steven Pinker, an author, cognitive scientist and Harvard professor, collaborated with the observatory team during development of the n-gram viewer. He said the new interface might be particularly useful for historians of science. “They will increasingly be able to test their explanations, conjectures and hypotheses by looking at the rise and fall of phrases in the scientific literature,” he said.

Dr. Aiden said the worlds of Google’s scanned books and arXiv’s papers were just the beginning for the observatory. “We plan on moving on soon to newspapers, blogs, tweets and other aspects of the historical record,” he said.

Dr. Ginsparg says he has already been trying out the topic modeling algorithms of Dr. Blei and others on arXiv’s collection of scientific papers.

“The technology gives you a way to home in for finer grains of similarity between articles,” Dr. Ginsparg said, “ones that might not be detected by a keyword search.”

E-mail: novelties@nytimes.com.

OPEN

MORE IN BUSINESS DAY (1 OF 28 ARTICLES)

Awaiting Health Law Ruling, and Preparing Plan B

Read More »