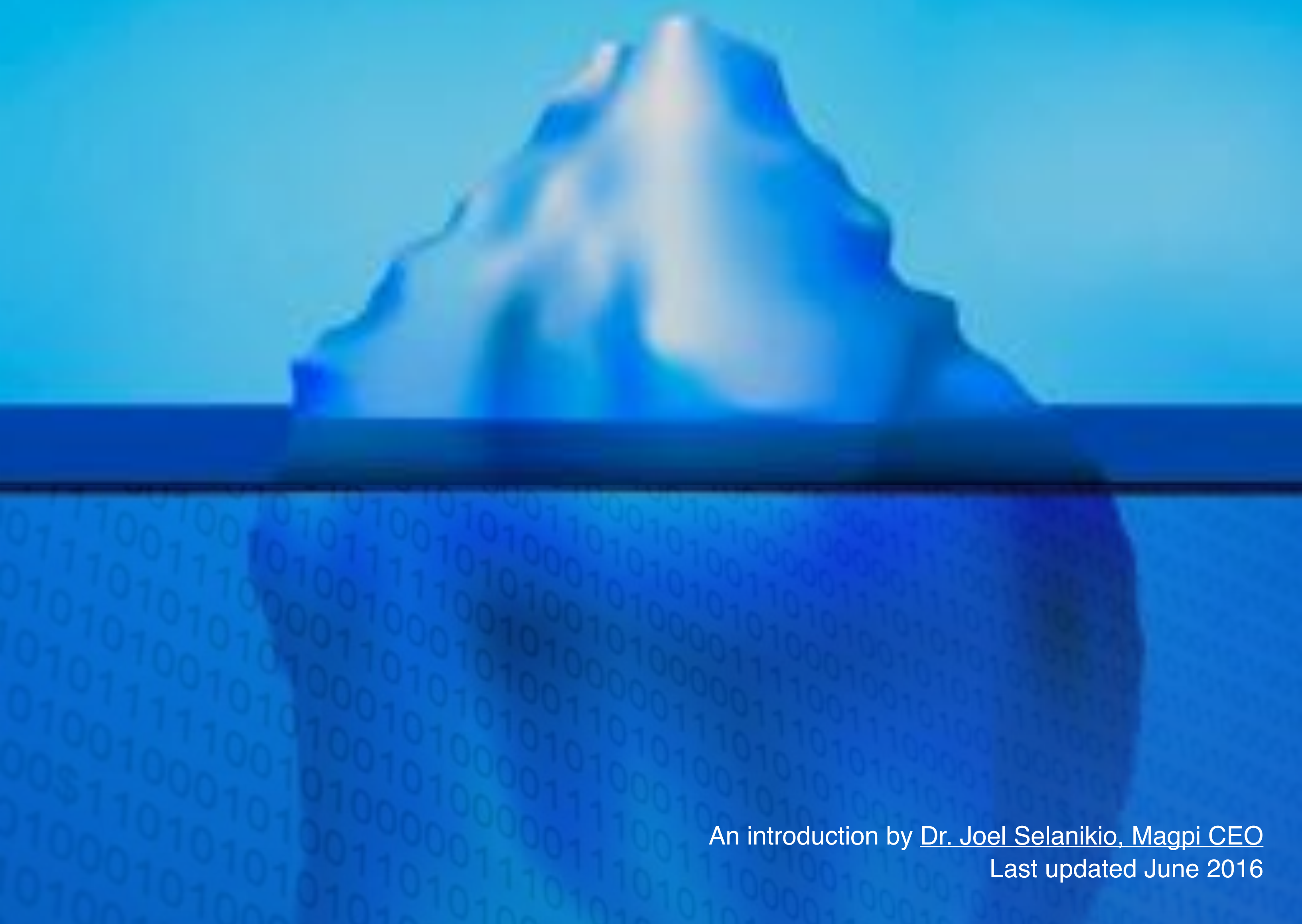
BIG DATA AND HEALTH

What Is Big Data, How Did It Happen, and What Does It Mean for Health?



An introduction by Dr. Joel Selanikio, Magpi CEO
Last updated June 2016

Table of Contents

What Does Big Data Mean?

There are almost as many definitions of big data as there are people writing about it:

"data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges"

"data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications"

"datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze"

No matter what the definition used, the key elements of big data are:

- increasing amounts of data being collected
- increasing connection of datasets
- increasing innovation to manage this data

Of course, there has always been a tremendous amount of information in the world because there has always been a tremendous amount of activity in the world. It is only as we have become better at collecting this information and recording it (thereby rendering it as "data"), communicating about it, and connecting it to other data that the term "big data" could really be considered appropriate.

Nonetheless, the idea that collected data is outstripping our ability to manage it — or at least forcing us to develop new methodologies for managing it — is not a new one. In fact, as early as 1880, it was noted that using the information technologies available at the time it would take 7 years to process the census data. This made it a certainty, given the growing population, that the 1890 census would not be analyzable within the ten year period before the 1900 census began. This prediction led the US government to hire Herman Hollerith to develop a punched card tabulation system to speed up the process.[1]

Hollerith was living in a great age of invention, however, and his tabulating machine was only one of many developments that have brought us to our current stage.

The idea that collected data is outstripping our ability to manage it — or at least forcing us to develop new methodologies for managing it — is not a new one.

[1] https://en.wikipedia.org/wiki/1880_United_States_Census
Hollerith's company, the Tabulating Machine Company, was later joined with several others to create what would become International Business Machines (IBM) in 1924.

How Did Big Data Happen?

Big data can be considered to be the result in large part of enormous advances in three interconnected fields:

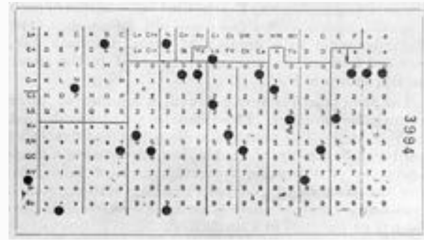
- Storage
- Connectivity
- Computing

It is possible to choose some representative advances — and inventors — in each of these fields, while at the same time emphasizing that these are just that: representative. For every person included, many are necessarily left out.

Storage

Let's begin with the development of modern data storage systems. As mentioned briefly above, in the 1880s the US and other governments were catalyzing development of tabulation systems for the purpose of managing the biggest datasets of that day. Those systems, though, were based on punched paper cards like those used by Herman Hollerith. The standard that developed was a punched card with 80 columns that stored about 70 bytes of data.[2]

Use of punched card systems by large businesses and the government spurred interest in more "data dense" means of data storage that might be more easily secured (much of the 1890 census data, the first to be tabulated using automated means, was destroyed by fire in 1921).



Hollerith punched card, 1895

In 1898, a Danish engineer named Valdemar Poulsen demonstrated a method of storing sound by means of magnetic wire, and later tape. This technology was adapted for storage of tabulation data, and by the end of the 1940s magnetic storage was the predominant mode of computer storage, and continues to be one of the dominant digital storage methods used today.

According to Parkinson's Law, "work expands so as to fill the time available for its completion", and it has been noted that data storage has its corollary: data expands to fill the space available for storage. This has certainly been the case in the last 4 decades: it is estimated that from the dawn of man until the 20th century a total of 5 exabytes of data (1 exabyte = 1 billion gigabytes) had ever been recorded worldwide,

[2] Note: Hollerith's punchcard above comprises ~150k bytes of data – the equivalent of more than 2000 cards worth of information.

but that we are now recording that same amount every few days. Luckily, advances in magnetic storage technology, along with other tech advances, have led to a dramatic plunge in storage costs. In fact, the price of one gigabyte of computer storage has fallen from approximately $700,000 U.S. dollars back in 1980, to less than one single dollar today!

Connectivity

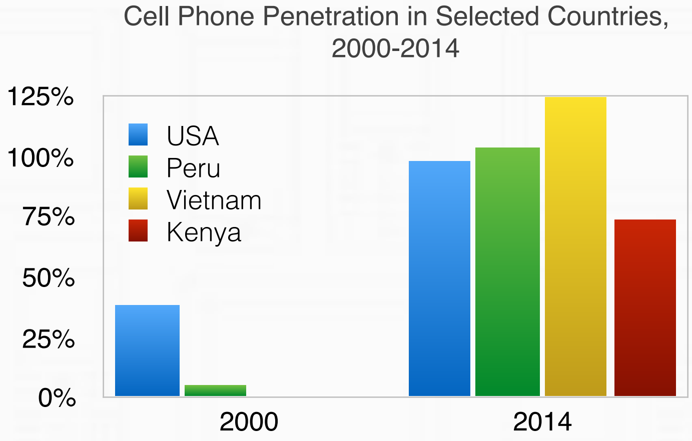
Martin Cooper invented the cellular telephone in 1973 while working for Motorola. The first commercial model, called the DynaTac 8000, was made available in 1984. It weighed just under 2 pounds (825g), took 10 hours to charge, offered 30 minutes of talk time, and cost the equivalent of about $10,000.

It took more than 20 years for miniaturization technology, cellular networks, economies of scale, and other related developments to create the near-universal current use of cellular phone technology.

As seen in the graph at right, as late as 2000 there was little or no cell phone penetration in poorer countries like Kenya and Vietnam, but currently it is estimated that more than 50% of the population in even the poorest countries has a mobile phone, or use of one.[3]

Economists from the Deloitte consultancy have estimated[4] that each this mobile utilization has tremendous effects on economic growth:

- A doubling of mobile data use leads to a growth in the GDP per capita growth rate of 0.5 percentage points
- Countries characterized by a higher level of data usage per 3G connection have seen an increase in their GDP per capita growth of up to 1.4 percentage points.
- A 10% rise in 3G penetration increases GDP per capita growth by 0.15 percentage points. In developing markets, a 10% expansion in mobile penetration increases productivity in the long run by 4.2 percentage points.



[3] Note that “penetration” displayed on the graph is measured by the number of SIM cards divided by the population, which overestimates the percentage of the population owning their own mobile phone (since many SIM cards are not yet assigned, and many wealthier individuals may have more than one SIM card). Data: http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx

[4] Report for the GSM Association: “The impact of mobile on economic growth". http://www2.deloitte.com/uk/en/pages/technology-media-and-telecommunications/articles/impact-of-mobile-telephony-on-economic-growth.html

Computing

As anyone with a smartphone knows, in the last decade Martin Cooper's mobile invention has merged inextricably with the "computer". Compare the IBM PC XT, the dominant business computing platform of the mid-1980s, with the Alcatel One Touch, one of the *least* capable and *least* expensive phones at the time of this writing.

The $10 Alcatel is fifty times faster, with about thirty-thousand times more memory.

And a built-in FM radio and flashlight.

In fact, an iPhone of today is far more powerful than the multi-million-dollar Cray supercomputers of just thirty years ago. Just as amazing as the remarkable increase in power has been the diminishing size of components, and the dramatic decrease in the price – and the fact that making telephone calls is a fast-diminishing use of such a "telephone," and usage shifts more to the computing side and away from telephony.[5]

1985

IBM
PC XT
128 KB
4.8 MHz
$3000

2015

Alcatel
One Touch
4GB
250 MHz
$10

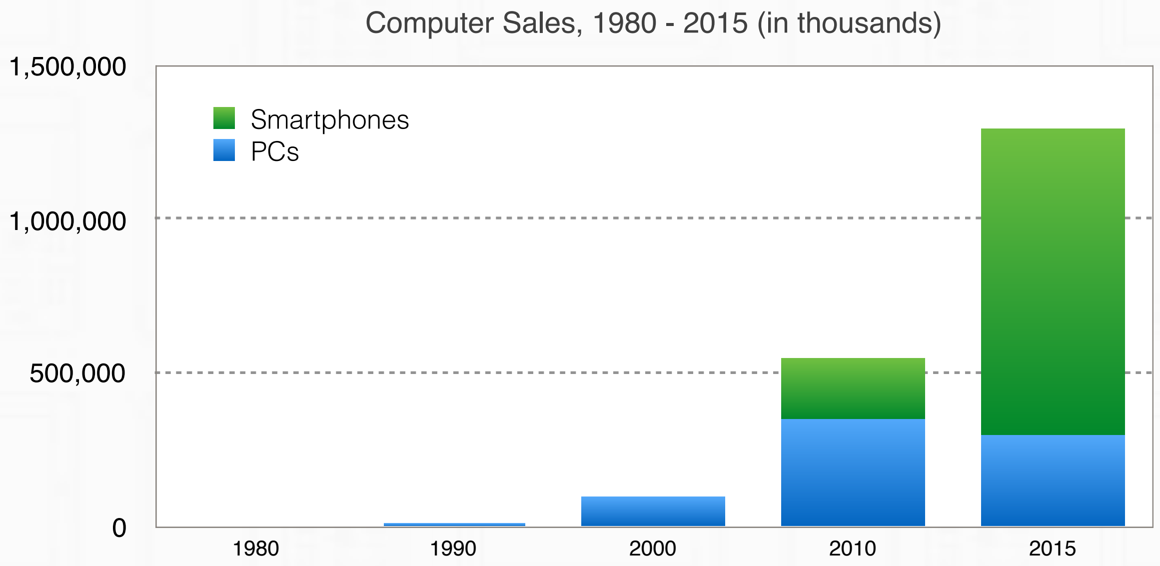
The trend towards smaller and cheaper electronics has been going on for a long time, and in the 1970s it finally produced computers that could be affordable not just for large institutions like the military or General Motors, but for individual consumers.

The Apple 1, 1976 (photo credit: Ed Uthman)

The Apple 1 (shown above), a device handmade by Steve Wozniak, one of the founders of Apple Computer, was arguably the first personal computer, aimed squarely at individuals for home computer purposes. An enormous hit in 1976, it helped to create the consumer computing market that later overlapped with business computing in the form of IBM's famed PC.

[5] http://www.theguardian.com/technology/2012/jul/18/ofcom-report-phone-calls-decline

Though there has been recent speculation and angst about the death of the PC – with sales of desktops and laptops appearing to have peaked in 2010 – it is clear when smartphones are added to the figures (see computer sales graph) that the personal computer has simply gotten much smaller and more mobile.



Big Data Use in Health and Medicine

In June 2011, the McKinsey & Company consulting company released a report "Big data: The next frontier for innovation, competition, and productivity".[6] What's remarkable just four years later is how much the report sets big data in the future, particularly in the section on big data in medicine.

In the nearly five years since the report was published, big data has come to be widely used in commerce, but is still in its infancy in medicine and healthcare. Below are a few examples of concrete initiatives currently moving forward in the field.

CDC Improving Ebola Control with Basic SMS Reporting

2014 and 2015 saw the largest outbreak of Ebola virus in history, in the region of West Africa. More than 10,000 deaths have been reported, and approximately 30,000 cases. Over the course of this event, surveillance of cases and possible cases has been difficult. Poor organization and communications in the region have combined with the loss of life among the very health worker population that might have been relied upon to help with this process.

Additionally, initial surveillance efforts involved the use of a small number of surveillance workers submitting data via paper forms on a weekly basis, and the inefficiencies in this process — still a very common approach in global public health — were felt to be simply not up to the massive task at hand. The result has been a poor, and delayed, understanding of the waxing and waning of the epidemic across the region.

The US Centers for Disease Control and Prevention (CDC), is working to improve "situational awareness"by operating a much more widespread surveillance network in some of the affected countries. This "zero reporting" approach exchanges paper forms for basic text messaging, with hundreds of health reporters in several countries sending a single daily text message sent into systems like Magpi, usually from very inexpensive cell phones. The message consists of a code word for the system (e.g. "ebola") and a number representing the number of suspected cases observed by that reporter for the previous 24 hours.

[6] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

This text-based approach combines a reduction in the amount of data per report (moving from an entire paper form to a single digit representing the number of suspected cases plus the location information where the data originated), but a dramatic increase in the number of reports.

This is a common approach in the new world of big data: accepting less complete data, or perhaps even less rigorously-randomly-sampled data, in exchange for much larger datasets:

> When the quantity of data is vastly larger and is of a new type, exactitude in some cases is no longer the goal so long as we can divine the general trend. Moving to a large scale changes not only the expectations of precision but the practical ability to achieve exactitude. Though it may seem counterintuitive at first, treating data as something imperfect and imprecise lets us make superior forecasts, and thus understand our world better. [7]

Certainly, it is impossible to argue, in the case of the Ebola surveillance, that blanketing the area with so many more sensors — no matter how simple the data they are reporting — will result in a less sensitive system. And in this case, the health reporters play the role of the fire alarm, not the fire department's investigation team: their role is simple to sound the alarm, and to cover as much of the community as possible. Sensitivity is the name of this game.

[7] Mayer-Schönberger, Viktor; Cukier, Kenneth (2013-03-05). Big Data: A Revolution That Will Transform How We Live, Work, and Think (Kindle Locations 619-622). Houghton Mifflin Harcourt. Kindle Edition.

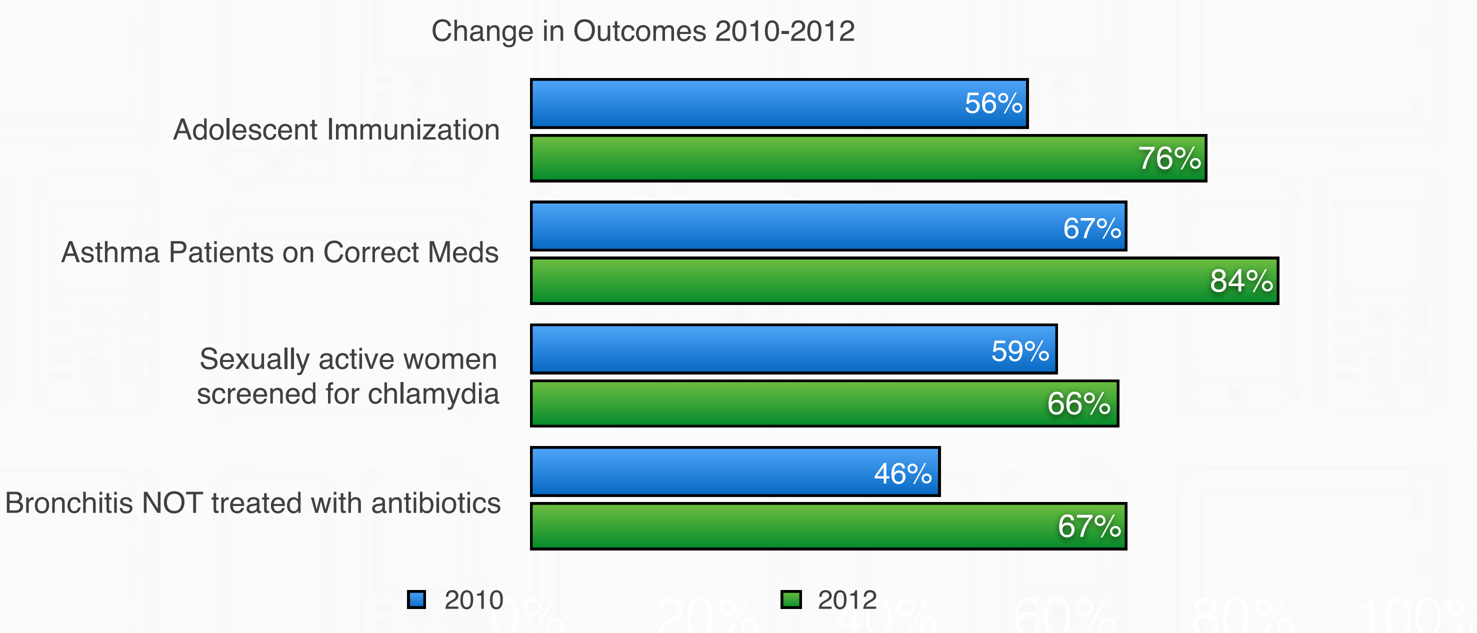
MemorialCare Health System Data Tracking System

> Technology is making it easier to monitor doctors' work as patients' details are compiled electronically instead of on paper charts. Software makers are selling new tools to crunch the data. Software called Crimson offered by the Advisory Board Co. now includes information on more than a half-million doctors, up from fewer than 50,000 in 2009.
>
> *– "Hospitals Prescribe Big Data to Track Doctors at Work", Wall Street Journal, July 11, 2013.*

As reported in the Wall Street Journal[8], MemorialCare Health System, a six-hospital nonprofit organization in California, is tracking and analyzing detailed information on physician performance at its facilities, including looking at rates of adolescent immunization, asthma patients on the correct medicines, patients with viral bronchitis treated (inappropriately) with antibiotics, and sexually active women screened for chlamydia.

MemorialCare makes the data available to each doctor, showing them how their practice activities, and outcomes, compare to their peers. While some doctors respond better to the new data than others, most are responsive to change if they see that they are not performing as well as their colleagues.

As shown in the graph below, MemorialCare saw sharp improvements in four metrics over the course of just two years: the rate of adolescent immunization (e.g. with HPV), the percentage of asthma patients currently taking the correct drug regimen, the percentage of sexually active women being screened for chlamydia, and the percentage of cases of probably viral bronchitis that were *not* treated with antibiotics.



[8] http://www.wsj.com/news/articles/SB10001424127887323551004578441154292068308

ResearchKit: A Mobile Tech Tool to Spur Medical Research

Announced in March 2015, Apple's ResearchKit is an open-source software framework that allows programmers to use the accelerometers and other sensors on iOS devices to create medical diagnostic and research apps. The idea is to use these very common devices to spur phone users to join medical studies, for their benefit and the benefit of others.

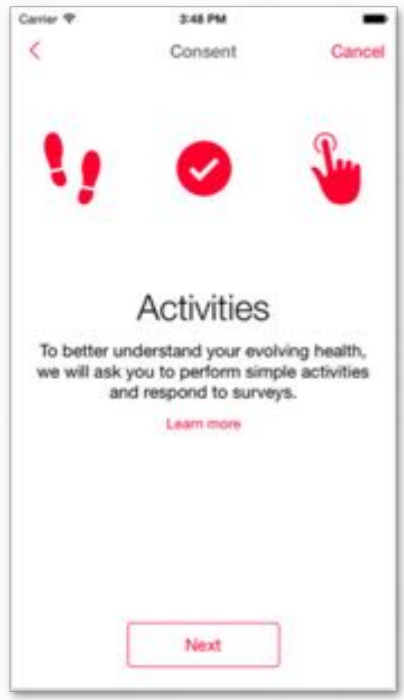
mPower Parkinson's Disease App

One of the first apps available utilizing ResearchKit was mPower, a "mobile Parkinson Disease Study," launched by Sage Bioworks, a nonprofit research organization:

> The mPower application uses a mix of surveys and tasks that activate phone sensors to collect and track health and symptoms of PD [Parkinson Disease] progression - like dexterity, balance or gait. Our goals are to learn about the variations of PD, to improve the way we describe and manage these variations, and to learn if mobile devices and sensors can help measure PD and its progression to ultimately improve the quality of life for people with PD.[9]

Amazingly, in the first 6 months more than 14,000 people have signed up for the app:

> Dr. Ray Dorsey, a neurologist at the University of Rochester who was involved in the creation of the app, said 14,000 people have downloaded it to date. "10 to15 percent of them have Parkinson's, so in the span of six months this is one of the largest clinical studies of Parkinson's ever conducted," he told ZDNet.[10]



Screenshot of mPower

EpiWatch Epilepsy Tracking App

Another early adopter of the ResearchKit framework is Johns Hopkins University, a leading research institution in Baltimore, USA. Released in 2015, Johns Hopkins' EpiWatch application utilizes ResearchKit on both the iPhone and the Apple Watch in order to collect data from epilepsy patients before, during, and after their seizures. As with mPower, EpiWatch uses an iPhone app to gain informed consent,

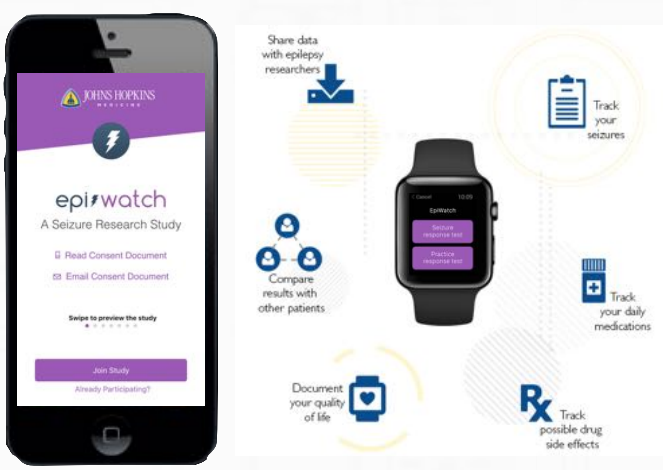
[9] http://parkinsonmpower.org/

[10] http://www.zdnet.com/article/move-over-healthkit-why-apples-researchkit-is-proving-the-real-hit-with-doctors/

dramatically simplifying the process of getting patients registered as participants in the ongoing study.

Once the patients are registered, they are able to activate the Apple Watch app to record a variety of measures (including heart rate, movement, etc. during a seizure). This is obviously more appropriate for patients who have a pre-seizure “aura” and can then start the watch app when they sense a seizure is coming.

> “The app then starts recording heart rate and movements. The app also requests patient participation in a specialized memory game to evaluate patient responsiveness during the seizure. It is the first medical research app to include a cognitive test of this nature…
>
> The app makes use of many of the Apple Watch sensors—including the accelerometer, which detects movements, and the gyroscope, which determines orientation in space—to measure and record movements and falls during seizures. The app also uses Apple Watch's heart rate monitor, since heart rate can rise significantly during seizures.[11]



IBM and NIH Working Together with Big Data

One of the most promising developments in big data has been the advent of IBM’s Watson technology, which seeks to allow researchers to much more easily find the answers to questions when those answers require multiple large and complex datasets. Part of Watson’s appeal is that it allows researchers to ask questions in “natural language” rather than having to learn a computer query language syntax.

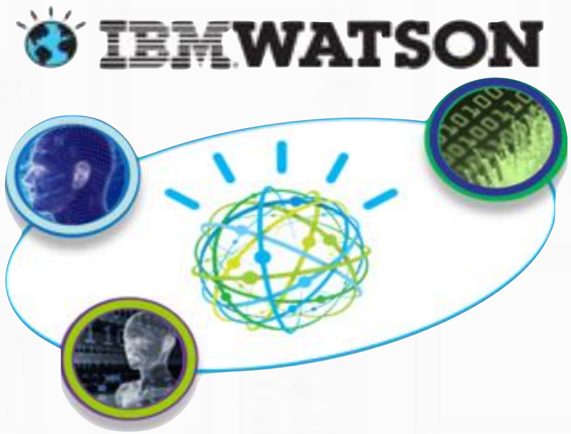
Watson has already been applied to many tasks, and was initially tested in 2008 against human contestants on the TV game show Jeopardy in answering a very wide variety of questions requiring access to very large databases. Watson beat both human champion players.

Since that time, IBM has been seeking to apply this technology within a wide range of medical applications. For example, oncologists at the Maine Center for Cancer Medicine began using Watson technology in 2013 in order to recommend cancer treatments based on large databases of up-to-date information:

> Thanks to its computing power Watson can sift through 1.5 million patient records and histories to provide treatment options in a matter of seconds based on previous treatment outcomes and patient histories. It has been fed with more

[11] http://hub.jhu.edu/2015/10/15/apple-watch-epiwatch

than 600,000 pieces of medical evidence, 2 million pages of text from 42 medical journals and clinical trials in the area of oncology research, IBM said.[12]



Seeking to capitalize on this groundbreaking approach to making large datasets more accessible, the U.S. National Institutes of Health — the preeminent U.S. government medical research organization, which oversees an annual $31 billion budget — is now working with IBM to connect a very wide variety of clinical and research datasets to the Watson system. This enables dramatically easier and faster querying of the data.[13]

Conclusion

Many recent technology advances in storage, computing, and networking — along with progress in wearable computing — have been brought together to enable the compilation, combination, and analysis of very large datasets in ways that were simply not possible before.

The field of medicine is only now beginning to grasp the potential for benefit to patients from big data tools, and only now beginning to use the tools on offer from large technology companies like IBM and Apple. By utilizing better data collection technologies, wearable computers (which can monitor patient vital signs continuously and for much lower cost than previously), better data visualization and analytics, a revolution is occurring in our ability as health workers, as researchers, and as patients to gather clinical information, to connect it to other information, and to put it to use in improving outcomes and growing our understanding.

For many public health and medical fields, these ideas have barely been applied — and this provides an incredible opportunity to view the many, many analogous examples of big data applied to commerce, and replicate the benefits for the sake of our patients, our practices, and our communities.

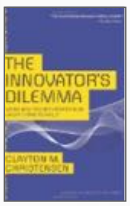
[12] http://in.reuters.com/article/2013/02/08/ibm-watson-cancer-idINDEE9170G120130208

[13] https://ibmecm.cloudant.com/wcm70x/_design/main/_show/detail/ECC-IMC14949USEN?instructions=false&lnk=ushpv18ce2

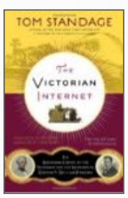
Resources and Further Reading

Big data timeline
http://www.winshuttle.com/big-data-timeline/

Big Data: A Revolution That Will Transform How We Live, Work, and Think
by Viktor Mayer-Schönberger and Kenneth Cukier
http://www.amazon.com/gp/product/0544002695



The Innovator's Dilemma
by Clayton Christiansen
http://www.amazon.com/Innovators-Dilemma-Technologies-Management-Innovation/dp/142219602X/



The Victorian Internet: The Remarkable Story of the Telegraph and the Nineteenth Century's On-line Pioneers
by Tom Standage
http://www.amazon.com/gp/product/162040592X

Big data: The next frontier for innovation, competition, and productivity
by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers
http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation/

TED Talk: The Surprising Seeds of a Big Data Revolution in Health
by Joel Selanikio
http://www.ted.com/talks/joel_selanikio_the_surprising_seeds_of_a_big_data_revolution_in_healthcare

About the Author

Dr. Joel Selanikio is an award-winning physician, inventor, emergency responder, and public speaker who leads the efforts of Magpi to develop and promote new technologies and business models for health and international development, including multiple-award-winning Magpi mobile data collection and messaging software – the most widely scaled mobile technology ever created for international development.

In December 2014 - January 2015, he was the lead physician at the IMC Ebola Treatment Center at Lunsar, Sierra Leone.

Dr. Selanikio is a frequent keynote speaker and consultant in the fields of social entrepreneurship, innovation, big data, public and global health, healthcare, child health, epidemiology, and the use of technology for development ("ICT4D") and for emergency and disaster preparedness and outbreak response, and has consulted or spoken on these topics at Davos, TEDx, Foo Camp, Harvard, MIT, Stanford, Google, DARPA, CNN, Fox News, the Clinton Global Initiative, the Royal Society of Medicine, and many other very prominent venues and organizations. He is a judge for the GSMA Global Mobile Awards, and is a winner of both the $100,000 Lemelson-MIT Award for Sustainability and the Wall Street Journal Technology Innovation Award.

Dr. Selanikio has been profiled by the Guardian, Wired, Forbes, The Economist, The Wall Street Journal, Fox News, the BBC, NPR, Information Week, and the Washington Post, among many other publications. He is a practicing pediatrician, and is a former Wall Street computer consultant, and former CDC epidemiologist. As an officer of the Public Health Service, Dr. Selanikio served as Chief of Operations for the HHS Secretary's Emergency Command Center after the 9/11 attacks. In 2005, he was given the Haverford Award for Humanitarian Service for his work in treating tsunami victims in Aceh.

View Joel Selanikio's profile