


VIRGINIA SUMMIT ON SCIENCE,
ENGINEERING, & MEDICINE

2014
MEETING **BIG**
DATA



Report of December 4-5, 2014

Virginia takes on Big Data



In the five seconds it takes you to read this sentence, the global human community will have generated another 800 trillion bytes of data. That's about 80 LOC's, where LOC stands for the printed collection of the Library of Congress. In 2015, the world will add five trillion billion bytes—that's five zetabytes or 5,000,000,000,000,000,000 bytes—to the ongoing, everywhere, all-the-time frenzy of data accumulation. This is the era of Big Data.

Financial transactions. Medical records. Census and labor statistics. Traffic patterns. Accelerator experiments. Drug trial results. Hubble Telescope imagery. Google search behavior. Phone and e-mail logs. Energy production and consumption. Ocean temperature measurements. Industrial process sensor data. Sports statistics. Personal fitness records.

This is just a tiny sampling of the ever-expanding list of data types becoming available. Al-

ready well on its way is the "Internet of Things" by which the constructed landscape of buildings, roads, trains, buses, fire hydrants, street lights, and millions of things inside of buildings and structures will feed data into a vast network of computers. For their parts, those machines will run a range of monitoring and analytic algorithms to make sense of all of these data to help in making decisions about pretty much everything.

On December 5, 2014, leaders in Virginia's science and technology communities gathered with other luminaries at the forefront of Big Data in technology, industry, health, medicine, and social and policy arenas at the National Academy of Sciences in Washington, DC, for the Virginia Summit on Science, Engineering and Medicine. The Summit was hosted by Senator Mark Warner, Virginia Tech, and the Virginia Academy of Science, Engineering and Medicine (VASEM), and marked the second gathering of VASEM members. (The inaugural VASEM Summit, held at the end of 2013, focused



on energy and health.) “The pace of data availability and acquisition is accelerating across all sectors of science and society,” according to the Summit organizers from Virginia Tech, the University of Virginia, and Virginia Commonwealth University. “This is challenging our ability to harness this information for exploration and discovery.”

An eye-opening, data-drenched portrait of the near future emerged throughout the day of talks by the likes of Vinton Cerf, one of the inventors of the Internet and now Google’s Vice President and “Chief Internet Evangelist,” Arati Prabhakar, Director of Defense Advanced Research Projects Agency (DARPA), and John Thompson, Director of the U.S. Census Bureau. Ever-growing flows of data from every sector of society will bring great opportunities for scientific discovery, technological innovation, economic growth, and personal and public health.

But data flows of this diversity and magnitude are new to humanity, including Virginians. Even as the Data Age promises an upside, it also carries with it potential downsides. This always has been so with new and powerful technologies. At this moment, the benefits and risks Big Data harbors for societies and individuals remain more speculative than known. The 2014 VASEM Summit was all about embracing that reality—in technology and industry, social science and policy making,

and health and medicine—with a goal of developing the means and ways of constructively controlling the coming rivers of data rather than being swept away by them.

“Virginia is home to some of the nation’s top technology companies and leading education and research institutions. That means the Commonwealth can play an important role in the rapidly growing sector of Big Data.”

~Senator Mark Warner

Emblematic of the gravity with which Virginia considers Big Data was the presence at the meeting of **Virginia Senator Mark Warner**. The Senator initiated the steps that would result in the creation of VASEM and its first Summit in 2013. The state-based organization, modeled after a similar program in Texas, brings together, in Warner’s words, “those great minds from across the Commonwealth, who have already been vetted for distinction by the National Academy of Sciences. Bringing together this intellectual firepower is a great asset.”

With the more intentional convening of the state’s science and technology talent that

VASEM embodies, Warner said he hopes the state will attract more than the 6% of the nation’s R&D funding that it received in the last budget cycle (Maryland scored 11%, he noted). He sees VASEM also as a means for inspiring coming generations of scientists and engineers in Virginia to excel. With institutions like the state’s universities as well as NASA’s Langley Research Center, DARPA, and the Howard Hughes Medical Institute’s Janelia Research Campus, Virginia is well placed for growth in the science and technology sectors. Big Data, Warner projected, will be a central component of this effort.

“In an era of diminished federal support for research and development, it is more important than ever for Virginia’s leading thinkers across institutions and disciplines to increase collaboration,” Senator Warner said. “Virginia is home to some of the nation’s top technology companies and leading education and research institutions. That means the Commonwealth can play an important role in the rapidly growing sector of Big Data. This second annual Summit provides an exciting venue for some of the best and brightest in Virginia to discuss the challenges and benefits of Big Data.”

The wise and skilled use of data, Warner noted at the Summit, ought to increase government efficiency, which is becoming more important than ever. “For every dollar you send to Washington, only 17 cents goes into the discretionary budget,” which includes funding for environment, education, and science R&D. “Some in Congress want to take the 17 cents down to nine cents,” he said. “I have been a business man longer than a politician and as an investor, I would never support an enterprise that spent less than 10% of revenue in R&D.” He implored those gathered at the Summit to find ways of leveraging data so that the diminishing discretionary budget still might buy more for the public than it has in the past.

“It is not often that a politician will fund-raise for you, rather than the other way around,” the Senator told the gathering, promising to help his in-state constituency to bring in more federal R&D dollars. As a start, he said, “I will be quiet and listen. This is great for Virginia. Great for the advancement of knowledge. And great for the country.”

The Landscape of Big Data

The opening talk by [Kenneth Prewitt, Carnegie Professor of Social Affairs at Columbia University's School of International and Public Affairs and former Director of the U.S. Census Bureau](#), would prove to be as provocative as the organizers had hoped. With the help of the "Internet In Real Time" website, Prewitt quickly gave attendees a feel for the data flows that already are streaming at astounding volumes and rates. "By the time you finish reading this sentence, there will have been 219,000 new Facebook posts, 22,800 new tweets, 7,000 apps downloaded, and about \$9,000 worth of items sold on Amazon," the site says. Meanwhile on the site, counters tally up, in real time, data traffic on the Internet, Facebook and Instagram postings, and money spent on Amazon.

Big Data has been a long time coming, Prewitt reminded the audience, pointing by way of example to the Census, which he described as "one of the original Big Data programs." With sources that include financial transactions, GPS systems, Internet searches, embedded sensors, and social media, data flows are growing at exponential rates. Cloud computing was a \$41 billion industry in 2011 and slated to grow to more than \$240 billion by 2020. "Fifty billion devices will be connected to the Internet of Things in 2020," he said, adding that the data "will steadily and dramatically increase its usefulness to science and policy."

He pointed to present day examples to back this claim. General Electric, he noted, fitted its jet engines with sensors that over 3.4 million miles of recent travel, allowed the company to detect possible defects 2,000 times faster than with prior checking practices. The second example had to do with predictions of quarterly

house sales. Trends in housing-related Google search queries are more accurate than forecasts of real estate economists, Prewitt said.

By the time you finish reading this sentence, there will have been 219,000 new Facebook posts, 22,800 new tweets, 7,000 apps downloaded, and about \$9,000 worth of items sold on Amazon

~"The Internet in Real Time" website

Compared to traditional data gathered from surveys, digital data offer high "granularity," by which Prewitt meant data derived from smaller geographic areas and shorter times between data gathering efforts. Add to higher granularity traits like low unit cost for the data, real-time availability, the possibility of continuous trend lines in the data, and powerful tools for data analytics and visualization, and the case for transforming traditional data-intensive programs (such as the once-a-decade census survey) become compelling. "We can reduce the time gap between events in the world and what we know about them," Prewitt said, pointing out the likely value of this timeliness when it comes to public policy and decision-making.

But there are problems to contend with. Unlike the well-defined and structured data that come from census surveys "we don't control commercial data," Prewitt said. "Google does. Apple does."



The Internet in Real Time

This figure represents 10 seconds of activity over the internet.



549,760
FACEBOOK POSTS



57,000
TWEETS



185,190
PHOTOS LIKED



23,140
HOURS WATCHED



2,380
PINS



1,820
USERS SEARCHES



5
REVIEWS



230
BLOG POSTS



34,027,780
EMAILS SENT



2,199,070
MESSAGES SENT



4,630
POSTS



231,480
MINUTES USED



\$23,590
MONEY SPENT



115,740
FILES SAVED



6,340
APPS DOWNLOADED



12,360
APPS DOWNLOADED



2,120
VOTES



350
CHECK-INS



3,551
HOURS STREAMED



3860
HOURS WATCHED



57,870
STORIES VIEWED



10,190
HOURS STREAMED

A stark illustration of this emerged from the recent revelations from the Edward Snowden experience in which the US government, in carrying out its most important role of protecting citizens, had to negotiate with private companies to do the sort of searching that, in Prewitt's words, "would work better than a million spies."

We can reduce the time gap between events in the world and what we know about them.

~ Kenneth Prewitt

"The government can't fulfill its most basic role without these commercial connections to vast digital data streams," he said. But that also means the government is relying, at least in part, on proprietary algorithms that filter, structure, and otherwise organize the data. Or consider a newspaper article based on the Twitterverse, that is, data about tweets associated with a particular topic or issue. This is not quite tapping into the pulse of the general population because the data come from a self-selected group that chooses to communicate via Twitter. So it becomes important to take into account the way information technologists, algorithm designers, and others control the character of the data we use.

A challenge in this context is that the methods of digital data acquisition and the design of data structures from private sources are often not transparent and users of the data cannot be certain about the reliability of the data. For science and other scholarly endeavors in which transparency about method and data is key, these fuzzy features of new digital data streams could be particularly problematic, Prewitt warned.

Another danger inherent in new data streams is that they can contain different types of hidden errors with different origins. The Census culture has always valued the pains it takes to identify and understand sampling, processing, inference-making and other kinds of errors that are part of the Census process. As Census professionals,

Prewitt said, "we show off how clever we are about finding our errors. It is in our DNA." When it comes to Google, Facebook, Twitter and other Big Data generators, he continued, "they don't know how to talk about errors yet." Google modifies its basic algorithm for carrying out keyword searches approximately 500 times in a year, which means it is not stable ground on which to base metrics for determining errors.

And errors can have consequences. Insurance companies might deploy an algorithm that taps into databases, such as toll data or transaction data, to identify drivers who are on the road late at night for many hours at a time under the premise that these likely are teenagers. They might try to jack up rates based on that inference. But many people driving at night are night shift workers, so this digital criterion is prone to false positives that in this case could hurt night shift workers. Another error prone area is the determination of health status based on third party data. "Error structure of digital data has a lot of consequences, and they can be negative," warns Prewitt.

Prewitt implored the audience to worry mightily about one more issue: who should have access to the data citizens provide, often with an expectation that the original recipient will not share the data with others. More and more, people have been giving up privacy and confidentiality in exchange for convenience of services, but now even the choice is disappearing. "I do not give consent to cameras or GPS tracking," Prewitt noted.

It is possible to refuse to use Google if one does not want his search behavior tracked, but is this practical? Prewitt asked rhetorically. "Confidentiality and consent regulations will be a key to working well with Big Data, given that the system is in place already and we really can't avoid using it unless we recede from society," he added. A legitimate concern of the public is that data, even census data, can be reused, reconfigured and end up harming them.

Amidst all of these concerns is a great opportunity to use new data sources for doing science, answering policy questions, and otherwise furthering the public good. This will take negotiating with private data generators for access to digital data along

with information about the data's weaknesses and error structures. "The task before us, it seems to me, is to be worried about this flood of data, but also to go to work," says Prewitt. "We want to get inside the data as scientists, to look at error structure. If we are using Google data 'scientifically,' and the company's algorithm is changing 500 times a year, then we need to account for the lack of data stability." That will require forging new understandings with private data generators.

The task before us, it seems to me, is to be worried about this flood of data, but also to go to work.

~ Kenneth Prewitt

Exactly how to adjudicate differences between government and private generators and users of data is not obvious, pointed out Vinton Cerf, one of the inventors of the Internet itself and now Google's Vice President and Chief Internet Evangelist. "If we were forced to make data available as Prewitt suggests, we would be out of business," Cerf said, arguing that the combination of fully transparent government data with proprietary data sources of data like Google's can give better answers to questions of public interest than either source alone. To blend these kinds of data sets with confidence, Prewitt stressed, the Googles of the world need to be more transparent about the strengths and weaknesses of their data. "Give us some report, as CDC does, about data quality," Prewitt suggested to Cerf.

Underlying the dialog between Prewitt and Cerf could be an ineluctable error source. "We observe content on the Net and we index it. We don't know the origin of all of it," Cerf explained, implying that Google cannot vouch for the accuracy of the content that the company indexes for its search services. "We do change the algorithm 500 times a year. This is driven by the metric: 'Does the user get what he wants?'" The bottom line, Cerf said, is that "errors in the data are not Google's fault," an answer that gave Prewitt, a champion of knowing all about the data one uses, no great comfort.

Big Data in Technology & Industry

Developing new technologies and industries always have been data-intensive endeavors. Measurements of everything from genetic sequences to steel furnace temperatures to supply chain inventories to marketing demand are part of the task.

One of Virginia's prized technology partners is the U.S. government's Defense Advanced Research Projects Agency (DARPA), a multi-billion dollar, visionary, technology-development organization with roots that date back to the surprise victory in the Space Race in 1957 by the Soviet Union with its launch of the world's first satellite, Sputnik. DARPA had a catalytic role in the creation of the Internet, which has been fomenting one of the most far reaching and dramatic waves of social change in history. And it is a primary part of the infrastructure without which the reach of Big Data would be greatly limited.

Arati Prabhakar, Director of DARPA with a long track record of technology leadership in academe and the private sector, focused the attention of attendees at the VASEM Summit on DARPA's charge of nurturing R&D that will yield breakthrough technologies for national security. Many in the audience had partnered with DARPA, among them Vinton Cerf and Robert Kahn, co-inventors of the software language known as Transmission Control Protocol/Internet Protocol, or TCP/IP that underlies the transmission of data packets, the *lingua franca* of the Internet.

DARPA's job has been to prevent Sputnik-like technological surprises for the country, Prabhakar said. "You do that by creating surprise yourself," she continued. "In the

Department of Defense, we did stealth, miniaturizing GPS, immense radar arrays, and other technologies that have changed how we fight and keep the country safe." Much of this technology development has had spinoffs in the private sector. DARPA has helped plant seeds that grew into the present day cybersphere, which is a system of technological systems. "Think of materials in batteries and displays and gallium arsenide chips in cell towers," Prabhakar said.

DARPA does not run its own laboratories, but the organization's program managers help to encourage and manage the innovation of others. "Our work gets done as we tap talents at universities and companies—public and private labs—and those organizations take nascent technology and turn it into products, sometimes world-changing ones. That is our ecosystem," Prabhakar explained. Central to DARPA's *modus operandi* is to make early investments in high-risk/high-payoff ideas that otherwise might never get beyond the concept stage.

Our work gets done as we tap talents at universities and companies—public and private labs—and those organizations take nascent technology and turn it into products, sometimes world-changing ones.

~ Arati Prabhakar

DARPA's mission has not changed in its nearly seven decades of existence, but the world has changed dramatically. In 1958, the USSR and the Cold War posed the greatest threat to the country. "Now it is ISIS and Ebola, violent extremists and new biological threats, advancing capabilities of nation states, like North Korea, Iran, and Russia, and criminal activity. These are what the national security threats are," Prabhakar said. "Very powerful technologies are now available to anyone, almost, and that adds to the security challenge," Prabhakar said at the Summit.

One of the data-intensive arenas that DARPA has been entering, Prabhakar noted, is the part of the cybersphere—many call it the Deep Web—that is not available on the Internet. These are the data that do not live in websites and so are not indexed and made searchable by Google. Sensor data associated with the Internet of Things is part of this larger cybersphere. One of the challenges in this context is to even know about the existence of these "dark" data, let alone how to make sense of them. "Can we make tools that take this data and turn it into information?" Prabhakar asked rhetorically.

One DARPA-relevant take on the Deep Web is a program called MEMEX, which has the objective of developing tools that can find signs of bad behavior in data patterns living in the Deep Web. "The vastness of information space creates places bad actors can hide," Prabhakar said, pointing as an illustration to bad actors involved in human trafficking. Law enforcers routinely turn to search engines like Google and Bing to scour through sex trade ads, but such searches can only gather data from indexed websites and with the algorithms available from the

commercial search engines they use. The MEMEX program is aimed at the Deep Web, not just the small part of the Web that is indexed, and it calls for the development of new search paradigms that eke security-related information from both the indexed and dark parts of the cybersphere. "An index curated for the counter-trafficking domain, along with configurable interfaces for search and analysis, would enable new opportunities to uncover and defeat trafficking enterprises," according to a description of MEMEX on DARPA's web site.

The vastness of information space creates places bad actors can hide.

~ Arati Prabhakar

In a demonstration project with law enforcement partners in the Dallas, Texas area, researchers supported by the MEMEX program developed tools to include, for

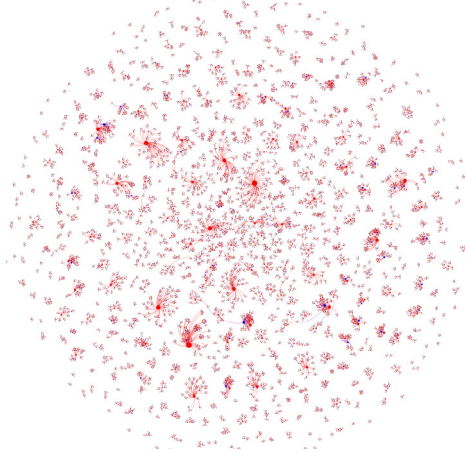


Figure 1. Cluster map created by MEMEX

example, phone numbers culled from the back page sex trade ads on publications and then used those in searches of Deep Web server cores that generally are not indexed by traditional search engines. "The use of forums, chats, advertisements, job postings, hidden services, etc., continues to enable a growing industry of modern slavery," according to DARPA's website. The MEMEX search protocols netted phone numbers showing up in ads and developed cluster maps showing where the numbers were active (Figure 1). "When we scooped this large data set, we found 600 phone num-

bers and threw those over the wall to law enforcement colleagues, and then let them deal with it within their rules of privacy engagement," Prabhakar explained. "They were stunned by the numbers we gave them." With the MEMEX approach, they found numbers linked to known criminal activity and even ones associated with North Korea that did not show up in their conventional searches, and were also able to tie these numbers to fund transfers, she said.

Prabhakar pointed to the national BRAIN Initiative, a large-scale and long-term federal R&D program, as another data-intensive domain where DARPA is investing resources. "It will take decades to understand what it means to be human, in a brain context," she noted, adding that a DARPA spin here is to learn enough about how the brain works to devise new ways of identifying, predicting, countering and otherwise managing threats. "This is not so much a story about Big Data as it about the power of using and manipulating data," she said.

One brain-related example that Prabhakar shared, from researchers at the University of Pittsburgh, involved implanting electrodes in the motor cortex of a paralyzed woman who was unable to use her hands. She learned how to commandeer the neural signals that the electrodes were picking up so that she could control robotic arms (which she named Hector and Lector, respectively) with enough finesse that she could shake hands or fist bump with visitors. "We have opened a door to restoration of function unlike anything before," Prabhakar said. "You can imagine ways in which we can use human brain function, because we tap into the brain's signals."

Ironically, this particular project has a small-data status: the electrodes tap into about 200 neurons and pull off a few kilobits of signal data per second. "Figuring how to take just that amount of data, and interpret the signals and translate these in real time to drive robot movements was amazing work," Prabhakar said. "Imagine when we move from 200 neurons into the vastness of something between 100 and 200 billion neurons. Imagine when it is not just electrical signaling we tap into, but also chemical signaling, and try to answer questions like 'What is memory?' The pathway to answer-

ing such questions is to master the ability to create and understand data at volumes and depths that we have not done before."

Defense and security are among the country's priorities, and so is energy. Speaking at the Summit only three days before the U.S. Senate confirmed her as Director of ARPA-E, the Department of Energy's version of DARPA (ARPA-E stands for Advanced Research Projects Agency—Energy), **Ellen Williams, Senior Advisor to the Secretary of Energy**, shared her vision of future innovation in the energy sector and the role that data will play in it. Clean, secure and affordable energy is the overall charge for the U.S. Department of Energy, remarked Williams, adding that meeting this responsibility is only possible with massive amounts of reliable and intelligently-used data.

The worldwide energy system has been immense and complex for a long time and it only will become more so. Data of many kinds will provide the intelligence to keep it all going. If all the world's energy were generated by oil, it would take 12 billion tons of this fossil fuel each year. These days, oil and the other two fossil fuels—coal and gas—each make up about one quarter of the global fuel mix. Renewable and nuclear fuels comprise the final quarter. Industry consumes about half of the power produced by these fuels (Figure 2). Electricity generation, transportation and a host of other human activities use the rest.

Producing energy requires systems of systems of systems, Williams said. The rail network for moving coal and fossil fuels from widely separated sources and power plants is part of this context of nested systems. Supply data show that some power stations are not receiving the suggested 30-day back-up supply of coal. Operating data indicate that train speeds are slowing down, which is part of the strain on the overall delivery system. Meanwhile, cargo data reveal that the number of carloads of grain being carried has been going up.

"Grain and coal are competing," Williams said, adding that the ability to characterize and quantify the effects "comes from data." The data come from lots of feedback and control systems and from lots of sensors throughout the system. And the

data are enabling management practices such as just-in-time coal delivery that can help one rail system meet the demands of both fuel and grain deliveries. With more and better data, new ways of optimizing systems become possible. “Maybe we don’t need a 30-day reserve of coal in an era of Big Data,” suggested Williams.

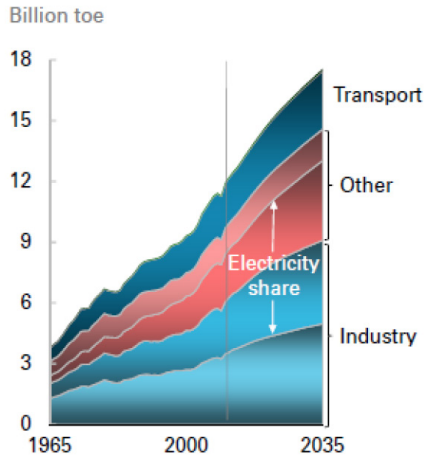


Figure 2. Energy consumption by sector

With its six million miles of distribution lines, 55,000 substations, 19,000 generating plants, 145 millions customers, and 3300 utility companies, the electric power grid is the most massive component of the overall energy system. Reliability is among the most important traits for the system, yet every day the grid needs to deliver more power while accepting varieties of power inputs from sources that include on-site electrical generators at industrial sites, wind generators atop mountains, and solar cells from individual homes. All of this extraction and injection of electricity adds to the dynamic complexity of the grid. To keep it stable, grid managers need to constantly match power, current, voltage, phase, and other parameters of high-volume electricity throughout the system.

“Getting that right is a big deal,” said Williams. “If we get it right, we will be able to integrate a diverse energy mix.” Getting it right is all about harvesting data throughout the system, understanding what the data are saying, and then knowing how to respond to those data-borne messages.

Data also will be a primary means of getting better at the energy use part of the equation. The manufacturing sector uses 25% of the country’s primary energy pro-

duction, but 2/3 of that energy is lost as waste heat. Big Data, in the form of signals from sensors feeding into analytic tools to make sense of the data, is part of the answer. Advanced controls, using sensors and process models, enable managers and technicians to know what is going on all of the time and what actions to take to optimize the energy use of their equipment. Williams said she has her eye on opportunities to leverage high-performance computing that can run higher-fidelity process models receiving higher-quality data streams from sensors throughout the manufacturing systems.

We can make gains in efficiency, integration, diversification of energy sources, and implantation of new physical approaches only because of intense instrumentation and control, and because of the data upon which those depend.

~Ellen Williams

New advanced sensors in drilling equipment promise to improve the efficiency of fuel exploration by seismic imaging, that is, by sending sound waves into the earth and seeking signs of fuel sources in the structure of the echoes that come back. These acoustic data have always been hard to interpret, especially during drilling when mud is flowing and debris is getting sloughed about. New sensors, however, can add to the acoustic data by sampling the mud for the presence of hydrocarbons, the molecules of oil. “The hope for the future is to develop better real time seismic imaging using passive sensors during drilling,” Williams said. “This will be huge data.”

In short, Williams said, “Big Data spans from physical modeling and analysis, through operational control, to supply chain and customer response. We can make gains

in efficiency, integration, diversification of energy sources, and implantation of new physical approaches only because of intense instrumentation and control, and because of the data upon which those depend.”

For Vinton Cerf, Google’s Chief Internet Evangelist, Williams’s descriptions of the electrical grid reminded him of both the circuit switching in the telephone system and packet switching underlying the Internet. He raised the questions of whether better battery technology would make the electrical grid more robust and easier to manage. Williams gave an unambiguous response: “Energy storage would be a Holy Grail.” Hydropower production, for one, varies greatly over the year and variability is one of the greatest challenges for managing the grid. Pop-up storage, which can buffer local variations in the system, would make the overall system more flexible.

Batteries are only one way to go, Williams noted, pointing also to converting surplus electricity into potential energy in the form of water pumped upward (that can later be shunted downward to drive turbines) or into chemical fuel, say, by hydrolyzing water to generate hydrogen, which then could be used regenerate electrical power. Senator Warner wondered aloud that, given how important new battery technology could be, whether it might be possible to develop an electrical “storage policy” to move battery technology forward. No one objected.

No one could have been more apropos for talking about one of the most vast and growing repositories of myriad types of data than Google’s Vinton Cerf, who is among those most responsible for the emergence and growth of the Internet.

“Scientists know more and more about less and less. Marketers know less and less about more and more, so they don’t really know anything,” Cerf quipped in his signature three-piece suit. Many arenas of science, among them computational biology, cosmology, and high-energy physics, are Big Data zones.

There are cases where so much data are being generated that there is a need to throw data away as fast as possible. An example is the NSF-sponsored IceCube Neutrino Observatory at the South Pole. The

observatory features 5200 sensor-equipped boreholes in the ice, which filter out most of the lower-energy and therefore less interesting neutrinos. So far, the huge instrument has detected within its vast data productions, in Cerf's words, "a few tens of significantly powerful neutrinos, which presumably originated outside of our galaxy."

There are so few of these detections that scientists have given these neutrinos proper names, among them Bert, Ernie, and Big Bird. To make way for the next detection, however, the system needs to make room to store newly-acquired data and that means purging some of the previously gotten data.

Think about all of the data, including the scientific data. Think about its value over time. Think about theories that might drive you to reanalyze data from the past.

~Vinton Cerf

As for Google, Cerf continued, "We have lots of data about lots of things from trawling the World Wide Web. Our job is to help people find data, so we index the data. It's a highly parallel process. The indexing programs are crawling and looking at every page, every hyperlink. We make big lists. We try not to go in circles. We compile huge piles of index info," said Cerf.

He explained that the indexes reside in a large number of different processors. "When you do a Google search, the question that goes out to the processors essentially is who has found these words anywhere in the World Wide Web? Those [processors] who found it, 'raise their hands.'" That yields a list of pages with the search words on them. To be useful, though, these hits have to be arranged in some kind of order. There are different ways of doing that and so making that decision catalyzes plenty of debate. The order of hits that shows up on browsers' displays is based on apparent importance, Cerf noted. It's a black art of sorts. Said Cerf, "We now use several

hundred signals [in the data] to figure ranking. This same mechanism works with ads." The process delivers giant lists of search hits and giant lists of ads, and algorithms work out the hit/ad dynamic in working memory, Cerf noted, not by pulling things off of disks.

Adding complexity and diversity of the data streams are data from sensors, which are becoming part and parcel of both natural (such as oceans and glaciers) and artificial (such as roadways and the electrical grid) landscapes. Ramping up the data streams in entirely new ways are data from sensors that riddle both natural and artificial environments, which is to say, the Internet of Things. This is leading to new services and capabilities. "We might be able to tell someone to leave 20 minutes earlier than they were planning to because of traffic patterns," Cerf offered as an example. That will require continuous monitoring of sensor data. This is akin to monitoring vital signs to determine base lines. In travel guidance applications, such base lines can provide a reference for determining when things like traffic or on-time performance of airlines are not normal. Eventually, algorithms that eke from data more information relevant to "situational awareness" and even people's intent could become routine.

Upping the chances that this sort of thinking is predictive and not just visionary is the emerging Internet of Things, which is projected to have 50 billion devices connected to the network in 50 years or sooner. "I hope you are worried like me that the 15 year-old next door does not take control of your heating, entertainment system and wine cellar," said Cerf, with a mix of humor and caution. Access control will become ever more important, he acknowledged.

Cerf is a data enthusiast, with an admixture of realism. "Think about all of the data, including the scientific data. Think about its value over time. Think about theories that might drive you to reanalyze data from the past," Cerf said. If it is possible to preserve all of the bits, the metadata that describes the provenance and features of the bits will have to be preserved as well. Is a temperature measurement in Fahrenheit, Celsius, or Kelvin degrees, for example? "We will need to preserve the application software, which means preserving operat-

ing systems, and also computers that ran the OS that ran the application software that made sense of the data," he continued. Failing that, he warned, future historians might not know a thing about us. If we do not want to be anonymous to our descendants, we had better do something."

In response to a question about storage challenges, Cerf noted that storage densities have been going up continuously and dramatically while costs have been going down and that has allowed the companies like Google to keep up... so far. But "...if we have to start throwing data away," he warned, "we will need protocols and decision tools for doing that." It will be a process wrought with risk, he noted: what we think is important now may not be what will be important later.

Robert Kahn, another Internet pioneer, reminded the gathering that the datascape is much larger than the portion of the Web Google indexes. "How do we take advantage of all the information in private data bases, which is bigger than the Web? What is technologically possible?" he asked. There are coding options now that essentially hang a "don't crawl" sign on data so that Google and other public search engines will not find and index it. "Lots of info is dark info," Cerf concurred. "We need to establish standards that will open indexing up to dark information."

Senator Warner observed that the government would stand to benefit from access to currently inaccessible data. "We know more about commercial transactions than we do about citizens' transactions with government at any level," he said. "We should have a common financial reporting system, but instead we have 200 just within DoD. We don't know enough about ourselves. Can we use Google to get the citizenry data that we can't get in government?"

Cerf mulled out loud over that possibility. He mused that the Senator push for funding projects that would develop the means for blending data in government databases with those in the hands of companies like Google.

Big Data, Smart Cities

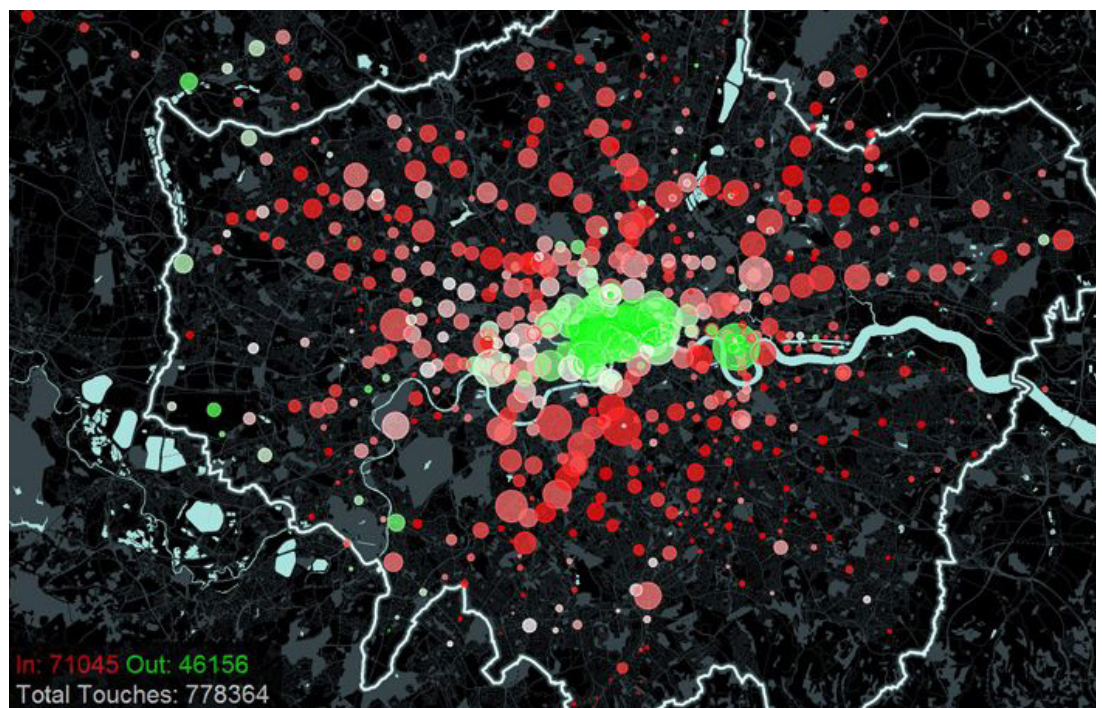


Figure 3. CASA, Flows of the London Underground

Over lunch, in the august Great Hall of the National Academy of Sciences, [Michael Batty, Bartlett Professor of Planning at the University of London and Chair of the Centre for Advanced Spatial Analysis](#) picked up on the theme of the Internet of Things by showing how embedded sensors and smart mobile devices are leading to what some refer to as “smart cities.” Data flowing from such sources as social media, buses and trains, and GPS sensors in smart phones, is amounting to a Big Data reservoir that can enable city managers, officials, and citizens to visualize the ebbs and flows in their urban environment.

A prescient thinker in this context was Patrick Geddes, whose 1915 book *Cities in Evolution: An*

Introduction to the Town Planning Movement and to the Study of Civics, applied Darwinian evolutionary theory to cities. “For most of 20th century, we thought of cities as machines,” Batty said, adding that this led to strategic platforms based on ripping cities apart and reassembling them as needed. “But cities are sensitive, subtle organizations. If you tinker with and disrupt them, they might not work well.” This is where data, lots of it, and finely resolved in space and time, can help cities to run and evolve more intelligently (Figure 3).

As an example, Batty showed how sensors that monitor usage in the London Underground harvest data that enable the generation of visualizations valuable for both managers and riders of

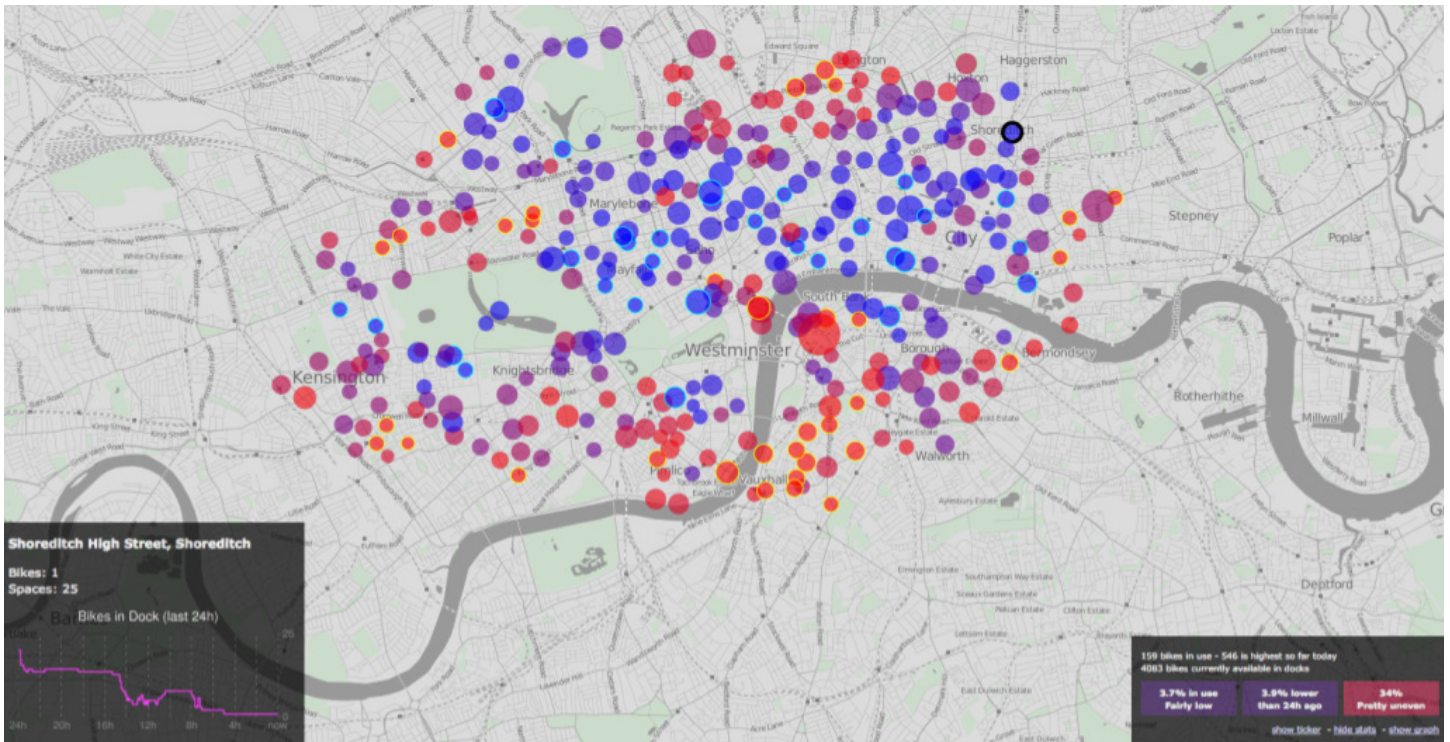


Figure 4. Bicycle use in London.

the system. Apparent in the dynamic map, in which dots and lines grow or shrink depending on traffic, are morning and evening peaks as well as a third peak that correlates with people going out for entertainment. The visualization depicts the city as pulsing, rather than static the way a traditional map of public transportation does.

For most of the 20th century, we thought of cities as machines. But cities are sensitive, subtle organizations. If you tinker with and disrupt them, they might not work well."

~Michael Batty

In another visualization, Batty showed patterns of people in London hiring bicycles from docking stations akin to the Bikeshare system in U.S. cities. In this depiction, red dots represent full docking stations and different sized blue dots represent different degrees of dock emptiness. The data are online and updated in real time. What unfolds are red and blue dots growing, shrinking, appearing,

and disappearing (Figure 4). In yet another visualization, nitrogen oxide (NO_x) sensors throughout London depict fluctuating air pollution levels (associated with starting and stopping traffic) on a second by second basis. "This helps to communicate things that are important to stakeholders," Batty stated.

"A new paradigm is to think of cities as organisms, not machines," Batty said. "To understand cities, we need to understand how energy flows. If we interfere in cities, we will alter them in ways we need to know." This organismal way of thinking draws upon a range of conceptual and mathematical frameworks—among them complexity theory, cellular automata, power laws, nonlinear dynamics, fractal geometry, scaling relationship, and statistical physics—that are amenable to computer modeling. Adding opportunity to enrich the organismal visualization of cities will be data from social media, such as the number of tweets in a given area about a particular city-relevant topic, such as "antisocial behavior," and even by using the Google Translate utility to map multilingual activity in the city.

Rendering these rich frameworks relevant and useful is that they are informed by the sensors and computers that city engineers have been embedding everywhere in public spaces to improve efficiency

by automating what previously were human-managed functions. This is opening up more opportunities to change management practices from near time to real time protocols, with decision- and action-taking cycles on the order of minutes, hours, and days rather than months and years, said Batty.

It will take better theories that more accurately relate real-world activity to virtual depictions of that activity to move into a future of data-driven smart cities, Batty noted, referring to data sources such as email, Skype, and retail transactions. More online shopping at home for example, could correspond to fewer Underground commutes to downtown stores. Also important, Batty said in response to a question by Robert Kahn, was to take on the technological challenge of collecting data that would reveal how people use energy in their city lives, a task that he says will take clever algorithms and insights about human behavior:

"We need good theory to make sense of all of this. We need to understand systems, like cities, of ever increasing complexity," said Batty, noting that it won't be easy. "As we get more wealthy, and install more technology, our cities will become faster moving targets of increasing complexity."

Big Data in Social & Economic Statistics

Of all phenomena that would be valuable to quantify and datify, human behavior might be the most consequential. To get better at that will take gathering data in a diversity of ways and from many sources, said [John Thompson, current Director of the U.S. Census Bureau](#), one of the most data-steeped agencies in all of government. “Most of what we know about the economy in US society flows from government data,” he said at the Summit. The paradigm that underlies much of the government’s data-gathering efforts derives from probability sampling of people, establishments, or other units of interest. In short, it is based on the assumption that getting data from a representative sample of a larger group will reflect the data that would be gathered were each and every member of the group providing data.

The survey instruments are designed by researchers, consistently applied in careful sampling procedures, and the resulting data are assembled and analyzed into sets of statistics that reflect the sampling process. Those data then are archived and made publicly available.

“We can look at anonymous data from the ‘50s and the error properties are usually measurable,” Thompson said. “But this system is really slow. We might get snapshots once a month or so, or once a year or more.” In other words, he said, “the temporal granularity of the data” is not good. Spatially, these surveys are weak as well. Noted Thompson, “We don’t know the unemployment of the block over there, but we do of cities, states, and the country.” Increasing the temporal and spatial granularity of census data will be an important challenge, Thompson said.

“In our world, we design measurements, and we measure tons of attributes on the same data records,” Thompson said. On the contrary, Big Data today in areas like genomics and energy use are not so multivariate—they are narrower in scope—but can be, in his words, “wonderfully timely.” With geo-location tools, such as GPS chips in mobile devices, the spatial granularity of surveys is getting better.

We don't know the unemployment of the block over there, but we do of cities, states, and the country.

~ John Thompson

The downside, Thompson said while invoking a caution voiced earlier in the day by former Census Bureau Director Kenneth Prewitt, is that analysts of the data are losing control over the measurements they work with. “Sometimes they are analyzing data with properties at the point of measurement that are unknowable,” he said. For example, they might not know all they would want to about how randomized a sampling of survey takers was. “The growing consensus in our world is that the future is neither exclusively surveys nor Big Data, but it will be putting them together,” Thompson said. That will require overcoming the cultural and sociological segregation between those with the skill sets it will take. It will mean that “...social scientists, computer scientists, and mathematicians will need to work together to transcend silos,” Thompson said.

Another cultural hurdle derives from the reality that the traditional ways and means for social and economic monitoring were put into place by institutions devoted to the common good. As the datasphere becomes more of a blend from government and private sources, “we will have to blend common good uses and private sector uses that might exist for profit and marketing,” Thompson said.

Picking up the baton from Thompson, [Erica Groshen, the 14th Commissioner of the Bureau of Labor Statistics](#), gently reminded private sector attendees at the Summit that, compared to agencies like hers, they actually are Johnny-come-latelies to the Big Data business. “But you have a lot to offer,” she assured them. Said Groshen: “The addition of data scientists, technology, and data itself will enrich the work of statistical agencies. And our experience, views, and history will advance work you are trying to do.”

As her Bureau moves forward into the emerging era of a Big Data landscape, Groshen highlighted two primary opportunities. One is to leverage new sources of data and new technologies to, in her words, “do what we do better, faster, cheaper.” The Bureau also is tapping into “corporate data dumps” to supplement government-initiated data gathering. “These steps can lower costs and lower the burden on respondents,” she said.

The addition of data scientists, technology, and data itself will enrich the work of statistical agencies. And our experience, views, and history will advance work you are trying to do.

~ Erica Groshen

The second major opportunity she called out was to fuse or otherwise blend Bureau-originated data with newly available data sets from different sources in ways that yield novel and useful products relevant to labor issues. An example here might be to merge hurricane maps and labor statistics to

reveal how weather events and trends influence subsequent job and labor trends.

“We see Big Data as an important means to an end, not an end unto itself,” Groshen pointed out, noting also that her Bureau also must balance costs, risks, and potential payoff. The quality, provenance, and reliability of data become primary concerns in this context, again mirroring words that Prewitt had voiced earlier: “We see all of the hubbub and furor about Big Data. We are thrilled that there will be more data scientists and better visualization of the data,” she said at the Summit. But “we won’t use Big Data to produce something that is not production grade,” she concluded, referring to a standard akin to an industrial one in which flawed units on a production line would be culled and discarded rather than used or sold.

A conundrum for data-intensive government bodies, commented U.S. Census Bureau Director John Thompson, is increasing pressure to produce data on a timelier basis but with no additional funding. To reduce the time it takes to produce products and the cost of that work, which Groshen had stressed as important, agencies will need to take advantage of new data resources. “One thing we are doing at the Census Bureau is giving data takers iPhones,” Thompson said, referring to the Bureau workers who make in-person visits to the people who do not self-respond to the surveys. “On a daily basis, we need to track takers,” Thompson said. “This generates huge amounts of data and a big challenge is to deploy those techniques to increase efficiencies.”

Consumer spending statistics is a category in bad need of modernization, Groshen added. The Bureau of Labor Statistics’ Consumer Expenditure Survey, which uses data that feed into inflation calculations, now relies on a decades old data-gathering strategy. Said Groshen: “We need to bring it into the 21st century using lots of new data capture techniques. We will run experiments this year.” Once again, the theme of developing protocols for blending those measurements that are under an agency’s control with those that are not came up. Social network data, a rich source of consumer self-reporting, might then become tappable, for example.

“I am excited about this overall, because if proper blends [of diverse data sets] could

be designed, then federal statistical systems could provide more timely indicators,” remarked [Robert Groves, Georgetown University’s Provost and a former Director of the U.S. Census Bureau](#). “We gather unemployment measures once a month even though we know there are short-term shocks, so people are making decisions on older data than they should be.”

Groves projected that those countries that become the most adept at coordinating and blending data sets from government, private, utility, think-tank, non-government organization (NGO), social network and other sources will be at a competitive advantage. “It would be great if our country could be [a leader],” he said.

We gather unemployment measures once a month even though we know there are short-term shocks, so people are making decisions on older data than they should be.

~ Robert Groves

Different sources have “independently created very rich data sources that could be exponentially good for the country if they were blended,” Groves said. For the public to embrace these uses of diverse data sets, people will need to develop a trust regarding the provenance and quality of the data, he noted, adding that privacy laws and issues will be an important part of the equations. Not to mention, according to Groshen, the question of the potential misuse of social data, perhaps even by “hostile governments who might attack us with help from the data.” Yet another concern, Thompson noted, is that it now is possible to produce statistics that look like good data but might be data with an agenda. “It is hard to see inside the black box,” he said, suggesting that norms of transparency, in which data providers open their black boxes, could help to understand and manage such risks.

Big Data in Health and Medicine

Early in the millennium when scientists cheered the completion of the public and private Big Biology initiatives to determine the sequence of the three billion genetic letters (nucleotides) distributed on the 46 chromosomes of a human being's genome, the notion of Big Data in health infiltrated the general public. But there was a downside too, said **Steven Woolf, Director of the Center on Society and Health at Virginia Commonwealth University**, as he opened up a session titled "Big Data in Health and Medicine."

"People now instinctively assume Big Data in health means genomics," he said. "But the health-relevant opportunities with Big Data are much broader than that, from neurons to neighborhoods," he noted, borrowing from the title of a report published by the National Academies Press. Even 20 years ago, Woolf said, bodies like the Institute of Medicine were explaining that health is shaped by many "layers of causal factors," and the modern era is bringing large amounts of data about the role of these determinants in shaping health outcomes.

"We want to improve the health of populations," said **Sandro Galea, Dean of the School of Public Health at Boston University**. "How can Big Data lend itself to this effort?" He envisions a society in which you don't need to treat conditions like schizophrenia, because you don't get schizophrenia to begin with. Identifying populations at risk for particular maladies and reducing those risks is central to this cause and a practice that has already scored some successes. Cigarette smoking rates in the U.S. are down from 43% in 1965 to 17% or so in 2011, noted Galea, who has an M.D. degree as well as a Dr.PH (Doctor of Public Health). But the country is not doing well in other areas, as evidenced by the country's rank for mortality of 16 out of 17 peer countries.

"Our focus on Big Data, as in much of health, has been very much on the individual," Galea noted. "It is all about you," he continued, pointing to the fitness and dieting industries as examples of the focus on individual action. But that message has costs in terms of public health, costs that bite back at the individual. "Public health is what we, as

a society, do collectively to assure the conditions for people to be healthy," Galea said.

A "geopopulation" comparison of the number of people affected, respectively, by flu, HIV, and Ebola is one way data can help people grasp relative threats and thereby decide where to focus resources (Figure 5). The example shows the population reach of flu in the country with a red-colored block of states that includes Indiana, Ohio, New York, West Virginia, Pennsylvania, and New Jersey. The population affected by HIV, meanwhile, corresponds only to those residing in the Chelsea South District (zip code 10011) of New York City. Ebola's reach is depicted as four red dots in Washington Square Park, which is just a tiny part of Chelsea South. Said Galea, "There are dozens and dozens of ways Big Data can improve public health."

The opportunities are expanding, he says, because the "availability of data is so dramatically changing." He draws historical inspiration from the famous success by the father of epidemiology, John Snow, to end a cholera epidemic in London by way of data. After creating a "ghost map" that showed where individuals had died in the epidemic, Snow could see a gradient of deaths that pointed, almost literally, to a public well on Broad Street. When he removed the handle of the well, and thereby a source of hand-to-hand transmission, the sickness stopped spreading.

That iconic cholera example was marked by a simple linear relationship between the incidences of death and the proximity to

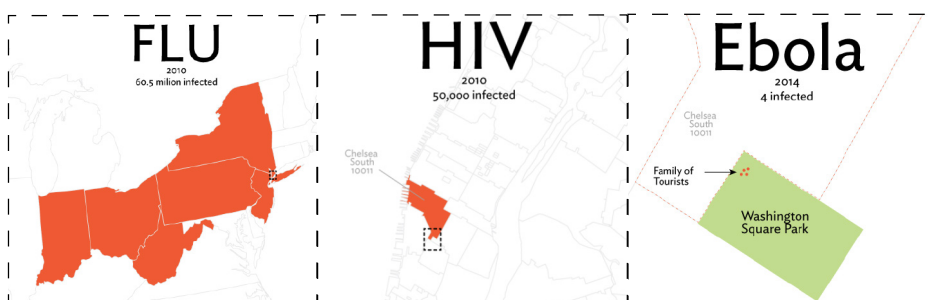


Figure 5. The example shows a "geopopulation" comparison

the public source of water. But many public health problems are too complex for their solutions to show up in simple mappings like the one that worked so well for Snow. A map showing the incidence of autism, for example, includes puzzling features such as adjacent states with rates that differ by over 1000% and other adjacent pairs with far more similar rates. "This argues that one-cause/one-outcome, linear thinking, is not applicable" for autism, Galea said.

Obesity is another nonlinear example in which multiple factors, among them physiological factors, food factors, physical activity, and psychological and cultural forces conspire into a rich and complex dynamic that maps out into networks that look like complicated subway systems. These maps illustrate that it is not possible to isolate a single deterministic cause. Rather, Galea explained, "there is a causal architecture with lots of factors that create conditions for health and illness."

There are dozens and dozens of ways Big Data can improve public health.

~ Sandro Galea

Complicated as they are, such data-derived mappings can provide guidance for enacting policies and devoting resources in ways that are likely to yield the greatest return on investment for public health. "We must use Big Data to get a handle on this," Galea said (Figure 6).

Understanding demographic trends and data is another pathway toward improving public health. "Urbanization has become the most common way to live," Galea said. "In 2050 more than 50% of the global population will live in cities." When overlaid onto city maps, for example, data about adverse living conditions (as in decrepit housing), obesity rates, and walkability, which is a surrogate for physical activity, can help public health professionals identify where they need to focus their attention and resources.

"Big Data has tremendous potential," Galea said, while cautioning that there is hubris among dataphiles who claim "we

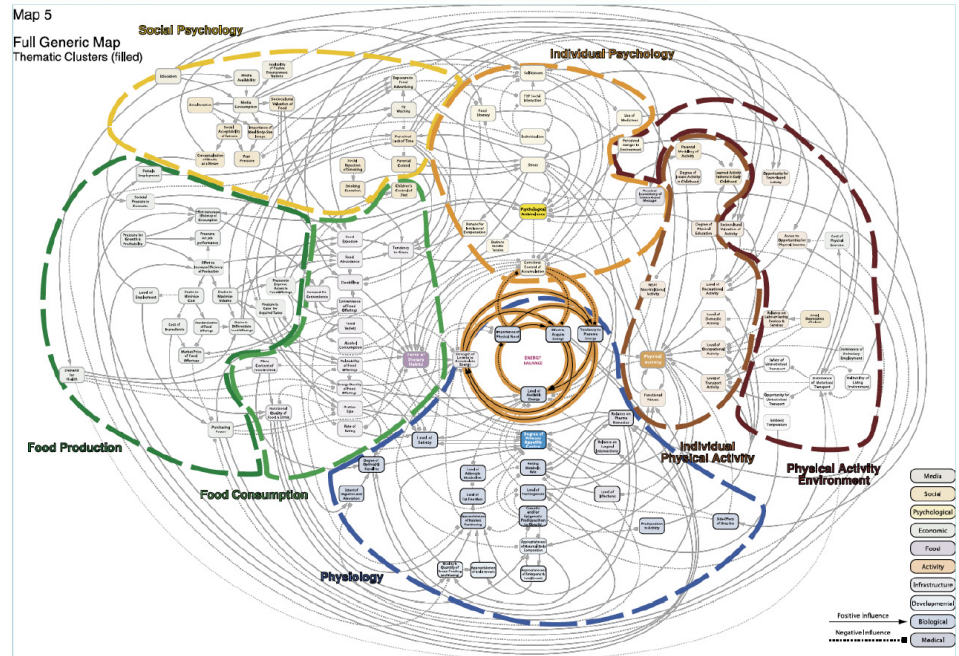


Figure 6. The full obesity system map with thematic clusters. Variables are represented by boxes, positive casual relationships are represented by solid arrows, and negative relationships by dotted lines. The central engine is highlighted in orange at the center of the map.

can use data to predict things." The myth of personalized medicine, in which personal genomic data would enable physicians to execute "precision medicine" tailored to each individual is a case in point, he said, pointing to work that has related genotypes to diabetes incidence. The resulting data show that ensembles of genes lead to a higher incidence of diabetes, but this is a purely descriptive result, not a predictive one, he stressed. The data do not provide guidance on actions to take to lower the incidence of diabetes.

Moving forward, Galea's prescription is this: leverage the emerging availability of diverse data types with the goal of "robust knowledge integration" within a strong foundation in population health principles and a "translational agenda." The result will be what he calls an "epidemiology of consequence." A challenge to complying with this prescription, Galea warned, is a cultural ethos of using Big Data in attempts to do predictive medicine for individuals. Argued Galea, "This has the effect of squeezing out public health, even though public health will do more for individuals than personalized medicine." A commenter in the audience commiserated with Galea about the government stress on individualized precision medicine, but noted

that there are efforts, including ones by the Robert Wood Johnson Foundation, to merge patient data and community data to move forward on public health.

There is plenty of incentive to get more out of our healthcare system, said Larry Green, Professor of Family Medicine at the University of Colorado in Denver and Chairman of the National Committee on Vital Health Statistics. In the eight hours of the VASEM Summit, "we will spend up to \$2.7 billion on healthcare," he calculated. "We spend like crazy but our expected longevity goes down, now all of the way to the bottom of 17 peer countries. Something is very, very wrong." Part of the problem, he said with a nod to Galea's theme, is the mistake of thinking of ourselves as predictable machines. Instead, he implored the audience, in agreement with Batty's view on cities, to "think of yourselves not as machines, but as complex adaptive organisms." What's more, he said with some humor, keep in mind that "you are very weird and you are very unpredictable."

For Green, the basic statistics on why Americans die prematurely should provide a guide for the sort of data society needs to improve public health. A pie chart showing these causes indicates that 40% of premature deaths in America are

due to behavior choices, 30% to genetics, 15% on socioeconomic, 10% on medical care, and 5% on environmental factors.

A strategic approach Green favors has roots that go back decades, in a now little-known book titled *Health is a Community Affair*. “You could rip the cover off and put today’s date on it,” he said. What is different now is that data that can help make community-based health approaches work are becoming available. Data can glue together all of the players, among them patients and health workers and those working in food, education, and city services. The strategy of creating community-based “health learning systems” has been building momentum, he noted, in part due to revisitations of the concept in the past few years in publications such as *The Annals of Family Medicine* and authoritative bodies like Department of Health and Human Services’ National Committee on Vital Health Statistics.

“In a learning system,” Green explained, “people, actions, results, and knowledge are connected in continuous feedback loops that enable improvement and change—learning—over time.” In this context, relevant data lead to revealing analytics, which lead to better local health. This constitutes a learning health system that works and improves by way of data. The learning unit is the local community.

Can we move into Big Data with a galvanizing idea that we are moving into this together and that we should all be stewards of the data?

~ Larry Green

Finding ways to handle privacy, confidentiality, and data security issues is a major challenge for establishing and running such learning health systems. These systems can only work with access to data, much of which is very personal and, in the wrong hands, could have negative consequences. As such, Green said, “health data stewardship” needs to become a priority of the overall project.

“Can we move into Big Data with a galvanizing idea that we are moving into this together and that we should all be stewards of the data?” asked Green. “We do not have the infrastructure to move forward well,” he said, suggesting that something like a federal level “Health Data Modernization Act” would help. “We in the health industry need to steward the resource of data,” he said, going so far as to say there ought to be a national “czar” in charge of this new health data stewardship paradigm. Without a chain of trust, community-based health learning systems cannot endure.

As daunting as such steps might seem, Green remains optimistic that community-based, data-driven health systems will, in his words, “change healthcare like you would not believe,” and for the better:

Data-gathering tools will be central components to build the sort of infrastructure Green described. And getting data “beyond the usual suspects” will expand the potential payoffs of Big Data-driven health systems, said [Kevin Patrick, Professor of Family and Preventive Medicine at the University of California, San Diego and Director of UCSD’s Center for Wireless and Population Health Systems](#).

“Everything matters in health on scales that range from cells to society,” said Patrick. Behavior on individual, community and national scales only adds to

the complexity. All of these factors influence one another, culminating in a complex dynamic of causality when it comes to both public and individual health.

Traditional health studies—including ones investigating new drugs, health behavior, and disease events and trends—have featured research designs such as randomized controlled trials, surveys, and surveillance based on self-reports. “Now we are moving toward wearable devices for tracking health-related states,” Patrick said. More and more, people are recording their food consumption, vital signs and activity. Worldwide, more than one billion people use social networks. As the Internet of Things grows, so too will knowledge about the state of health-relevant parts of the constructed landscape. Smartphones that can feed all these data into databases and analytic algorithms already are everywhere. “It is stunning,” Patrick said.

“We are creating digital tracks of everyday life and a diverse ecosystem of devices is generating massive amounts of data,” continued Patrick, who is the Principal Investigator of the Health Data Exploration Project supported by the Robert Wood Johnson Foundation. Among the questions the project is addressing are how data can be better used to promote the public good and what kinds of models of inquiry will lead to data harvests that are most consequential for health issues.

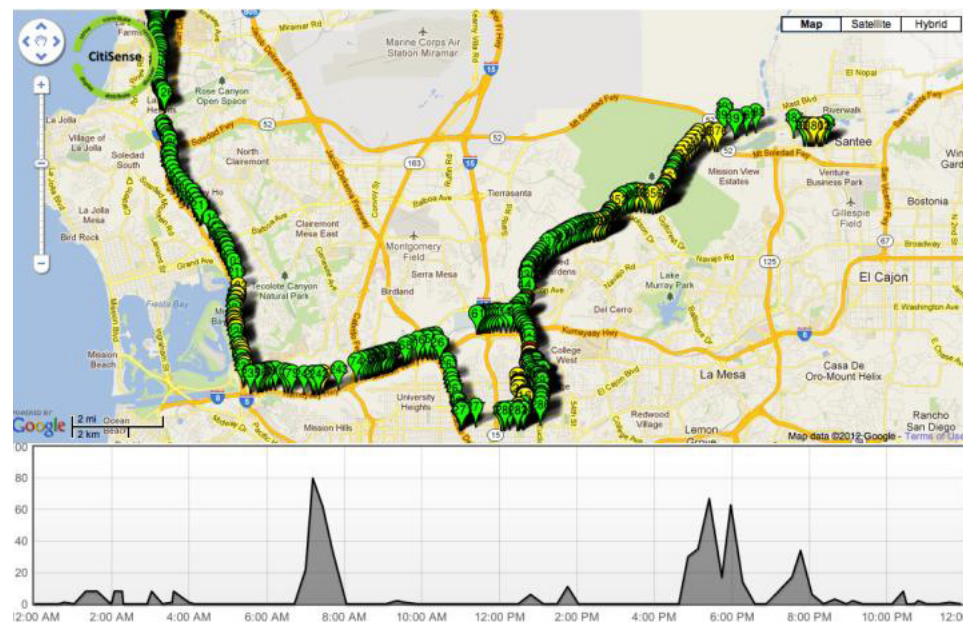


Figure 7. CitiSense System Overview

Echoing concerns raised by others at the VASEM Summit, Patrick spoke about the need to deal with vexing challenges to acquiring and using data in new and more extensive ways. For one, he noted, “we will be dealing with the haves and the have-nots,” that is, disparities within communities in how beneficial a more data-driven health system will be. There is also a large-scale ethical issue at play, he noted, pointing to the runaway costs of healthcare. By 2025, he said, the costs could grow from the already outrageous \$3 trillion annual expense of today to more than \$5.2 trillion.

We are creating digital tracks of everyday life and a diverse ecosystem of devices is generating massive amounts of data.

~ Kevin Patrick

Environmental components of health constitute another aspect of the dynamic for which better use of more data ought to improve public health. That is where another project Patrick works on, called CitiSense, could make a difference (Figure 7). The basic idea of the NSF-funded program is to fit mobile phones with pollution sensor boards (designed by UCSD engineering students) for “always-on participatory sensing for air quality.” Highlighting the importance of this type of sensing, Patrick noted that in 2012 diesel exhaust was designated a carcinogen by the World Health Organization. Other sensors enable the phones to measure carbon monoxide, nitrogen emissions, ozone and other chemicals of concern.

“We are swimming in these,” Patrick said, referring to the pollutants. “So we made a sensor board, connected to cell phones, and put them into circulation.” By importing the data into the Cloud and applying visualization methods, people can develop knowledge of the exposures they face. It is possible too to track exposure in a particular person during a particular outing, but the data also are fully shareable. Remarked Patrick, if you uncouple your health from where you live, you will miss critical environmental determinants.

“Data from just a few users can lead to higher granularity exposure maps,” Patrick said. This is how Big Data can be translated into solutions for individuals who, for example, might choose alternative routes or travel times based on the exposure data.

Entirely new categories of health-relevant data will come on-line. One category that Patrick finds tantalizing is based on a growing understanding of the relationship between health and a person’s microbiome—the trillions of bacteria that live on and in a human body. “Ten years ago, there was no ability to understand this, but now we can,” said Patrick. “These new data may become as important as genomic information for health.”

In time, he continued, it might be possible to integrate the entire ecosystem of data—whether they are genomic or microbiome data, environmental exposure data or behavior data—to provide highly personalized guidance about what communities and individuals should do for the cause of achieving and maintaining good health.

A step toward that direction is underway with the National Science Foundation (NSF)-funded DELPHI (Data e-Platform

to Leverage Multilevel Personal Health Information) project, Patrick said. Medical, personal health, public health, environmental, genomic, microbiomic, behavioral and social determinant data are all on the project’s agenda. “If you are in a high violence area, you will not go outside to exercise,” Patrick offers as an illustration of the sort of unconventional and fact-of-life data that DELPHI will deploy. “If your healthcare providers know that, they might want to develop an indoor exercise regimen for you.”

The more general goal of DELPHI, he said, is “the creation of a ‘Whole Health Information Platform’ that takes into account everything from the genome to the exposome.” For someone with asthma, for example, the DELPHI system might combine air quality data, personal data like self-reported peak airflow, and hospitalization and other medical records. Those data would go through a layer of analytic algorithms and protocols to determine such things as the probability of danger of an asthma attack for a particular person in his or her particular environmental setting, whether a doctor should be alerted, and if medication is needed.

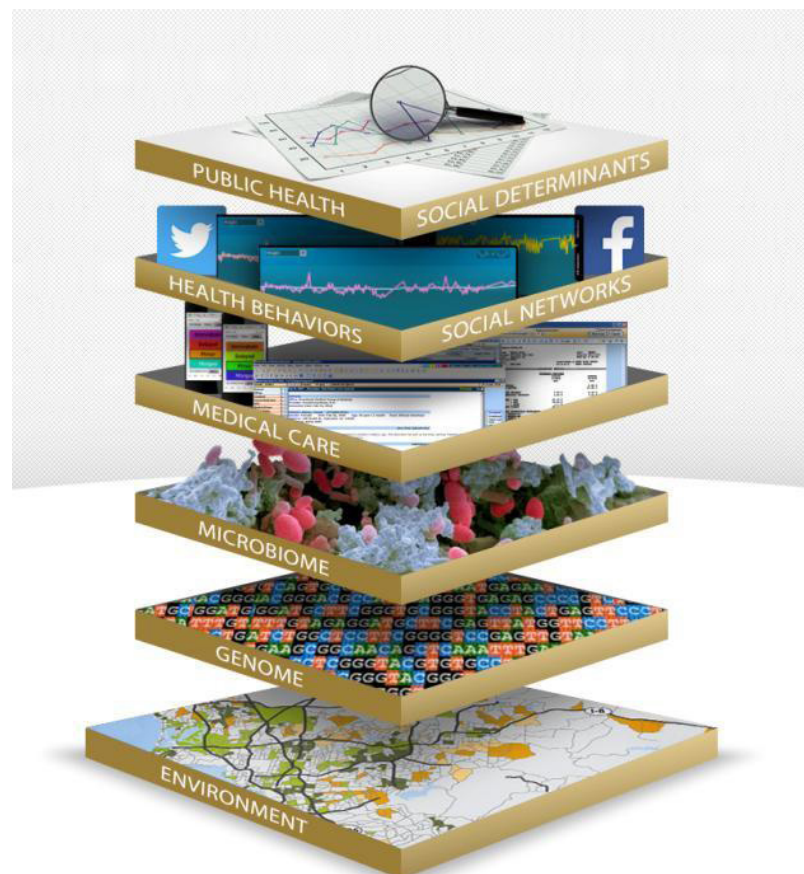


Figure 8. List of major influences on health

Big Data for the People of Virginia

In her closing remarks at the 2014 VASEM Summit on Big Data, Summit organizer [Sallie Keller, a Professor of Statistics at Virginia Tech and Director of Virginia Bioinformatics Institute's Social Decision Analytics Laboratory](#), reminded the gathering that "there is a lot of work ahead of us." Keller said she looked forward to the collaborations the Summit might catalyze and to develop the Virginia Academy of Science, Engineering and Medicine, VASEM, into a body that proves to be routinely valuable to the Commonwealth of Virginia. Earlier during the Summit, Virginia's Secretary of Technology, Karen Jackson, read a letter that Virginia Governor Terry McAuliffe wrote for the Summit that presaged Keller's sentiment. "Big Data plays a critical role in amplifying the success of technology, business, health, medicine and science," the Governor wrote. That list covers a lot of territory, but the Governor also expects the new era of Big Data to help him do his job. "The reliable practice and usage of data can promote effective, evidence-based decision-making throughout Virginia." The ever-growing amount and scope of medically-related data—from sources that range in scale from the molecular (think genomic and blood tests) to the societal (think traffic and air quality monitoring) all of the way to global sources of, say epidemiological trends and adverse drug response occurrences—is opening a pathway to what Patrick and his colleagues envision will be a new era of health and medicine.