The Promise and Peril of Big Data

David Bollier
Rapporteur

THE ASPEN INSTITUTE

Communications and Society Program
Charles M. Firestone
Executive Director
Washington, DC
2010

To purchase additional copies of this report, please contact:

The Aspen Institute
Publications Office
P.O. Box 222
109 Houghton Lab Lane
Queenstown, Maryland 21658
Phone: (410) 820-5326
Fax: (410) 827-9174
E-mail: publications@aspeninstitute.org

For all other inquiries, please contact:

The Aspen Institute
Communications and Society Program
One Dupont Circle, NW
Suite 700
Washington, DC 20036
Phone: (202) 736-5818
Fax: (202) 467-0790

Charles M. Firestone
Executive Director

Patricia K. Kelly
Assistant Director

The Aspen Institute
One Dupont Circle, NW
Suite 700
Washington, DC 20036

Published in the United States of America in 2010
by The Aspen Institute

Printed in the United States of America

ISBN: 0-89843-516-1

10-001

1762/CSP/10-BK

Contents

This report is written from the perspective of an informed observer at the Eighteenth Annual Aspen Institute Roundtable on Information Technology. Unless attributed to a particular person, none of the comments or ideas contained in this report should be taken as embodying the views or carrying the endorsement of any specific participant at the Conference.

Foreword

According to a recent report[1], the amount of digital content on the Internet is now close to five hundred billion gigabytes. This number is expected to double within a year. Ten years ago, a single gigabyte of data seemed like a vast amount of information. Now, we commonly hear of data stored in terabytes or petabytes. Some even talk of exabytes or the yottabyte, which is a trillion terabytes or, as one website describes it, "everything that there is."[2]

The explosion of mobile networks, cloud computing and new technologies has given rise to incomprehensibly large worlds of information, often described as "Big Data." Using advanced correlation techniques, data analysts (both human and machine) can sift through massive swaths of data to predict conditions, behaviors and events in ways unimagined only years earlier. As the following report describes it:

> Google now studies the timing and location of search-engine queries to predict flu outbreaks and unemployment trends before official government statistics come out. Credit card companies routinely pore over vast quantities of census, financial and personal information to try to detect fraud and identify consumer purchasing trends.
>
> Medical researchers sift through the health records of thousands of people to try to identify useful correlations between medical treatments and health outcomes.
>
> Companies running social-networking websites conduct "data mining" studies on huge stores of personal information in attempts to identify subtle consumer preferences and craft better marketing strategies.
>
> A new class of "geo-location" data is emerging that lets companies analyze mobile device data to make

1. See http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm.
2. See http://www.uplink.freeuk.com/data.html.

> intriguing inferences about people's lives and the economy. It turns out, for example, that the length of time that consumers are willing to travel to shopping malls—data gathered from tracking the location of people's cell phones—is an excellent proxy for measuring consumer demand in the economy.

But this analytical ability poses new questions and challenges. For example, what are the ethical considerations of governments or businesses using Big Data to target people without their knowledge? Does the ability to analyze massive amounts of data change the nature of scientific methodology? Does Big Data represent an evolution of knowledge, or is *more actually less* when it comes to information on such scales?

The Aspen Institute Communications and Society Program convened 25 leaders, entrepreneurs, and academics from the realms of technology, business management, economics, statistics, journalism, computer science, and public policy to address these subjects at the 2009 Roundtable on Information Technology.

This report, written by David Bollier, captures the insights from the three-day event, exploring the topic of Big Data and inferential software within a number of important contexts. For example:

- Do huge datasets and advanced correlation techniques mean we no longer need to rely on hypothesis in scientific inquiry?

- When does "now-casting," the search through massive amounts of aggregated data to estimate individual behavior, go over the line of personal privacy?

- How will healthcare companies and insurers use the correlations of aggregated health behaviors in addressing the future care of patients?

The Roundtable became most animated, however, and found the greatest promise in the application of Big Data to the analysis of systemic risk in financial markets.

A system of streamlined financial reporting, massive transparency, and "open source analytics," they concluded, would serve better than past regulatory approaches. Participants rallied to the idea, furthermore, that a National Institute of Finance could serve as a resource for the financial regulators and investigate where the system failed in one way or another.

Acknowledgements

We want to thank McKinsey & Company for reprising as the senior sponsor of this Roundtable. In addition, we thank Bill Coleman, Google, the Markle Foundation, and Text 100 for sponsoring this conference; James Manyika, Bill Coleman, John Seely Brown, Hal Varian, Stefaan Verhulst and Jacques Bughin for their suggestions and assistance in designing the program and recommending participants; Stefaan Verhulst, Jacques Bughin and Peter Keefer for suggesting readings; and Kiahna Williams, project manager for the Communications and Society Program, for her efforts in selecting, editing, and producing the materials and organizing the Roundtable; and Patricia Kelly, assistant director, for editing and overseeing the production of this report.

Charles M. Firestone
Executive Director
Communications and Society Program
Washington, D.C.
January 2010

The Promise and Peril of Big Data

David Bollier

The Promise and Peril of Big Data

David Bollier

It has been a quiet revolution, this steady growth of computing and databases. But a confluence of factors is now making Big Data a powerful force in its own right.

Computing has become ubiquitous, creating countless new digital puddles, lakes, tributaries and oceans of information. A menagerie of digital devices has proliferated and gone mobile—cell phones, smart phones, laptops, personal sensors—which in turn are generating a daily flood of new information. More business and government agencies are discovering the strategic uses of large databases. And as all these systems begin to interconnect with each other and as powerful new software tools and techniques are invented to analyze the data for valuable inferences, a radically new kind of "knowledge infrastructure" is materializing. A new era of Big Data is emerging, and the implications for business, government, democracy and culture are enormous.

…a radically new kind of "knowledge infrastructure" is materializing. A new era of Big Data is emerging….

Computer databases have been around for decades, of course. What is new are the growing scale, sophistication and ubiquity of data-crunching to identify novel patterns of information and inference. Data is not just a back-office, accounts-settling tool any more. It is increasingly used as a real-time decision-making tool. Researchers using advanced correlation techniques can now tease out potentially useful patterns of information that would otherwise remain hidden in petabytes of data (a petabyte is a number starting with 1 and having 15 zeros after it).

Google now studies the timing and location of search-engine queries to predict flu outbreaks and unemployment trends before official

government statistics come out. Credit card companies routinely pore over vast quantities of census, financial and personal information to try to detect fraud and identify consumer purchasing trends.

Medical researchers sift through the health records of thousands of people to try to identify useful correlations between medical treatments and health outcomes.

Companies running social-networking websites conduct "data mining" studies on huge stores of personal information in attempts to identify subtle consumer preferences and craft better marketing strategies.

A new class of "geo-location" data is emerging that lets companies analyze mobile device data to make intriguing inferences about people's lives and the economy. It turns out, for example, that the length of time that consumers are willing to travel to shopping malls—data gathered from tracking the location of people's cell phones—is an excellent proxy for measuring consumer demand in the economy.

The inferential techniques being used on Big Data can offer great insight into many complicated issues, in many instances with remarkable accuracy and timeliness. The quality of business decision-making, government administration, scientific research and much else can potentially be improved by analyzing data in better ways.

But critics worry that Big Data may be misused and abused, and that it may give certain players, especially large corporations, new abilities to manipulate consumers or compete unfairly in the marketplace. Data experts and critics alike worry that potential abuses of inferential data could imperil personal privacy, civil liberties and consumer freedoms.

Because the issues posed by Big Data are so novel and significant, the Aspen Institute Roundtable on Information Technology decided to explore them in great depth at its eighteenth annual conference. A distinguished group of 25 technologists, economists, computer scientists, entrepreneurs, statisticians, management consultants and others were invited to grapple with the issues in three days of meetings, from August 4 to 7, 2009, in Aspen, Colorado. The discussions were moderated by Charles M. Firestone, Executive Director of the Aspen Institute Communications and Society Program. This report is an interpretive synthesis of the highlights of those talks.

How to Make Sense of Big Data?

To understand implications of Big Data, it first helps to understand the more salient uses of Big Data and the forces that are expanding inferential data analysis. Historically, some of the most sophisticated users of deep analytics on large databases have been Internet-based companies such as search engines, social networking websites and online retailers. But as magnetic storage technologies have gotten cheaper and high-speed networking has made greater bandwidth more available, other industries, government agencies, universities and scientists have begun to adopt the new data-analysis techniques and machine-learning systems.

Certain technologies are fueling the use of inferential data techniques. New types of remote censors are generating new streams of digital data from telescopes, video cameras, traffic monitors, magnetic resonance imaging machines, and biological and chemical sensors monitoring the environment. Millions of individuals are generating roaring streams of personal data from their cell phones, laptops, websites and other digital devices.

The growth of cluster computing systems and cloud computing facilities are also providing a hospitable context for the growth of inferential data techniques, notes computer researcher Randal Bryant and his colleagues.[1] Cluster computing systems provide the storage capacity, computing power and high-speed local area networks to handle large data sets. In conjunction with "new forms of computation combining statistical analysis, optimization and artificial intelligence," writes Bryant, researchers "are able to construct statistical models from large collections of data to infer how the system should respond to new data." Thus companies like Netflix, the DVD-rental company, can use automated machine-learning to identify correlations in their customers' viewing habits and offer automated recommendations to customers.

Within the tech sector, which is arguably the most advanced user of Big Data, companies are inventing new services such that give driving directions (MapQuest), provide satellite images (Google Earth) and consumer recommendations (TripAdvisor). Retail giants like Wal-Mart assiduously study their massive sales databases—267 million transactions a day—to help them devise better pricing strategies, inventory control and advertising campaigns.

Intelligence agencies must now contend with a flood of data from its own satellites and telephone intercepts as well as from the Internet and publications. Many scientific disciplines are becoming more computer-based and data-driven, such as physics, astronomy, oceanography and biology.

Data Correlation or Scientific Models?

As the deluge of data grows, a key question is how to make sense of the raw information. How can researchers use statistical tools and computer technologies to identify meaningful patterns of information? How shall significant correlations of data be interpreted? What is the role of traditional forms of scientific theorizing and analytic models in assessing data?

Chris Anderson, the Editor-in-Chief of *Wired* magazine, ignited a small firestorm in 2008 when he proposed that "the data deluge makes the scientific method obsolete."[2] Anderson argued the provocative case that, in an age of cloud computing and massive datasets, the real challenge is not to come up with new taxonomies or models, but to sift through the data in new ways to find meaningful correlations.

> At the petabyte scale, information is not a matter of simple three and four-dimensional taxonomy and order but of dimensionally agnostic statistics. It calls for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later. For instance, Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising—it just assumed that better data, with better analytic tools, would win the day. And Google was right.

Physics and genetics have drifted into arid, speculative theorizing, Anderson argues, because of the inadequacy of testable models. The solution, he asserts, lies in finding meaningful correlations in massive piles of Big Data, "Petabytes allow us to say: 'Correlation is enough.' We can stop looking for models. We can analyze the data without

hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."

J. Craig Venter used supercomputers and statistical methods to find meaningful patterns from shotgun gene sequencing, said Anderson. Why not apply that methodology more broadly? He asked, "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?"

Conference participants agreed that there is a lot of useful information to be gleaned from Big Data correlations. But there was a strong consensus that Anderson's polemic goes too far. "Unless you create a model of what you think is going to happen, you can't ask questions about the data," said William T. Coleman. "You have to have some basis for asking questions."

Researcher John Timmer put it succinctly in an article at the Ars Technica website, "Correlations are a way of catching a scientist's attention, but the models and mechanisms that explain them are how we make the predictions that not only advance science, but generate practical applications."[3]

Hal Varian, Chief Economist at Google, agreed with that argument, "Theory is what allows you to extrapolate outside the observed domain. When you have a theory, you don't want to test it by just looking at the data that went into it. You want to make some new prediction that's implied by the theory. If your prediction is validated, that gives you some confidence in the theory. There's this old line, 'Why does deduction work? Well, because you can prove it works. Why does induction work? Well, it's always worked in the past.'"

Extrapolating from correlations can yield specious results even if large data sets are used. The classic example may be "My TiVO Thinks I'm Gay." *The Wall Street Journal* once described a TiVO customer who gradually came to realize that his TiVO recommendation system thought he was gay because it kept recommending gay-themes films. When the customer began recording war movies and other "guy stuff" in an effort to change his "reputation," the system began recommending documentaries about the Third Reich.[4]

Another much-told story of misguided recommendations based on statistical correlations involved Jeff Bezos, the founder of Amazon. To demonstrate the Amazon recommendation engine in front of an audience, Bezos once called up his own set of recommendations. To his surprise, the system's first recommendation was *Slave Girls from Infinity*—a choice triggered by Bezos' purchase of a DVD of *Barbarella*, the Jane-Fonda-as-sex-kitten film, the week before.

Using correlations as the basis for forecasts can be slippery for other reasons. Once people know there is an automated system in place, they may deliberately try to game it. Or they may unwittingly alter their behavior.

It is the "classic Heisenberg principle problem," said Kim Taipale, the Founder and Executive Director of the Center for Advanced Studies in Science and Technology. "As soon as you put up a visualization of data, I'm like—whoa!—I'm going to 'Google bomb' those questions so that I can change the outcomes." ("Google bombing" describes concerted, often-mischievous attempts to game the search-algorithm of the Google search engine in order to raise the ranking of a given page in the search results.[5])

The sophistication of recommendation-engines is improving all the time, of course, so many silly correlations may be weeded out in the future. But no computer system is likely to simulate the level of subtlety and personalization that real human beings show in dynamic social contexts, at least in the near future. Running the numbers and finding the correlations will never be enough.

Theory is important, said Kim Taipale, because "you have to have something you can come back to in order to say that something is right or wrong." Michael Chui, Senior Expert at McKinsey & Company, agrees: "Theory is about predicting what you haven't observed yet. Google's headlights only go as far as the data it has seen. One way to think about theories is that they help you to describe ontologies that already exist." (Ontology is a branch of philosophy that explores the nature of being, the categories used to describe it, and their ordered relationships with each other. Such issues can matter profoundly when trying to collect, organize and interpret information.)

Jeff Jonas, Chief Scientist, Entity Analytic Solutions at the IBM Software Group, offered a more complicated view. While he agrees

that Big Data does not invalidate the need for theories and models, Jonas believes that huge datasets may help us "find and see dynamically changing ontologies without having to try to prescribe them in advance. Taxonomies and ontologies are things that you might discover by observation, and watch evolve over time."

John Clippinger, Co-Director of the Law Lab at Harvard University, said: "Researchers have wrestled long and hard with language and semantics to try to develop some universal ontologies, but they have not really resolved that. But it's clear that you have to have some underlying notion of mechanism. That leads me to think that there may be some self-organizing grammars that have certain properties to them—certain mechanisms—that can yield certain kinds of predictions. The question is whether we can identify a mechanism that is rich enough to characterize a wide range of behaviors. That's something that you can explore with statistics."

How Should Theories be Crafted in an Age of Big Data?

If correlations drawn from Big Data are suspect, or not sturdy enough to build interpretations upon, how then shall society construct models and theories in the age of Big Data?

Patrick W. Gross, Chairman of the Lovell Group, challenged the either/or proposition that either scientific models or data correlations will drive future knowledge. "In practice, the theory and the data reinforce each other. It's not a question of data correlations versus theory. The use of data for correlations allows one to test theories and refine them."

That may be, but how should theory-formation proceed in light of the oceans of data that can now be explored? John Seely Brown, Independent Co-Chair of Deloitte Center for the Edge, believes that we may need to devise new methods of theory formation: "One of the big problems [with Big Data] is how to determine if something is an outlier or not," and therefore can be disregarded. "In some ways, the more data you have, the more basis you have for deciding that something is an outlier. You have more confidence in deciding what to knock out of the data set—at least, under the Bayesian and correlational-type theories of the moment."

But this sort of theory-formation is fairly crude in light of the keen and subtle insights that might be gleaned from Big Data, said Brown: "Big Data suddenly changes the whole game of how you look at the ethereal odd data sets." Instead of identifying outliers and "cleaning" datasets, theory formation using Big Data allows you to "craft an ontology and subject it to tests to see what its predictive value is."

He cited an attempt to see if a theory could be devised to compress the English language using computerized, inferential techniques. "It turns out that if you do it just right—if you keep words as words—you can compress the language by *x* amount. But if you actually build a theory-formation system that ends up discovering the morphology of English, you can radically compress English. The catch was, how do you build a machine that actually starts to invent the ontologies and look at what it can do with those ontologies?"

"…The more data there is, the better my chances of finding the 'generators' for a new theory."

John Seely Brown

Before huge datasets and computing power could be applied to this problem, researchers had rudimentary theories about the morphology of the English language. "But now that we have 'infinite' amounts of computing power, we can start saying, 'Well, maybe there are many different ways to develop a theory.'"

In other words, the data once perceived as "noise" can now be reconsidered with the rest of the data, leading to new ways to develop theories and ontologies. Or as Brown put it, "How can you invent the 'theory behind the noise' in order to de-convolve it in order to find the pattern that you weren't supposed to find? The more data there is, the better my chances of finding the 'generators' for a new theory."

Jordan Greenhall suggested that there may be two general ways to develop ontologies. One is basically a "top down" mode of inquiry that applies familiar philosophical approaches, using *a priori* categories. The other is a "bottom up" mode that uses dynamic, low-level data and builds ontologies based on the contingent information identified through automated processes.

For William T. Coleman, the real challenge is building new types of machine-learning tools to help explore and develop ontologies: "We

have to learn how to make data tagged and self-describing at some level. We have to be able to discover ontologies based on the questions and problems we are posing." This task will require the development of new tools so that the deep patterns of Big Data can be explored more flexibly yet systematically.

Bill Stensrud, Chairman and Chief Executive Officer of InstantEncore, a website that connects classical music fans with their favorite artists, said, "I believe in the future the big opportunity is going to be non-human-directed efforts to search Big Data, to find what questions can be asked of the data that we haven't even known to ask."

"The data *is* the question!" Jeff Jonas said. "I mean that seriously!"

Visualization as a Sense-Making Tool

Perhaps one of the best tools for identifying meaningful correlations and exploring them as a way to develop new models and theories, is computer-aided visualization of data. Fernanda B. Viégas, Research Scientist at the Visual Communications Lab at IBM, made a presentation that described some of the latest techniques for using visualization to uncover significant meanings that may be hidden in Big Data.

Google is an irresistible place to begin such an inquiry because it has access to such massive amounts of timely search-query data. "Is Google the ultimate oracle?" Viégas wondered. She was intrigued with "Google Suggest," the feature on the Google search engine that, as you type in your query, automatically lists the most-searched phrases that begin with the words entered. The feature serves as a kind of instant aggregator of what is on people's minds.

Viégas was fascinated with people using Google as a source of practical advice, and especially with the types of "why?" questions that they asked. For example, for people who enter the words "Why doesn't he…" will get Google suggestions that complete the phrase as "Why doesn't he *call*?", "Why doesn't he *like me*?" and "Why doesn't he *love me*?" Viégas wondered what the corresponding Google suggestions would be for men's queries, such as "Why doesn't *she*…?" Viégas found that men asked similar questions, but with revealing variations, such as "Why doesn't she *just leave*?"

Viégas and her IBM research colleague Martin Wattenberg developed a feature that visually displays the two genders' queries side by side, so that the differences can be readily seen. The program, now in beta form, is meant to show how Google data can be visually depicted to help yield interesting insights.

While much can be learned by automating the search process for the data or by "pouring" it into a useful visual format, sometimes it takes active human interpretation to spot the interesting patterns. For example, researchers using Google Earth maps made a striking discovery—that two out of three cows (based on a sample of 8,510 cattle in 308 herds from around the world) align their bodies with the magnetic north of the Earth's magnetic field.[6] No machine would have been capable of making this starting observation as something worth investigating.

Viégas offered other arresting examples of how the visualization of data can reveal interesting patterns, which in turn can help researchers develop new models and theories. Can the vast amount of data collected by remote sensors yield any useful patterns that might serve as building blocks for new types of knowledge? This is one hope for "smart dust," defined at Wikipedia as a "hypothetical wireless network of tiny microelectromechanical (MEMS) sensors, robots, or devices that can detect (for example) light, temperature, or vibration."

To test this idea with "dumb dust"—grains of salt and sand—scientists put the grains on the top of a plate to show how they respond when the frequency of audio signals directed at the bottom of the plate is manipulated. It turns out that the sand self-organizes itself into certain regular patterns, which have huge implications for the study of elasticity in building materials. So the study of remote sensor data can "help us understand how vibration works," said Viégas. It engendered new models of knowledge that "you could take from one domain (acoustics) and sort of apply to another domain (civil engineering)."

Visualization techniques for data are not confined to labs and tech companies; they are becoming a popular communications tool. Major newspapers such as *The New York Times* and *The Washington Post* are using innovative visualizations and graphics to show the significance of otherwise-dry numbers. Health websites like "Patients Like Me" invite people to create visualizations of their disease symptoms, which then become a powerful catalyst for group discussions and further scrutiny of the data.

Visualizations can help shine a light on some improbable sorts of social activity. Viégas describes a project of hers to map the "history flow" of edits made on Wikipedia articles. To learn how a given Wikipedia entry may have been altered over the course of months or years, Viégas developed a color-coded bar chart (resembling a "bar code" on products) that illustrates how many people added or changed the text of a given entry. By using this visualization for the "abortion" entry, Viégas found that certain periods were notable for intense participation by many people, followed by a blank "gash" of no color. The gash, she discovered, represented an "edit war"—a period of intense disagreement about what the text should say, followed by vandalism in which someone deleted the entire entry (after which Wikipedia editors reverted the entry to the preexisting text).

The visualizations are useful, said Viégas, because they help even the casual observer see what the "normal" participation dynamics are for a given Wikipedia entry. They also help researchers identify questions that might be explored statistically—for example, how often does vandalism occur and how quickly does the text get reverted? "This visualization tool gave us a way to do data exploration, and ask questions about things, and then do statistical analyses of them," said Viégas.

Stensrud agreed that visualization of Big Data gives you a way "to find things that you had no theory about and no statistical models to identify, but with visualization it jumps right out at you and says, 'This is bizarre.' "

Or as Lise Getoor, Associate Professor in the Department of Computer Science at the University of Maryland, articulated, visualizations allows researchers to "'explore the space of models' in more expansive ways. They can combine large data sets with statistical analysis and new types of computational resources to use various form functions in a systematic way and explore a wider space."

After exploring the broader modeling possibilities, said Getoor, "you still want to come back to do the standard hypothesis testing and analysis, to make sure that your data is well-curated and collected. One of the big changes is that you now have this observational data that helps you develop an initial model to explore."

Kim Taipale of the Center for Advanced Studies in Science and Technology warned that visualization design choices drive results every bit as much as traditional "data-cleaning" choices. Visualization tech-

niques contain embedded judgments. In Viégas' visualization models of Wikipedia editing histories, for example, she had to rely upon only a fraction of the available data—and the choices of which entries to study ("abortion" and "chocolate," among others) were idiosyncratic. Taipale believes disputes about the reliability of visualization designs resemble conversations about communications theory in the 1950s, which hosted similar arguments about how to interpret signal from noise.

Jesper Andersen, a statistician, computer scientist and Co-Founder of Freerisk, warned about the special risks of reaching conclusions from a single body of data. It is generally safer to use larger data sets from multiple sources. Visualization techniques do not solve this problem. "When you use visualization as an analytic tool, I think it can be very dangerous," he said. "Whenever you do statistics, one of the big things you find is spurious correlations"—apparent relationships or proximities that do not actually exist.

"You need to make sure the pattern that you *think* is there, is actually there," said Andersen. "Otherwise, the problem gets worse the bigger your data is—and we don't have any idea how to handle that in visualization because there is a very, very thin layer of truth on the data, because of tricks of the eye about whether what you see is actually there. The only way that we can solve this problem right now is to protect ourselves with a model."

So how can one determine what is accurate and objective? In a real-world business context, where the goal is to make money, the question may be moot, said Stephen Baker, *Business Week* journalist and author of *The Numerati.* "The companies featured in Amazon's recommendations don't have to be right. They just have to be better than the status quo and encourage more people to buy books—and in that way, make more money for the company," he said.

Baker noted that companies are often built "on revenue streams that come from imprecise data methods that are often wrong." The company may or may not need to decide whether to "move from what works to truth." It may not be worth trying to do so. This leads Baker to wonder if "truth could be just something that we deal with in our spare time because it's not really part of the business model."

Bias-Free Interpretation of Big Data?

Andersen's point is part of a larger challenge for those interpreting Big Data: How can the numbers be interpreted accurately without unwittingly introducing bias? As a large mass of raw information, Big Data is not self-explanatory. And yet the specific methodologies for interpreting the data are open to all sorts of philosophical debate. Can the data represent an "objective truth" or is any interpretation necessarily biased by some subjective filter or the way that data is "cleaned?"

"Cleaning the data"—i.e., deciding which attributes and variables matter and which can be ignored—is a dicey proposition, said Jesper Andersen, because "it removes the objectivity from the data itself. It's a very opinionated process of deciding what variables matter. People have this notion that you can have an agnostic method of running over data, but the truth is that the moment you touch the data, you've spoiled it. For any operation, you have destroyed that objective basis for it."

"'Bad data' is good for you."

Jeff Jonas

The problems of having an objective interpretation of data are made worse when the information comes from disparate sources. "Every one of those sources is error-prone, and there are assumptions that you can safely match up two pieces together. So I think we are just magnifying that problem [when we combine multiple data sets]. There are a lot of things we can do to correct such problems, but all of them are hypothesis-driven."

Responding to Andersen, Jeff Jonas of the IBM Software Group believes that "'bad data' is good for you. You *want* to see that natural variability. You *want* to support dissent and disagreement in the numbers. There is no such thing as a single version of truth. And as you assemble and correlate data, you have to let new observations change your mind about earlier assertions."

Jonas warned that there is a "zone" of fallibility in data, a "fuzzy line" between actual errors and what people choose to hear. For example, he said, "My brother's name is 'Rody' and people often record this as 'Rudy' instead. In this little zone, you can't do peer review and you can't read everybody's mind. And so to protect yourself, you need to keep natural variability and know where every piece of data comes from—and then

allow yourself to have a complete change of mind about what you think is true, based on the presence of new observations."

Or as Bill Stensrud of InstantEncore put it, "One man's noise is another man's data."

Is More Actually Less?

One of the most persistent, unresolved questions is whether Big Data truly yields new insights—or whether it simply sows more confusion and false confidence. Is more actually less?

Perhaps Big Data is a tempting seduction best avoided, suggested Stefaan Verhulst, Chief of Research at the Markle Foundation. Perhaps "less is more" in many instances, he argued, because "more data collection doesn't mean more knowledge. It actually means much more confusion, false positives and so on. The challenge is for data holders to become more constrained in what they collect." Big Data is driven more by storage capabilities than by superior ways to ascertain useful knowledge, he noted.

"One man's noise is another man's data."

Bill Stensrud

"The real challenge is to understand what kind of data points you need in order to form a theory or make decisions," said Verhulst. He recommends an "information audit" as a way to make more intelligent choices. "People quite often fail to understand the data points that they actually need, and so they just collect everything or just embrace Big Data. In many cases, less is actually more—if data holders can find a way to know what they need to know or what data points they need to have."

Hal Varian, Chief Economist at Google, pointed out that small samples of large data sets can be entirely reliable proxies for the Big Data. "At Google, we have a system to look at all the data. You can run a day's worth of data in about half an hour. I said, no, that's not really necessary. And so the engineers take one-third of a percent of the daily data as a sample, and calculate all the aggregate statistics off my representative sample."

"I mean, the reason that you've got this Big Data is you want to be able to pick a random sample from it and be able to analyze it. Generally,

you'll get just as good a result from the random sample as from looking at everything—but the trick is making sure that it's *really* a random sample that is representative. If you're trying to predict the weather in New England from looking at the weather patterns in California, you'll have a problem. That's why you need the whole system. You're not going to need every molecule in that system; you might be able to deal with every weather station, or some degree of aggregation that's going to make the analysis a lot easier."

Bill Stensrud took issue with this approach as a general rule: "If you know what questions you're asking of the data, you may be able to work with a 2 percent sample of the whole data set. But if you don't know what questions you're asking, reducing it down to 2 percent means that you discard all the noise that could be important information. What you really want to be doing is looking at the whole data set in ways that tell you things and answers questions that you're not asking."

"The more people you have playing with the data, the more people are going to do useful things with it."

Kim Taipale

Abundance of data in a time of open networks does have one significant virtue—it enables more people to crunch the same numbers and come up with their own novel interpretations. "The more people you have playing with the data, the more people are going to do useful things with it," argued Kim Taipale.

The paradox of Big Data may be that it takes more data to discover a narrow sliver of information. "Sometimes you have to use *more* to find *less*," said Jeff Jonas of IBM Software Group. "I do work helping governments find criminals within. You really don't want to stare at less data. You want to use more data to find the needle in the haystack, which is really hard to find without a lot of triangulation. But at some point, less becomes more because all you are interested in doing is to prune the data, so that you can stare at the 'less.'"

Esther Dyson, Chairman of EDventure Holdings, believes that sifting through "more" to distill a more meaningful "less" represents a huge market opportunity in the future. "There is a huge business for third parties in providing information back to consumers in a form that is meaningful," she said. One example is a company called Skydeck,

which helps you identify your cell phone calling patterns, based on the data that your phone company provides on your behalf.

…sifting through "more" to distill a more meaningful "less" represents a huge market opportunity….

Esther Dyson

The lesson of Big Data may be "the more abundance, the more need for mediation," said Stefaan Verhulst. There is a need for a "new mediating ecosystem."

John Liechty, Associate Professor of Marketing and Statistics at Pennsylvania State University, agreed: "It really comes down to what tools we have to deal with Big Data. We're trying to get to a system where you can begin to extract meaning from automated systems…. Less is more only if we are able to reduce large sets of data down, and find ways to think about the data and make decisions with it. Ultimately, you have to have some extraction of the data in order to deal with it as a human being."

Correlations, Causality and Strategic Decision-making

The existence of Big Data intensifies the search for interesting correlations. But correlation, as any first-year statistics student learns, does not establish causality. Causality requires models and theories—and even they have distinct limits in predicting the future. So it is one thing to establish significant correlations, and still another to make the leap from correlations to causal attributes. As Bill Stensrud put it, "When you get these enormously complex problems, I'm not sure how effective classic causal science ends up being. That's because the data sets are so large and because it is difficult to establish causality because of the scale of the problem."

That said, there are many circumstances in which correlations by themselves are eminently useful. Professor Lise Getoor of the University of Maryland pointed out that for tasks like collaborative filtering, group recommendations and personalization, "correlations are actually enough to do interesting things."

For Sense Networks, Inc., which evaluates geo-location data for mobile phone providers, establishing correlations is the primary task. "We analyze really large data sets of location data from mobile

phones and carriers and handset manufacturers," said Greg Skibiski, Co-Founder and Chief Executive Officer of the company. "So we see tens of millions of people moving around, and really all we care about, in the end, is the correlations. The problem is, we have to make some really core theory decisions at the very, very beginning of [analyzing] these data sets."

For example, said Skibiski, how should analysts define "place?" Is place defined by the amount of time that people spend there, or by the number of visits that they make daily, weekly or monthly? Or does one try to establish a more subjective, idiosyncratic definition?

In the end, Skibiski said the size of a database tends to resolve such definitional quibbles: "If the 'lift curves' look good and the false-negatives and false-positives match up, that's the end of the story for us." However the techniques are refined, it is clear that they are enabling new sorts of inferences to emerge. A recent M.I.T. study found that geo-location data patterns can successfully predict people's future locations and social interactions.[7]

Correlations can be functional and useful in stimulating sales and making money, but they can also be highly imperfect. "If I order books for myself and my wife through Amazon," said John Seely Brown, "there are two different sets of data to be looked at, so sometimes Amazon will have to decide that, well, maybe there are two sets of buyers here." Bill Stensrud agreed, "Everything I buy from Amazon is a present for somebody else, and so their recommendation engine is meaningless to me. Some day they'll figure that out."

"Google doesn't know how many Jeff Jonas's there are," said Jeff Jonas of IBM Software Systems. "If you can't do correlations at atomic level construction [counting discrete identifiable units], you can't really do any form of meaningful predictions because you can't get trajectory or velocity [from the data]."

Even though correlations are inherently limited as predictors, they can be useful in many different ways. Some new businesses are based on sharing real-time correlations of data. City Sense tabulates the most active night life spots in a city, based on mobile phone and taxi traffic, giving subscribers a near real-time idea of "where the action is" at that very moment. Rapleaf, a San Francisco company, sifts through its massive pile of data from social networking websites to make suggestive

correlations, such as the association between one's friends and credit risk, reports Greg Skibiski. Even if your personal financial indicators give you a credit risk score of 550, but your friends have scores of 650, then your actual credit risk may well be closer to 650, according to some data analysts.

Data correlations are also useful in provoking interpretive stories, said John Seely Brown. "The quality of the story really matters in making sense of the data. You start to see stories clash against each other and get passed around, so that they may actually generate new insights." Data correlations can provoke people to develop a "new ecology of stories," which itself may shed light on the numbers.

For Joi Ito, the Chief Executive Officer of Creative Commons, the search for correlations is a trap to be avoided, at least in his capacity of a computer security expert and a venture capitalist. Ito says he is "always looking for unpredictable things that you can use opportunistically." As a venture capitalist, he is looking for the "subversive outlier" whose ideas could have a big upside. From a security perspective, Ito says he wants to be alert to the *unexpected* forms of intrusion and deceit, not to the ones whose correlations can be easily discovered using computers.

"I'm always very aggressively going outside of my comfort zone and looking for that tiny little data point that the statisticians won't see," said Ito. "Then I amplify it like crazy so that I can make as much money as quickly as possible. When you do that kind of analysis on, say, terrorist networks, you have to understand that Hezbollah is actively trying to continuously come up with patterns that they think you won't predict."

"Remember," said Ito, "the same technology that we're using to analyze Big Data enables these other actors to become more actively random. The people who are outliers, who used to sort of behave randomly, now have access to the same tools as the rest of us and are looking at the same data.

"People like me don't even look at the data. We go randomly to places that are completely absent of any data and we test and then we jump. That's why I'm in the Middle East—because it's completely random. [Ito recently moved to Abu Dhabi.] It's a big hole in which you can mess around and maybe find something. If you do find something,

then you start creating your own patterns and hook them back into the [mainstream]. *Then* you create this huge arbitrage. But the way that I do it is completely non-analytical. The more analytical you become, the more likely you're going to end up bumping into somebody who is already there. So it's much better to be completely random."

…linear extrapolations are "a little bit like saying, in 1950, 'What's the business model for big band music?'"

Jordan Greenhall

Ito's *modus operandi* may be especially well-suited to our times, in which data trends are not necessarily linear. As Jordan Greenhall explained, linear extrapolations are "a little bit like saying, in 1950, 'What's the business model for big band music?'" The human brain has a relatively high elasticity, he said, and the different experiences and technologies of today's generation mean that its brain neurology actually differs from that of previous generational cohorts. "As a result," said Greenhall, "we can't extrapolate from our own expectations of what makes sense to us, to what may make sense to a younger generation. So any decision made today has to plan forward ten years to make sure that it makes sense in the future reality."

To Greenhall, the pace of cultural (if not neurological) change is so great that we must recognize "the non-linearity of the space that we are dealing with." Simple correlations will be very crude tools—sensible by the terms of a more stable, classical worldview, but more problematic in their capacity to limn the future.

"Big Data is about exactly *right now*, with no historical context that is predictive," said Ito. "It's predictive of a linear thing—but you can use data collection to discover non-linearity as well. For example, I look on Twitter every day for any word that I don't know. Then I search Twitter for a whole feed and go to all the links. I always discover a new trend every morning that is completely non-linear from anything that I expected—because if I expect it, I just gloss over it…. It's important not to be obsessed with the old models that come from the old data. It's more important to be ignorant enough to come up with a new model of the future."

Business and Social Implications of Big Data

However one may quarrel about interpretive methodologies, there is little question that Big Data can help identify emerging trends, improve business decision making and develop new revenue-making strategies. Hal Varian, Chief Economist at Google, says that the growth of "computer-mediated transactions"—in which a computer sits between every buyer and seller—means that companies can "do many, many extra things."

"We have a lot of tools to look at data," said Varian. "Our basic operating procedure is to come up with an idea, build a simulation, and then go out and do the experimentation. At any given moment, Google is running hundreds and hundreds of experiments on both the search side [of data] and the ad side. We use a lot of different variables, and trade them off against others, sometimes explicitly and sometimes implicitly. If you're getting a 1 percent or 2 percent lift every two or three weeks, then after a few years you can build up an advantage."

Many innovative uses of Big Data could be called "now-casting," said Varian. This term refers to the use of real-time data to describe contemporaneous activities *before* official data sources are available. "We've got a real-time variable, Google search queries, which are pretty much continuous," said Varian. "Even if all you've got is a contemporaneous correlation, you've still got a six-week lead on the reported values" for certain types of data.

One of the most noted examples of now-casting is a service known as Google Flu Trends. By tracking the incidence of flu-related search terms, this Google spinoff service can identify possible flu outbreaks one to two weeks earlier than official health reports. When the Google data are correlated with actual flu cases compiled by the Centers for Disease Control, the Google estimates are 97 percent to 98 percent accurate.[8]

Varian noted that Google search queries for jobs and welfare can also indicate future economic trends. When first-time claims for unemployment benefits drop, for example, it has historically signaled the end of a recession. Google data can reveal such trends a week or two earlier than official government statistics. In fact, Varian has made the rounds in Washington "to make the case that government agencies should use Google tools to better draw current snapshots of consumer sentiment, corporate health and social interests," according to *The Washington Post*.[9]

As Varian told the Aspen conference participants, "In about nine months, Fed Chairman Ben Bernanke is going to have to decide whether to raise interest rates. He will look at a whole lot of variables—the last month's economic reports, retail sales data, initial claims from unemployment, and so on. To the extent that this data is more up-to-date, you could potentially make a better decision about whether to move on one issue or another."

"To make money, you've got to predict two things—what's going to happen and what people *think* is going to happen."

Hal Varian

American Express has used its sizeable database of consumer behavior to identify early trends and craft appropriate responses. For example, Amex has found that people who run up large bills on their American Express card and then register a new forwarding address in Florida have a greater likelihood to declare bankruptcy. That is because Florida has one of the most liberal bankruptcy laws, which makes it a favorite destination for debtors who are financially troubled. Identifying such correlations in the data—a soaring credit card balance and a re-location to Florida—can trigger an inquiry into the actual solvency of the cardholder.

There are many types of real-time data streams that can now be assembled and analyzed. Besides search engine queries, data for credit card purchases, the trucking and shipping of packages, and mobile telephone usage are all useful bodies of information. Much of this data is becoming available on a near real-time basis, which leads Varian to predict that credit card data will be compiled and sold on a daily basis at some point in the future. "Real-time economic indicators" will be possible, he said. "The hope is that as you take the economic pulse in real time, you'll be able to respond to anomalies more quickly."

By revealing a new genre of ultra-timely information, now-casting enables new types of arbitrage. If a company or investor can use real-time data to out-perform the market by just a few percentage points, it can reap that much more revenue in the marketplace. "The problem is not whether your predictions are more accurate, it's whether they beat the consensus," said Varian. "To make money, you've got to predict two things—what's going to happen and what people *think* is going to happen. You only make money by beating that spread."

"Playing the percentages" can be especially important in advertising and marketing, several participants noted. As more consumers migrate from traditional mass media to online media, questions arise. How should marketing budgets be allocated in this marketing landscape? What has greater influence on consumer purchases—public relations or advertising?

Aedhmar Hynes, Chief Executive Officer of Text 100 Public Relations, reported that "for high-impact brands" in the United States—that is, brands where the purchase is a large expenditure that consumers may research or ponder—"the impact of public relations on the brand was 27 percent, whereas the impact of advertising on such purchases was less than 1 percent. The reverse was true for low-impact buying decisions. If you're going to buy a piece of gum, the likelihood is that advertising will influence that decision far more than if you're going to buy a notebook computer."

Hynes speculated that data-analysis might be especially influential in making more effective media buys. But she also wondered if the interpretations would be useful in non-U.S. countries where the use of computers and search engines is less pervasive than in the U.S.

Certainly one new marketing frontier that Big Data will enable is the use of real-time data correlations to drive business decisions. Jacques Bughin of McKinsey & Company reported that his firm had discovered "that the time frame between search and buy has been reduced in some market segments." This suggests a greater opportunity to influence potential buyers.

It is also possible to discover some non-intuitive correlations in consumer buying patterns, such as the fact that people who do search-engine queries for "weddings" also tend to do searches for "diets." (An apocryphal correlation asserts that people who search for "diapers" also search for "six packs of beer," a hypothetical correlation made in a speech that later ripened into an urban legend.)

Some correlations are entirely verified, yet extremely difficult to interpret. What are we to make of the fact that in the three days preceding the bankruptcy of Bear Stearns, the nightlife patterns in seven cities—people staying out late in restaurants and bars—was a "five Sigma event," according to Sense Networks, Inc., meaning a significant deviation from the statistical mean. "A lot of people were out extremely late on those evenings, like you've never seen in years of data," said Greg Skibiski.

According to Jacques Bughin of McKinsey and Company, the real insight is that Big Data has the potential to discover new laws of macrobehaviors totally overlooked with the paucity of data of the past. Although social influence is and remains large, a new type of market power behavior has emerged, one that is not necessarily firm driven, but consumer driven. Today, relatively few consumers are able to influence others via social media and other interactive platforms. Businesses that are able to target or link to those influencers will have an edge, says Bughin. For instance, more and more firms using this new power curve are opening their business systems to users and suppliers to co-create products and services, or they are leveraging their brand and platforms for third parties (think Apple with the iphone). The more companies that master the skills for open collaboration with users—and successfully deliver—the higher the probability to leverage the influencers to their benefit.

Social Perils Posed by Big Data

As Big Data becomes a more common tool in corporate decisions, a number of new social perils arise. The most obvious is the risk of privacy violations. "Is personalization something that is done *to* you or *for* you?" wondered Kim Taipale of the Center for Advanced Studies in Science and Technology. A business with economic motives is driving the process of data-driven personalization, but consumers have far less knowledge of what is going on, and have far less ability to respond. The benefits of personalization tend to accrue to businesses but the harms are inflicted on dispersed and unorganized individuals, Taipale noted.

Marc Rotenberg, Executive Director of the Electronic Privacy Information Center, admits that there are two sides to personalization. When Amazon and iTunes use their databases of consumer purchases to make recommendations to prospective customers, most people welcome the advice. It may help them identify just the book or music that they want. On the other hand, "people start getting very uneasy when buying suggestions are made based on how much we know about this particular person, a practice that takes us into the realm of behavioral targeting"—the "my TiVO thinks I'm gay" phenomenon.

One independent survey of adult Internet users—by two professors at the University of Pennsylvania and the University of California, Berkeley, in September 2009—found that two-thirds of users object

to online tracking by advertisers. Respondents particularly disliked behavioral advertising, in which commercial websites tailor ads based on an individual's Web behavior. "I do think we're at the cusp of a new era, and the kinds of information that companies share and have today is nothing like we'll see ten years from now," said Professor Joseph Turow, the lead author of the study. "The most important thing is to bring the public into the picture, which is not going on right now."[10]

Citizens are also legitimately worried about Internet service providers (ISPs) who may use "deep packet inspection" techniques to analyze the data flowing through their wires, to determine what websites you may be visiting and what purchases you may be making. "In recent years," said Rotenberg, "ISPs have recognized that there is commercial value in their networks, beyond any security issues. Some of the same tools that can be used to identify spam can be used to figure out who's interested in buying a new SUV or who's planning on traveling." One solution might be to allow ISPs to use deep packet inspection for assuring the security of their networks, but to prohibit use of the data for commercial purposes, he said.

"Vendors are using Big Data to try to acquire the consumer," said Bill Stensrud, Chairman and Chief Executive Officer of InstantEncore, "and they are doing that by using technologies that are beyond the reach of the consumer by orders of magnitude." He noted that three responses have been suggested—counter-responses by hackers to harass companies that violate their privacy; schemes to "monetize" people's private data so that they can control it and sell it; and government regulation to protect individual privacy.

"None of these answers leave me very satisfied," said Stensrud. "I guess one of the questions that I have is how does the consumer get armed to fight on an even ground with the big companies who can make nano-second stock market trades and personalize their marketing?"

Joi Ito of Creative Commons took issue with the whole framing of the discussion as one between vendors and consumers; there are non-market actors who are influential as well. "We don't use the word 'consumer' in our group [Creative Commons], which acts collectively in the way that people in the hacker, open source and Wikipedia worlds do. We believe that there are ways for us to take control, so that your business models don't matter. *We're* in charge."

Ito cited the hundreds of thousands of fake Twitter accounts that hackers created in the course of a few hours after Ashton Kutcher and CNN announced their hopes of being the first to amass one million Twitter followers. A website of disruptive hackers, 4chan, quickly gamed the system to amass a huge following for a notorious criminal. For Ito, attacks by "smart mobs" demonstrate the ability of non-commercial actors to influence trends—"a power that is getting stronger and stronger and stronger," he said.

Big Data and Health Care

As researchers apply extreme inference techniques to public health, disease research, drug research, genetic engineering, and much else, the implications are both heartening and frightening. Identifying new correlations in data can improve the ways to develop drugs, administer medical treatments and design government programs. But it can also introduce new frustrations and complexities because solutions must overcome existing incentive systems for physicians, insurers and patients.

Patients are beginning to take charge of their own health care by doing research about their injuries and illnesses and joining social networks….

Stefaan Verhulst, the Chief of Research at the Markle Foundation, gave an overview of the health care trends that involve Big Data. The first, most significant is the American Recovery and Reinvestment Act, the so-called ARRA, which is President Obama's stimulus package for dealing with the financial crisis. Approximately $19 billion of that stimulus legislation is earmarked to encourage physicians to adopt electronic medical recordkeeping systems. (Some people argue that, if additional funds managed by the Department of Health and Human Services and other sources are included, some $26 billion is being directed at this issue.)

This law is potentially significant because the American health care system is plagued by a fragmented, inefficient system of paper-based recordkeeping. Digitizing records could make health care recordkeeping vastly more efficient and versatile, especially in assembling large pools of data and evaluating them for new insights.

Currently, about 15 percent of office-based physicians have electronic medical records systems, said Verhulst. He believes that the ARRA "will transform the way that patient information is stored and shared. Many people hope that we will finally have some level of 'health information liquidity' that will let physicians and providers share health information across jurisdictions."

The other big effect of digital technologies has been called "Health 2.0," or "participatory health care." Patients are beginning to take charge of their own health care by doing research about their injuries and illnesses and joining social networks on which they can exchange information and provide support to each other. A survey by the Pew Charitable Trust, "the Social Life of Health Information," found that 61 percent of American adults look for health information online. Yahoo reports that there are some 2,000 Yahoo groups that focus on health-related issues, and Google finds that an impressive number of search queries are for health-related topics.

Verhulst believes that Big Data could be useful in improving health in two significant ways—population health and personalized health care. In terms of public health, the Centers for Disease Control are eager to improve their "syndromic surveillance," their ability to monitor the spread of certain diseases such as the flu. Its revised BioSense program tries to tap into patient information from hospitals and emergency rooms in a more efficient, distributed way. In a less rigorous but more timely way, Google Flu helps identify the diffusion of contagious diseases as well. Other initiatives such as Health Map try to amass geographic portraits of disease incidence.

Big Data is also helping consumers acquire more reliable and timely information about the cost and quality of health care, said Verhulst. *Consumer Reports* magazine recently started a ratings system that tries to rate medical care according to specific indicators. Studies of "comparative effectiveness" are also attempting to use Big Data to reach better conclusions about patient outcomes. What works best and what treatments are cheapest? While there is a strong consensus that comparative effectiveness studies are worth doing, there is great disagreement about how they should be designed. Should the system be centralized or distributed? What should be the priority research areas?

Yet another area of population health where new pools of health data will be helpful is in drug research. Clinical drug research often relies upon surprisingly small sets of data, especially after the drugs are

introduced to the marketplace. In response, the FDA recently initiated the Sentinel Initiative, a system to improve post-marketing drug surveillance. The envisioned system will be a distributed data network, in which participating organizations will each maintain control of their data, but share them via standardized formats and computer programs to build a network-wide database.[11] Using more closed, proprietary sets of data, companies like Patients Like Me and 23andme provide drug companies with information about patient populations, their health conditions and the specific drugs that people are using.

Finally, data is an important tool in developing new types of personalized health care. Companies like 23andme and Navigenics can provide highly detailed genetic analysis of a person's genome—information that may be helpful in customizing certain types of medical care. There is also a lot of business activity in managing personal health records, including Google Health, and Microsoft HealthVault.

Esther Dyson, an investor in technology- and health-related businesses, believes that the "grassroots Internet" may do for health what it is already doing for politics—empower individuals by making information more accessible and transparent. "The Internet is changing people's expectations of what they have a right to know and say," Dyson wrote in the *Financial Times*. "Just as they expect to know more about their politicians, they expect to know more about their own health institutions—and to criticize them publicly. Websites let people rate their own doctors and hospitals, even as public pressure and occasionally public rules demand more and more transparency about performance and outcomes."[12]

For all the promise of improving health care through data, the surprising reality is that Big Data does not really exist in health care settings. Researchers in life sciences have significant bodies of data, but health data about populations and individuals are far more limited. "Most of the applications of great data computing are actually in the life sciences and biology-based areas," said Verhulst of the Markle Foundation. This is largely because patient data records cannot be so easily collected and shared; there are all sorts of technical, ethical and public policy barriers to making such data "liquid."

And yet patient data—if properly compiled and analyzed with consent from the individuals involved—could yield many useful insights.

"Some 97 percent of people who have cancer in the United States are not part of any clinical trials," said Joi Ito of Creative Commons. "Just providing some basic tools like websites could provide you with tons of data." Health information could also be used to try to improve people's health-related behaviors, said Stefaan Verhulst. "Some 75 percent of health care costs are related to chronic diseases, and most of these diseases have important behavioral aspects."

Big Data as a Disruptive Force (Which is therefore Resisted)

Notwithstanding the enormous potential benefits of using data to improve the health care system and individual health, intelligent uses of Big Data are frequently resisted. Why? Because health insurers, pharmaceutical companies and patients often believe that their self-interests will be harmed by the collection and use of data. Examples of this include:

…health insurers, pharmaceutical companies and patients often believe that their self-interests will be harmed by the collection and use of data.

- Pharmaceutical companies are not eager to do more aggressive post-marketing surveillance of how their drugs work lest they discover "adverse events" that might trigger legal liability or depress sales.

- Physicians are traditionally rewarded by *frequent* visits by patients (which generate more revenue for them). Preventive care or better health outcomes are not necessarily remunerative.

- Most oncologists decline to participate in clinical studies because, according to *The New York Times*, "They make little or nothing on trials and, in fact, often lose money. These doctors also may discourage patients from going elsewhere to enter a trial: if a patient leaves, the doctor loses business."[13]

- Consumers could benefit from comparative ratings of doctors or hospitals, which *Consumer Reports* and other organizations are starting to compile, but doctors and hospitals generally do not welcome such ratings.

- Insurers and health care providers might be able to provide more tailored and effective care if they knew the genetic background of a person, but many patients worry that such disclosures could result in discriminatory treatment or a termination of their insurance.

Any attempts to use Big Data to improve health care will have to grapple with the self-interests of different players involved. "If having more information about me enables me to live longer and be healthier," said Marc Rotenberg of the Electronic Privacy Information Center, "then I don't think I'm going to object. If having more information about me means my insurance rates go up, or that I'm excluded from coverage, or other things that might adversely affect me, then I may have a good reason to be concerned."

Many doctors, said Stefaan Verhulst, "are freaking out" about the various rating systems that are being developed to judge their medical treatment. So some are now requiring prospective patients to sign a "no complaints contract" as a condition of receiving treatment.

John Seely Brown, Independent Co-Chair of the Deloitte Center for the Edge, stated the dilemma succinctly: "In some sense, in order to get to Big Data, we are going to have to decide whether we will be able to find market mechanisms to do this, or whether we will have to have the government assist it, by passing legislation."

A pivotal question, according to Stefaan Verhulst of the Markle Foundation, is whether data sets should be regarded as a market resource to be exploited for private, proprietary gain or whether they should be regarded as a public good.

Recent Attempts to Leverage Big Data

Notwithstanding some resistance, there are many important efforts afoot to leverage data to improve care. One of the most important, as previously mentioned, is the Health 2.0 or the participatory health care movement. Patient-driven websites and advocates for access to data are a growing force. For example, HealthDataRights.org has issued "A Declaration of Health Data Rights" that asserts that people should:

1. Have the right to our own health data.

2. Have the right to know the source of each health data element.

3. Have the right to take possession of a complete copy of our individual health data, without delay, at minimal or no cost; if data exist in computable form, they must be made available in that form.

4. Have the right to share our health data with others as we see fit.

These principles express basic human rights as well as essential elements of health care that is participatory, appropriate and in the interests of each patient. *No law or policy should abridge these rights.*

Another significant force for opening up data is the Science Commons, an offshoot of Creative Commons, which is dedicated to addressing the barriers to sharing information in scientific contexts. One Science Commons project is dedicated to tackling the barriers of technical standards, copyright and institutional rules that prevent different databases from sharing their information. This is particularly important in the life sciences where innovation is being stymied by an inability to mix different data sets.

"The transaction costs of having to negotiate a contract to put two databases together usually exceeds the value of that transaction," said Joi Ito of Creative Commons. It is not enough to make databases "open," he said, because you cannot just mix proprietary information with "open" information, and hope to allow that information to be shared in turn.

Ito explained that Science Commons is trying to develop a "legal layer of interoperability for data" that would function much as the TCP/IP protocols for the Internet enable different computer networks to communicate with each other. Science Commons is currently building a "knowledge system" known as Neurocommons that attempts to connect all sorts of scientific ontologies, databases and other systems into an open-source platform. When researchers can share neurological research data easily, it will accelerate discovery and innovation.

Protecting Medical Privacy

Of course, one person's "data liquidity" is another person's privacy violation. While scientists may need easier sharing of data, most people do not want their medical data to flow too easily and without controls. One way to deal with this understandable concern is for data holders to "de-identify" and then "re-identify" the data. Personal privacy can be protected by stripping away personal identifiers in the data, and then aggregate it without such markers.

Marc Rotenberg of EPIC is skeptical about the actual efficacy of de-identification techniques, however. He cited a case involving the State of New Hampshire and its system for protecting the privacy of doctors' drug prescribing practices and patients' records. "Even though we were told that there was a de-identification system in place," said Rotenberg, "it didn't work. In theory, I think de-identification is an excellent approach, and it is one of the things that we [EPIC] continue to propose because it is one way to reconcile the public benefit while minimizing private harm. But it has to work."

"…when they go and correct something, they actually don't know who to tell to correct it downstream."

Jeff Jonas

De-identification may be problematic precisely because of the size of the databases. Big Data offers greater opportunities for "re-identifying" the data, i.e., linking a given set of medical information to a specific person.

Unauthorized "secondary use" of data—i.e., the re-use of data for purposes that a patient did not originally authorize—is a related privacy problem that will require public policy intervention. Stefaan Verhulst said that secondary use of data is not addressed by any data policy regimes right now. Yet another problem is "dirty data"—data whose integrity and reliability are dubious because of sloppy practices and protocols.

Jeff Jonas of IBM Software Group noted, that "There is no outbound record-level accountability—organizations transfer data out and they don't know where they sent it. That means when they go and correct something, they actually don't know who to tell to correct it downstream. That's how things work in most every information-sharing system.

Recipients know where they get their records, for the most part, but the issuer doesn't. So they can't keep the ecosystem of data current."

One partial solution to the problems information sharing and data protection, suggested Jonas, is a process that he calls "analytics in the anonymized data space." By this, he means that data holders who intend to share data with another party must deliberately "anonymize" the data first, before sharing it. Then analytics are performed on the data while it remains in the anonymized form. Such analytics are emerging and in some settings they are producing materially similar results as analytics performed on clear text.

Another approach that could work for personal health data, said Esther Dyson, is to develop schemes that can "securitize" people's health. "The challenge in health insurance from the insurers' point of view is that it's not really worth it to pay for prevention because some other insurance company is going to reap the benefits later on," she said. So if a person's health were assigned a monetary value, much as financial instruments "securitize" intangible risks, then insurance companies would have a financial motive to provide preventive care. The financial value of a person's health would be reflected in a schedule of health care fees; if the actual cost of medical care were *less* than the predicted cost, then the insurer would make money after providing preventive care.

Stephen Baker, the *Business Week* reporter and author of *The Numerati,* likens such a scheme to an unusual auto insurance scheme offered by Norwich Union in Great Britain: "The company puts a black box in your car and then monitors your driving patterns, and offers you a rate based on your actual driving behavior. Now, that data belongs to the company. They control it and it's up to them to figure out how it works for them. But imagine a system in which they gave you back your driving data, and you could ask companies to offer you bids for the best insurance deal. That scheme resembles Dyson's 'securitization' of people's health. Not really… though there are some parallels. The point is that you have to work with risk-adjusted groups, not single individuals, since so much of health outcomes depends on the person's initial condition rather than his behavior."

Two objections were raised to the idea, however. How would it benefit people with adverse health conditions (who would presumably have to pay more)? As noted, the government would subsidize—"risk-

adjust" for—those people in poor health and likely to cost more. And do people truly wish to take control of their "data identities" in this way? The problem of a plentitude of choice and a paucity of time or information to make good choices, could be crippling for some people.

Notwithstanding the considerable cost of deploying electronic health records in the U.S.—in the range of $100 billion—interest in doing so is likely to accelerate because of the insights that may be gleaned from Big Data. "In many industries, we collect a lot of data, and just haven't learned how to analyze and use it," said Michael Chui of McKinsey & Co. "In U.S. health care, arguably, we don't collect Big Data at all!"

How Should Big Data Abuses be Addressed?

The rise of large pools of databases that interact with each other clearly elevates the potential for privacy violations, identity theft, civil security and consumer manipulation. A significant portion of the conference therefore dealt with how public policy ought to respond—or how other schemes might deal effectively with these problems.

Marc Rotenberg of the Electronic Privacy Information Center noted that misuses and abuses of data are not a new problem. He cited the U.S. government's reliance on census data to help identify American citizens of Japanese ancestry, so that they could be confined in interment camps during World War II. During George W. Bush's administration, Admiral Poindexter sought to institute the Total Information Awareness program, which would have amassed unprecedented amounts of data from diverse sources. The goal was to analyze data patterns in an attempt to predict future terrorist and criminal activity.

The mindset behind the Total Information Awareness program—that Big Data can yield meaningful and predictive insights that protect our civil order and national security—is surely useful in certain respects. Yet many people fear that that mindset can lead us closer to dystopian scenarios. The touchstone for such fears is Phillip Dick's book *Minority Report* (later a movie by Stephen Spielberg) in which the government uses elaborate computer analyses to identify and arrest "criminals" before they are able to commit a crime.

The problem is that Big Data enables authorities to make inferences that amount to "probabilistic cause," but U.S. law currently requires a judge's finding of "probable cause" before a search or seizure may be conducted.

Data may provide suggestive *statistical* evidence that certain events may occur in the future—evidence that may, upon further investigation, meet the legal standards for finding probable cause. But probabilistic cause remains a less reliable and more abstract predictive standard.

That is because the premises of suspicion are not necessarily discernible when databases, using undisclosed algorithms, identify patterns of inference and assert probabilistic cause. "If you're going to make decisions about people—such as preventing them from boarding a plane or detaining them as a security risk—then there has to be some fact that someone will stand behind that provides the basis of the decision," said Rotenberg. Law enforcement authorities naturally would like to shift to the easier standard of probabilistic cause, he said. But this should require greater transparency, such as disclosing the computing algorithms and inferential reasoning that suggest a security risk.

Stephen Baker of *Business Week* warned that computer algorithms are not infallible just because they are computerized: "The prejudices of a society are reflected in the algorithms that are searched." When researching his book, *The Numerati*, he encountered a story about an FBI agent who supposedly correlated a neighborhood's consumption of hummus with the likelihood that it could be a haven for terrorists. Baker could not verify the story, and it may well be apocryphal, but he said the story nonetheless illustrates how stupid assumptions, once plugged into a database, can be taken seriously.

On the other hand, why should probabilistic cause not be a legitimate tool for persuading a judge that there is indeed probable cause for a search? "The idea that you can't use probabilistic data to get a probable cause standard is silly," said Kim Taipale, the Center for Advanced Studies in Science and Technology Founder and Executive Director. He added that "if you start from the premise that the data is going to exist and the data may be relevant for making a judgment that is important to society," then the goal should not be to ban the use of correlations and data analysis. The goal should be to monitor it properly. "We don't take guns away from policemen. We try to control abuse and misuses of them. To say that government can't use these tools to do legitimate things is silly."

Regulation, Contracts or Other Approaches?

There was a range of perspectives among conference participants about how exactly the misuse or abuse of Big Data should be addressed. The traditional approach is congressional statutes and regulation, which may well be necessary and effective in certain respects. But some participants argued that there are other approaches that need to be explored further because traditional regulation is ill-equipped to oversee electronic networks and large databases.

There was general consensus among participants about the importance of mandatory transparency. Just as credit-rating agencies such as Equifax must provide individuals with copies of their credit records upon request, so companies that hold people's personal data ought to make similar disclosures, said Esther Dyson. It's a standard of "transparency back to you," she said, which amounts to saying, "This is what we know about you."

"But where does that get you [in terms of preventing abuses]?" asked Charles Firestone, Executive Director of the Aspen Institute Communications and Society Program. Such disclosures may or may not advance larger policy principles or data-handling practices to protect privacy. Bill Strensrud of InstantEncore worried that "if there's an enormous number of companies collecting information on you, and they're all reporting back to you, you could be overwhelmed with too much information, so that you couldn't effectively do anything."

Dyson replied, "The basic principles should be that data collection is transparent and accessible in a meaningful way, rather than in a vague and unspecific way." John Clippinger, Co-Director of the Law Lab at the Harvard University Berkman Center for Internet and Society stressed that, while not everyone will pore through all their personal data, the point is to set down a set of architectural principles for how data will be handled and how that handling will be disclosed. There was agreement, too, that disclosure is not likely to be enough on its own. Public policies will be needed, and different standards will be needed for different types of data and circumstances.

Greg Skibiski of Sense Networks believes we need a "New Deal on Data."[14] By this, he means that the end users should be able to own their data and dictate how it is used. This should apply to "any data that we collect about you and metadata that we make out of it," he said.

He also urged that data should have a "lifespan," so that it is routinely purged after a given period of time. Otherwise, data that is saved is more likely to be abused.

But Kim Taipale has doubts that such an "ownership" standard would be practical given the scale and range of Big Data uses today. The point should be to empower users to have greater control over the access and use of their data.

One of the best ways to prevent abuses, said Stefaan Verhulst, is for companies to conduct "information audits" so that they only collect the data that they need in the first place. By asking better questions upfront, companies will decrease their legal vulnerability and the likelihood of privacy violations. But some participants objected that information audits clash with the very premise of Big Data, that "more is better." The very point of looking to Big Data, said Aedhmar Hynes of Text100 Public Relations, is "to identify patterns that create answers to questions you didn't even know to ask." So limiting data-collection in the first place could undercut the potential benefits that Big Data might deliver.

The very point of looking to Big Data is "to identify patterns that create answers to questions you didn't even know to ask."

Aedhmar Hynes

A number of conference participants had doubts that traditional regulatory structures—hierarchical, centralized, rule-driven—could adequately patrol the uses and abuses of databases. "We are sneaking all these antiquated [legal and regulatory] architectures into the future and trying to figure out how we make them work," said Jordan Greenhall, formerly CEO of DivX, "but in fact the future is not necessarily going to include the legacy structures. I think we need to start with thinking about the most appropriate social structures [for using data] and then figure out what the new model should look like."

John Clippinger agreed: "The idea of taking existing institutional structures, and simply pushing them into the future, won't work." He cited the Department of Defense's recognition that top-down strategies are not terribly effective in dealing with situations of distributed control and asymmetric warfare. "Instead of the omniscient eye, you're going to have to rely upon the edge for discipline and control."

Clippinger urged that breaches of data privacy and security—beyond the special cases of terrorism and other catastrophes—should be addressed through "new types of social ordering and social norms that are self-enforcing." Greenhall agreed: "You can't just try to reinvigorate some regulatory institution. You have to go back to brass tacks and build it back up on an entirely different basis—one that makes sense in light of the networked culture."

"Instead of the omniscient eye, you're going to have to rely upon the edge for discipline and control."

John Clippinger

There was disagreement about whether this was indeed possible or practical. Some regarded it as too visionary; others as inescapable given the trendlines of electronic culture. Still others believe that no serious reform will occur until there is a major crisis or data breach that causes economic harm.

A final disagreement centered on when regulators should step in. "The scale of the problems [with data abuse] are growing and the timeframe over when they operate are shrinking," said Bill Stensrud, "so what we did 100 years ago isn't very relevant to what will be needed in the future." The fact that "bad actors" always tend to out-pace regulatory controls suggests that regulation should try to anticipate new problems, and not wait for them to materialize.

Yet others argued that private law regimes—vendor contracts with other businesses and individuals, for example—would provide swifter responses to emerging abuses. For example, Hal Varian of Google noted that Google has an Ad Preferences Manager that enables people to opt in or opt out of various information-retention choices. "I think it is a kind of model for what will become an industry standard in this area, either through industry self-regulation or government regulation," said Varian.

Open Source Analytics for Financial Markets?

One of the more intriguing frontiers is creating large, publicly accessible databases to help identify problems in financial markets. Instead of relying on the mystical Invisible Hand of the market or the Heavy Hand of government regulation, perhaps large quantities of data, made

available in standardized formats to anyone, could serve as a way to protect consumers and investors involved in financial transactions.

In a 2009 article in *Wired* magazine, Daniel Roth explored how a glut of financial data is allowing problems to hide in plain sight: "Between 1996 and 2005 alone, the federal government issued more than 30 major rules requiring new financial disclosure protocols, and the data has piled up. The SEC's public document database, Edgar, now catalogs 200 gigabytes of filings each year—roughly 15 million pages of text—up from 35 gigabytes a decade ago."[15] Even regulators are choking on the data.

"Financial markets are at least as complicated and important as the weather, but we don't have the equivalent of a national weather service or a national hurricane center, for the financial markets."

John Liechty

The most promising solution is to make the data more flexible and useful, says Roth, "by requiring public companies and all financial firms to report more granular data online—and in real time, not just quarterly—uniformly tagged and exportable into any spreadsheet, database, widget or Web page."

One innovation in this regard is XBRL, a set of tags invented by accountant Charlie Hoffman to standardize financial information. The tags radically reduce the time it takes to audit financial data, and makes the data easier to many people to access and interpret. The SEC now requires companies with a market capitalization above $5 billion to use the format; all publicly traded companies and mutual funds will do so by 2011.

The hope and expectation is that "open source analytics" will allow many more people to start scrutinizing financial data. Just as the blogosphere has served as a fact-checker on the press and a source of new reporting and insights, so open-source financial data could yield red flags about corporate conduct and financial transactions.

John Liechty of Pennsylvania State University described his efforts to persuade Congress to authorize the creation of a National Institute

of Finance. The envisioned technical agency would provide regulators with new analytical capabilities to monitor and safeguard the financial system as a whole. The project is being pushed by a diverse set of academics, regulators and concerned industry professionals.

Liechty described the calamitous ignorance of federal regulators and financial officials when Lehman Brothers and AIG were failing in September 2008. "We really didn't know what was going on," said Liechty. "That's the point. The regulators didn't have any idea because they didn't have the right tools…. Financial markets are at least as complicated and important as the weather, but we don't have the equivalent of a national weather service or a national hurricane center, for the financial markets."

A website for the Committee to Establish the National Institute of Finance states the rationale for the new agency this way:

> While financial institutions already report a great deal of data to federal regulators, they don't report the types of data needed at the level of detail required that would enable a holistic assessment of firms' exposures to each other. More fundamentally, firms currently report data in a diversity of formats that are often mutually incompatible and require conversions that are difficult, expensive and error-prone. With no established and enforced standards in place, data from different sources cannot readily be linked, compared or analyzed in an integrated fashion. Consequently, it is currently impossible to create a comprehensive picture of the whole financial system that identifies the sources of potential instabilities.[16]

By streamlining the process by which the federal government collects financial data, standardizing their formats and performing holistic analyses—and enabling others to do so—the National Institute of Finance would help identify emerging systemic risks and run "stress tests" of financial institutions. Liechty and others are currently trying to incorporate the Institute into a pending package of regulatory reform ideas for the financial sector.

Conclusion

Big Data presents many exciting opportunities to improve modern society. There are incalculable opportunities to make scientific research more productive, and to accelerate discovery and innovation. People can use new tools to help improve their health and well-being, and medical care can be made more efficient and effective. Government, too, has a great stake in using large databases to improve the delivery of government services and to monitor for threats to national security.

Large databases also open up all sorts of new business opportunities. "Now-casting" is helping companies understand the real-time dynamics of certain areas of life—from the diffusion of diseases to consumer purchases to night-life activity—which will have many long-term reverberations on markets. New types of data-intermediaries are also likely to arise to help people make sense of an otherwise-bewildering flood of information. Indeed, data-intermediaries and interpreters could represent a burgeoning segment of the information technology sector in the years ahead.

But Big Data also presents many formidable challenges to government and citizens precisely because data technologies are becoming so pervasive, intrusive and difficult to understand. How shall society protect itself against those who would misuse or abuse large databases? What new regulatory systems, private-law innovations or social practices will be capable of controlling anti-social behaviors—and how should we even define what is socially and legally acceptable when the practices enabled by Big Data are so novel and often arcane?

These are some of the important open questions posed by the rise of Big Data. This report broaches some of the more salient issues that should be addressed. In the coming years, government, business, consumers and citizen groups will need to devote much greater attention to the economic, social and personal implications of large databases. One way or another, our society will need to take some innovative, imaginative leaps to ensure that database technologies and techniques are used effectively and responsibly.

Notes

1. Randal Bryant, Randy H. Katz and Edward D. Lazowska, "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society," December 2008, pp. 1-15, at http://www.cra.org/ccc/docs/init/Big_Data.pdf.

2. Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, June 23, 2008, at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

3. John Timmer, "Why the Cloud Cannot Obscure the Scientific Method," Ars Technica, June 25, 2008, at http://arstechnica.com/old/content/2008/06/why-the-cloud-cannot-obscure-the-scientific-method.ars.

4. Jeffrey Zaslow, "If TiVO Thinks You Are Gay, Here's How to Set It Straight," *Wall Street Journal*, November 26, 2002, p. 1, at http://online.wsj.com/article_email/SB1038261936872356908.html.

5. See http://en.wikipedia.org/wiki/Google_bomb.

6. Thomas H. Maugh II, "Cows Have Magnetic Sense, Google Earth Images Indicate," *Los Angeles Times*, August 26, 2008, at http://articles.latimes.com/2008/aug/26/science/sci-cows26.

7. N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255-268.

8. Alice Park, "Is Google Any Help in Tracking an Epidemic?" *Time* magazine, May 6, 2009, at http://www.time.com/time/health/article/0,8599,1895811,00.html.

9. Cecilia Kang, "Google Economist Sees Good Signs in Searchers," *The Washington Post*, September 12, 2009, at http://www.washingtonpost.com/wp-dyn/content/article/2009/09/11/AR2009091103771.html?hpid%3Dmoreheadlines&sub=AR.

10. Stephanie Clifford, "Tracked for Ads? Many Americans Say No Thanks," September 30, 2009, p. B3.

11. See, e.g., Richard Platt, M.D., et al., "The New Sentinel Network—Improving the Evidence of Medical-Product Safety," *New England Journal of Medicine*, August 13, 2009, p. 645-647.

12. Esther Dyson, "Health 2.0 could shock the system," *Financial Times*, August 12, 2009, p. 41.

13. Gina Kolata, "Forty Years War: Lack of Study Volunteers Hobbles Cancer Fight," *The New York Times*, August 3, 2009, p. A1, at http://www.nytimes.com/2009/08/03/health/research/03trials.html?scp=1&sq=lack+of+study+volunteers&st=nyt.

14. For more, see http://www.sensenetworks.com/press/wef_globalit.pdf.

15. Daniel Roth, "Road Map for Financial Recovery: Radical Transparency Now!" *Wired*, February 23, 2009, available at http://www.wired.com/techbiz/it/magazine/17-03/wp_reboot.

16. See http://www.ce-nif.org.

APPENDIX

The Eighteenth Annual Aspen Institute Roundtable on Information Technology

Extreme Inference: Implications of Data Intensive Advanced Correlation Techniques

Aspen, Colorado • August 4–7, 2009

Roundtable Participants

Jesper Andersen
Co-Founder
Freerisk

Stephen Baker
Senior Writer
BusinessWeek

David Bollier
Independent Journalist
and Consultant
Onthecommons.org

John Seely Brown
Independent Co-Chair
Deloitte Center for the Edge

Jacques Bughin
Director, Belgium
McKinsey & Company Inc.

Michael Chui
Senior Expert
McKinsey & Company, Inc.

John Clippinger
Co-Director, Law Lab
Berkman Center for Internet
and Society
Harvard Law School

William (Bill) T. Coleman III

Esther Dyson
Chairman
EDventure Holdings

Charles M. Firestone
Executive Director
Communications and Society
Program
The Aspen Institute

Note: Titles and affiliations are as of the date of the conference.

Lise Getoor
Associate Professor
Department of Computer Science
University of Maryland

Jordan Greenhall

Patrick W. Gross
Chairman
The Lovell Group

Aedhmar Hynes
Chief Executive Officer
Text100 Public Relations

Joi Ito
Chief Executive Officer
Creative Commons

Jeff Jonas
IBM Distinguished Engineer
and
Chief Scientist, Entity Analytic Solutions
IBM Software Group

John Liechty
Associate Professor,
Marketing and Statistics
Pennsylvania State University

Vijay Ravindran
Senior Vice President
and
Chief Digital Officer
The Washington Post Company

Marc Rotenberg
Executive Director
Electronic Privacy Information Center

Greg Skibiski
Chief Executive Officer
and Co-Founder
Sense Networks, Inc.

Bill Stensrud
Chairman and Chief Executive Officer
InstantEncore

Kim Taipale
Founder and Executive Director
The Center for Advanced Studies in Science and Technology

Hal Varian
Chief Economist
Google

Stefaan Verhulst
Chief of Research
John and Mary R. Markle Foundation

Fernanda B. Viégas
Research Scientist
Visual Communication Lab
IBM

Staff:

Kiahna Williams
Project Manager
Communications and Society Program
The Aspen Institute

Note: Titles and affiliations are as of the date of the conference.

About the Author

David Bollier (www.bollier.org) is an author, activist, blogger and consultant who has served as rapporteur for Aspen Institute Communications and Society conferences for more than 20 years.

Much of Bollier's work over the past ten years has been devoted to exploring the commons as a new paradigm of economics, politics and culture. He has pursued this work as an editor of Onthecommons.org, a leading website about commons-based policy and politics and in collaboration with various international and domestic partners.

Bollier's first book on the commons, *Silent Theft: The Private Plunder of Our Commons Wealth*, is a far-ranging survey of market enclosures of shared resources, from public lands and the airwaves to creativity and knowledge. *Brand Name Bullies: The Quest to Own and Control Culture* documents the vast expansion of copyright and trademark law over the past generation. Bollier's latest book, *Viral Spiral: How the Commoners Built a Digital Republic of Their Own*, describes the rise of free software, free culture, and the movements behind open business models, open science, open educational resources and new modes of Internet-enabled citizenship.

Since 1984, Bollier has worked with American television writer/ producer Norman Lear and served as Senior Fellow at the Norman Lear Center at the USC Annenberg School for Communication. Bollier is also co-founder and board member of Public Knowledge, a Washington policy advocacy organization dedicated to protecting the information commons. Bollier lives in Amherst, Massachusetts.

Previous Publications from the Aspen Institute Roundtable on Information Technology

Identity in the Age of Cloud Computing: The next-generation Internet's impact on business, governance and social-interaction (2008)

J.D. Lasica, rapporteur

The Seventeenth Annual Roundtable on Information Technology brought together 28 leaders and experts from the ICT, financial, government, academic, and public policy sectors to better understand the implications of cloud computing and, where appropriate, to suggest policies for the betterment of society. Participants discussed the migration of information, software and identity into the Cloud and explored the transformative possibilities of this new computing paradigm for culture, business and personal interaction. The report of the roundtable, written by J.D. Lasica, offers insights from the roundtable and includes a set of policy recommendations and advice for the new presidential administration. 2009, 98 pages, ISBN Paper 0-89843-505-6, $12 per copy.

Beyond the Edge: Decentralized Co-creation of Value (2007)

David Bollier, rapporteur

The 2007 Roundtable convened 27 leaders to analyze the current and future social and economic impacts the co-creation of knowledge across networks made possible with new communications and information technologies. While collaborative engagement encourages increased productivity and creativity, it can also lead to mass chaos from the co-creation process. The roundtable participants discussed what separates successes from failures in the new collaborative era by reviewing business and organizational models and the implications of new models. 2007, 64 pages, ISBN Paper 0-89843-481-5, $12.00 per copy.

The Mobile Generation: Global Transformations at the Cellular Level (2006)

J.D. Lasica, rapporteur

The 2006 Roundtable examined the profound changes ahead as a result of the convergence of wireless technologies and the Internet. The Roundtable addressed the technological and behavioral changes already taking place in the United States and other parts of the world as a result of widespread and innovative uses of wireless devices; the trends in these behaviors, especially with the younger generation; and what this could mean for life values in the coming decade. The Roundtable tackled new economic and business models for communications entities, social and political ramifications, and the implications for leaders in all parts of the world. 66 pages, ISBN Paper 0-89843-466-1, $12.00 per copy.

When Push Comes to Pull: The New Economy and Culture of Networking Technology (2005)

David Bollier, rapporteur

The author considers how communications, economics, business, cultural, and social institutions are changing from mass production to an individualized "pull" model. *When Push Comes to Pull* describes the coexistence of both push (top down or hierarchical) and pull (bottom up or networked) models—how they interact, evolve, and overlay each other in the networked information economy. The report explores the application of "pull" to the worlds of business and economics; the content and intellectual property industries; the emergence of an economy of the commons; and personal and social dynamics, including leadership in a pull world. It also touches on the application of the pull model to learning systems; the military, in the form of network-centric warfare; and the provision of government services. 78 pages, ISBN Paper 0-89843-443-2, $12.00 per copy.

Information Technology and the New Global Economy: Tensions, Opportunities, and the Role of Public Policy (2004)

David Bollier, rapporteur

This report provides context and insight into the unfolding of new economic realities arising from the information revolution—how the world's players will live, learn, innovate, offer, consume, thrive, and die in the new global economic landscape. *Information Technology and the*

New Global Economy draws a portrait of a changing global economy by describing new business models for the networked environment, exploring topics of innovation and specialization. Among the more creative concepts propounded at the Roundtable was an analysis of the world's economy in terms of video game theory that suggests that if developing countries are not incorporated into the world economic community in some acceptable way—if they cannot make economic progress—they could become disrupters to the entire economic or communications system. The report also explores issues of outsourcing and insourcing in the context of digital technologies moving work to the worker instead of vice versa. Participants concentrated on developments in India and China, taking note of some of the vulnerabilities in each of those countries as well as the likely impact of their rapid development on the broader global economy. 57 pages, ISBN Paper: 0-89843-427-0, $12.00 per copy.

People / Networks / Power: Communications Technologies and the New International Politics (2003)

David Bollier, rapporteur

This report explores the sweeping implications of information technology for national sovereignty, formal and informal diplomacy, and international politics. Bollier describes the special challenges and new rules facing governments and nongovernmental organizations in projecting their messages globally. The author further explores the relationships between the soft power of persuasion and the more traditional hard power of the military and discusses how governments will have to pay close attention to newly burgeoning social communities in order to prosper. 68 pages, ISBN Paper: 0-89843-396-7, $12.00 per copy.

The Rise of Netpolitik: How the Internet Is Changing International Politics and Diplomacy (2002)

David Bollier, rapporteur

How are the Internet and other digital technologies changing the conduct of world affairs? What do these changes mean for our understanding of power in international relations and how political interests are and will be pursued? *The Rise of Netpolitik* explores the sweeping implications of information technology for national sovereignty, formal and informal international diplomacy, politics, commerce, and cultural identity. The

report begins with a look at how the velocity of information and the diversification of information sources are complicating international diplomacy. It further addresses geopolitical and military implications, as well as how the Internet is affecting cross-cultural and political relationships. It also emphasizes the role of storytelling in a world in which the Internet and other technologies bring our competing stories into closer proximity with each other and stories will be interpreted in different ways by different cultures. 69 pages, ISBN Paper: 0-89843-368-1, $12.00 per copy.

The Internet Time Lag: Anticipating the Long-Term Consequences of the Information Revolution (2001)

Evan Schwartz, rapporteur

Some of the unintended consequences of the Internet and the freedoms it symbolizes are now rushing to the fore. We now know that the network of terrorists who attacked the World Trade Center and the Pentagon made full use of communication technologies, including email, Travelocity.com, automatic teller machines (ATMs), data encryption, international money transfers, cell phones, credit cards, and the like. Is the Internet an epochal invention, a major driver of the economy for many years to come, or just a passing fad? Will the new phenomena of recent years—such as the contraction of hierarchies, instant communication, and lightning-fast times to market—last beyond the funding bubble? What is the next new economy? What are the broader social consequences of the answers to those earlier questions? This report takes a wide-ranging look at the economic, business, social, and political consequences of the Internet, as well as its ramifications for the process of globalization. 58 pages, ISBN Paper: 0-89843-331-2, $12.00 per copy.

Uncharted Territory: New Frontiers of Digital Innovation (2001)

David Bollier, rapporteur

This report looks critically at key insights on the new economy and its implications in light of the digital revolution. The report begins with an examination of the interplay between the current economy and the capital economy and then probes the emerging world of mobile com-

merce and its potential for driving the next great boom in the economy. It further explores new business models resulting from the combination of mobile communications and the new economy. 68 pages, ISBN Paper: 0-89843-307-X, 12.00 per copy.

Ecologies of Innovation: The Role of Information and Communications Technologies (2000)

David Bollier, rapporteur

This report explores the nature of innovation and the role of the information and communications sectors in fostering ecologies of innovation. In this context, the report examines the ways in which the creation of new ecologies is affecting significant societal institutions and policies, including foreign policies, industry and business structures, and power relationships. 44 pages, ISBN Paper: 0-89843-288-X, $12.00 per copy.

Reports can be ordered online at *www.aspeninstitute.org* or by sending an email request to *publications@aspeninstitute.org.*

About the Communications and Society Program

www.aspeninstitute.org/c&s

The Communications and Society Program is an active venue for global leaders and experts from a variety of disciplines and backgrounds to exchange and gain new knowledge and insights on the societal impact of advances in digital technology and network communications. The Program also creates a multi-disciplinary space in the communications policy-making world where veteran and emerging decision-makers can explore new concepts, find personal growth and insight, and develop new networks for the betterment of the policy-making process and society.

The Program's projects fall into one or more of three categories: communications and media policy, digital technologies and democratic values, and network technology and social change. Ongoing activities of the Communications and Society Program include annual roundtables on journalism and society (e.g., journalism and national security), communications policy in a converged world (e.g., the future of video regulation), the impact of advances in information technology (e.g., "when push comes to pull"), advances in the mailing medium, and diversity and the media. The Program also convenes the Aspen Institute Forum on Communications and Society, in which chief executive-level leaders of business, government and the non-profit sector examine issues relating to the changing media and technology environment.

Most conferences utilize the signature Aspen Institute seminar format: approximately 25 leaders from a variety of disciplines and perspectives engaged in roundtable dialogue, moderated with the objective of driving the agenda to specific conclusions and recommendations.

Conference reports and other materials are distributed to key policymakers and opinion leaders within the United States and around the world. They are also available to the public at large through the World Wide Web, *www.aspeninstitute.org/c&s.*

The Program's Executive Director is Charles M. Firestone, who has served in that capacity since 1989, and has also served as Executive

Vice President of the Aspen Institute for three years. He is a communications attorney and law professor, formerly director of the UCLA Communications Law Program, first president of the Los Angeles Board of Telecommunications Commissioners, and an appellate attorney for the U.S. Federal Communications Commission.