Forbes

Gil Press, Contributor
I write about technology, entrepreneurs and innovation.

TECH | 5/09/2013 @ 9:45AM | 33,881 views

A Very Short History Of Big Data

The story of how data became big starts many years before the current buzz around big data. Already seventy years ago we encounter the first attempts to quantify the growth rate in the *volume of data* or what has popularly been known as the "information explosion" (a term first used in 1941, according to the *Oxford English Dictionary*). The following are the major milestones in the history of sizing data volumes plus other "firsts" in the evolution of the idea of "big data" and observations pertaining to data or information explosion.

Last Update: December 21, 2013

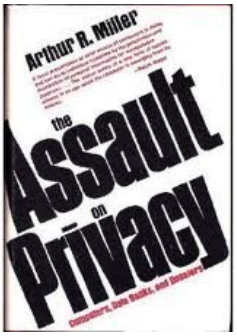
1944 Fremont Rider, Wesleyan University Librarian, publishes *The Scholar and the Future of the Research Library*. He estimates that American university libraries were doubling in size every sixteen years. Given this growth rate, Rider speculates that the Yale Library in 2040 will have "approximately 200,000,000 volumes, which will occupy over 6,000 miles of shelves... [requiring] a cataloging staff of over six thousand persons."

1961 Derek Price publishes *Science Since Babylon*, in which he charts the growth of scientific knowledge by looking at the growth in the number of scientific journals and papers. He concludes that the number of new journals has grown exponentially rather than linearly, doubling every fifteen years and increasing by a factor of ten during every half-century. Price calls this the "law of exponential increase," explaining that "each [scientific] advance generates a new series of advances at a reasonably constant birth rate, so that the number of births is strictly proportional to the size of the population of discoveries at any given time."

November 1967 B. A. Marron and P. A. D. de Maine publish "Automatic data compression" in the *Communications of the ACM*, stating that "The 'information explosion' noted in recent years makes it essential that storage requirements for all information be kept to a minimum." The paper describes "a fully automatic and rapid three-part compressor which can be used with 'any' body of information to greatly reduce slow external storage requirements and to increase the rate of information transmission through a computer."

1971 Arthur Miller writes in *The Assault on Privacy* that "Too many information handlers seem to measure a man by the number of bits of storage capacity his dossier will occupy."

1975 The Ministry of Posts and Telecommunications in Japan starts conducting the Information Flow Census, tracking the volume of information circulating in Japan (the idea was first suggested in a 1969 paper). The census introduces "amount of words" as the unifying unit of measurement across all



media. The 1975 census already finds that information supply is increasing much faster than information consumption and in 1978 it reports that "the demand for information provided by mass media, which are one-way communication, has become stagnant, and the demand for information provided by personal telecommunications media, which are characterized by two-way communications, has drastically increased.... Our society is moving toward a new stage... in which more priority is placed on segmented, more detailed information to meet individual needs, instead of conventional mass-reproduced conformed information." [Translated in Alistair D. Duff 2000; see also Martin Hilbert 2012 (PDF)]

April 1980 I.A. Tjomsland gives a talk titled "Where Do We Go From Here?" at the Fourth IEEE Symposium on Mass Storage Systems, in which he says "Those associated with storage devices long ago realized that Parkinson's First Law may be paraphrased to describe our industry—'Data expands to fill the space available'.... I believe that large amounts of data are being retained because users have no way of identifying obsolete data; the penalties for storing obsolete data are less apparent than are the penalties for discarding potentially useful data."

1981 The Hungarian Central Statistics Office starts a research project to account for the country's information industries, including measuring information volume in bits. The research continues to this day. In 1993, Istvan Dienes, chief scientist of the Hungarian Central Statistics Office, compiles a manual for a standard system of national information accounts. [See Istvan Dienes 1994 (PDF), and Martin Hilbert 2012 (PDF)]

August 1983 Ithiel de Sola Pool publishes "Tracking the Flow of Information" in *Science*. Looking at growth trends in 17 major communications media from 1960 to 1977, he concludes that "words made available to Americans (over the age of 10) through these media grew at a rate of 8.9 percent per year... words actually attended to from those media grew at just 2.9 percent per year.... In the period of observation, much of the growth in the flow of information was due to the growth in broadcasting... But toward the end of that period [1977] the situation was changing: point-to-point media were growing faster than broadcasting." Pool, Inose, Takasaki and Hurwitz follow in 1984 with *Communications Flows: A Census in the United States and Japan*, a book comparing the volumes of information produced in the United States and Japan.

July 1986 Hal B. Becker publishes "Can users really absorb data at today's rates? Tomorrow's?" in *Data Communications*. Becker estimates that "the recoding density achieved by Gutenberg was approximately 500 symbols (characters) per cubic inch—500 times the density of [4,000 B.C. Sumerian] clay tablets. By the year 2000, semiconductor random access memory should be storing 1.25X10^11 bytes per cubic inch."

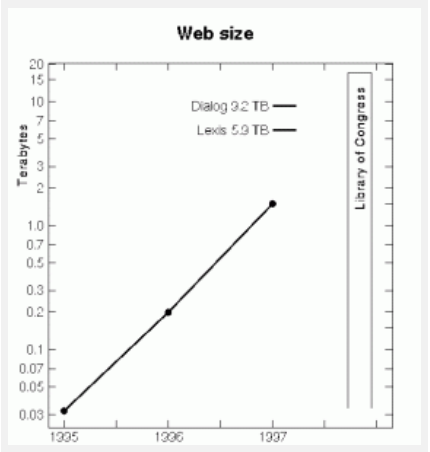
September 1990 Peter J. Denning publishes "Saving All the Bits" (PDF) in *American Scientist*. Says Denning: "The imperative [for scientists] to save all the bits forces us into an impossible situation: The rate and volume of information flow overwhelm our networks, storage devices and retrieval systems, as well as the human capacity for comprehension... What machines can we build that will monitor the data stream of an instrument, or sift through a database of recordings, and propose for us a statistical summary of what's there?... it is possible to build machines that can recognize or predict patterns in data without understanding the meaning of the patterns. Such machines may eventually be fast enough to deal with large data streams in

real time… With these machines, we can significantly reduce the number of bits that must be saved, and we can reduce the hazard of losing latent discoveries from burial in an immense database. The same machines can also pore through existing databases looking for patterns and forming class descriptions for the bits that we've already saved."

1996 Digital storage becomes more cost-effective for storing data than paper according to R.J.T. Morris and B.J. Truskowski, in "The Evolution of Storage Systems," *IBM Systems Journal*, July 1, 2003.

October 1997 Michael Cox and David Ellsworth publish "Application-controlled demand paging for out-of-core visualization" in the Proceedings of the IEEE 8th conference on Visualization. They start the article with "Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*. When data sets do not fit in main memory (*in core*), or when they do not fit even on local disk, the most common solution is to acquire more resources." It is the first article in the ACM digital library to use the term "big data."

1997 Michael Lesk publishes "How much information is there in the world?" Lesk concludes that "There may be a few thousand petabytes of information all told; and the production of tape and disk will reach that level by the year 2000. So in only a few years, (a) we will be able [to] save everything–no information will have to be thrown out, and (b) the typical piece of information will never be looked at by a human being."



Source: Michael Lesk

April 1998 John R. Masey, Chief Scientist at SGI, presents at a USENIX meeting a paper titled "Big Data… and the Next Wave of Infrastress."

October 1998 K.G. Coffman and Andrew Odlyzko publish "The Size and Growth Rate of the Internet." They conclude that "the growth rate of traffic on the public Internet, while lower than is often cited, is still about 100% per year, much higher than for traffic on other networks. Hence, if present growth trends continue, data traffic in the U. S. will overtake voice traffic around the year 2002 and will be dominated by the Internet." Odlyzko later established the Minnesota Internet Traffic Studies (MINTS), tracking the growth in Internet traffic from 2002 to 2009.

August 1999 Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, and Robert Haimes publish "Visually exploring gigabyte data sets in real time" in the *Communications of the ACM*. It is the first CACM article to use the term "Big Data" (the title of one of the article's sections is "Big Data for Scientific Visualization"). The article opens with the following statement: "Very powerful computers are a blessing to many fields of inquiry. They are also a curse; fast computations spew out massive amounts of data. Where megabyte data sets were once considered large, we now find data sets from individual simulations in the 300GB range. But understanding the data resulting from high-end computations is a significant endeavor. As more than one scientist has put it, it is just plain difficult to look at all the numbers. And as Richard W. Hamming, mathematician and pioneer computer scientist, pointed out, the purpose of computing is insight, not numbers."

October 1999 Bryson, Kenwright and Haimes join David Banks, Robert van

Liere, and Sam Uselton on a panel titled "Automation or interaction: what's best for big data?" at the IEEE 1999 conference on Visualization.

October 2000 Peter Lyman and Hal R. Varian at UC Berkeley publish "How Much Information?" It is the first comprehensive study to quantify, in computer storage terms, the total amount of new and original information (not counting copies) created in the world annually and stored in four physical media: paper, film, optical (CDs and DVDs), and magnetic. The study finds that in 1999, the world produced about 1.5 exabytes of unique information, or about 250 megabytes for every man, woman, and child on earth. It also finds that "a vast amount of unique information is created and stored by individuals" (what it calls the "democratization of data") and that "not only is digital information production the largest in total, it is also the most rapidly growing." Calling this finding "dominance of digital," Lyman and Varian state that "even today, most textual information is 'born digital,' and within a few years this will be true for images as well." A similar study conducted in 2003 by the same researchers found that the world produced about 5 exabytes of new information in 2002 and that 92% of the new information was stored on magnetic media, mostly in hard disks.

November 2000 Francis X. Diebold presents to the Eighth World Congress of the Econometric Society a paper titled "'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting (PDF)," in which he states "Recently, much good science, whether physical, biological, or social, has been forced to confront—and has often benefited from—the "Big Data" phenomenon. Big Data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology."

February 2001 Doug Laney, an analyst with the Meta Group, publishes a research note titled "3D Data Management: Controlling Data Volume, Velocity, and Variety." A decade later, the "3Vs" have become the generally-accepted three defining dimensions of big data, although the term itself does not appear in Laney's note.

September 2005 Tim O'Reilly publishes "What is Web 2.0" in which he asserts that "data is the next Intel inside." O'Reilly: "As Hal Varian remarked in a personal conversation last year, 'SQL is the new HTML.' Database management is a core competency of Web 2.0 companies, so much so that we have sometimes referred to these applications as 'infoware' rather than merely software."

March 2007 John F. Gantz, David Reinsel and other researchers at IDC release a white paper titled "The Expanding Digital Universe: A Forecast of Worldwide Information Growth through 2010 (PDF)." It is the first study to estimate and forecast the amount of digital data created and replicated each year. IDC estimates that in 2006, the world created 161 exabytes of data and forecasts that between 2006 and 2010, the information added annually to the digital universe will increase more than six fold to 988 exabytes, or doubling every 18 months. According to the 2010 (PDF) and 2012 (PDF) releases of the same study, the amount of digital data created annually surpassed this forecast, reaching 1227 exabytes in 2010, and growing to 2837 exabytes in 2012.

January 2008 Bret Swanson and George Gilder publish "Estimating the Exaflood (PDF)," in which they project that U.S. IP traffic could reach one

zettabyte by 2015 and that the U.S. Internet of 2015 will be at least 50 times larger than it was in 2006.

June 2008 Cisco releases the "Cisco Visual Networking Index – Forecast and Methodology, 2007–2012 (PDF)" part of an "ongoing initiative to track and forecast the impact of visual networking applications." It predicts that "IP traffic will nearly double every two years through 2012" and that it will reach half a zettabyte in 2012. The forecast held well, as Cisco's latest report (May 30, 2012) estimates IP traffic in 2012 at just over half a zettabyte and notes it "has increased eightfold over the past 5 years."

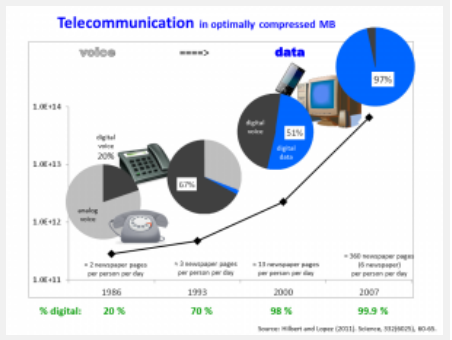
September 2008 A special issue of *Nature* on Big Data "examines what big data sets mean for contemporary science."

December 2008 Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska publish "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society (PDF)." They write: "Just as search engines have transformed how we access information, other forms of *big-data computing* can and will transform the activities of companies, scientific researchers, medical practitioners, and our nation's defense and intelligence operations.... Big-data computing is perhaps the biggest innovation in computing in the last decade. We have only begun to see its potential to collect, organize, and process data in all walks of life. A modest investment by the federal government could greatly accelerate its development and deployment."

December 2009 Roger E. Bohn and James E. Short publish "How Much Information? 2009 Report on American Consumers." The study finds that in 2008, "Americans consumed information for about 1.3 trillion hours, an average of almost 12 hours per day. Consumption totaled 3.6 Zettabytes and 10,845 trillion words, corresponding to 100,500 words and 34 gigabytes for an average person on an average day." Bohn, Short, and Chattanya Baru follow this up in January 2011 with "How Much Information? 2010 Report on Enterprise Server Information," in which they estimate that in 2008, "the world's servers processed 9.57 Zettabytes of information, almost 10 to the 22nd power, or ten million million gigabytes. This was 12 gigabytes of information daily for the average worker, or about 3 terabytes of information per worker per year. The world's companies on average processed 63 terabytes of information annually."

February 2010 Kenneth Cukier publishes in *The Economist* a Special Report titled, "Data, data everywhere." Writes Cukier: "...the world contains an unimaginably vast amount of digital information which is getting ever vaster more rapidly... The effect is being felt everywhere, from business to science, from governments to the arts. Scientists and computer engineers have coined a new term for the phenomenon: 'big data.'"

February 2011 Martin Hilbert and Priscila Lopez publish "The World's Technological Capacity to Store, Communicate, and Compute Information" in *Science*. They estimate that the world's information storage capacity grew at a compound annual growth rate of 25% per year between 1986 and 2007. They also estimate that in 1986, 99.2% of all storage capacity was analog, but in 2007, 94% of storage capacity was digital, a complete reversal of roles (in 2002, digital information storage surpassed non-digital for the first time).



Hilbert and Lopez 2011

May 2011 James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers of the McKinsey Global Institute publish "Big data: The next frontier for innovation, competition, and productivity." They estimate that "by 2009, nearly all sectors in the US economy had at least an average of 200 terabytes of stored data (twice the size of US retailer Wal-Mart's data warehouse in 1999) per company with more than 1,000 employees" and that the securities and investment services sector leads in terms of stored data per firm. In total, the study estimates that 7.4 exabytes of new data were stored by enterprises and 6.8 exabytes by consumers in 2010.

April 2012 The *International Journal of Communications* publishes a Special Section titled "Info Capacity" on the methodologies and findings of various studies measuring the volume of information. In "Tracking the flow of information into the home (PDF)," Neuman, Park, and Panek (following the methodology used by Japan's MPT and Pool above) estimate that the total media supply to U.S. homes has risen from around 50,000 minutes per day in 1960 to close to 900,000 in 2005. And looking at the ratio of supply to demand in 2005, they estimate that people in the U.S. are "approaching a thousand minutes of mediated content available for every minute available for consumption." In "International Production and Dissemination of Information (PDF)," Bounie and Gille (following Lyman and Varian above) estimate that the world produced 14.7 exabytes of new information in 2008, nearly triple the volume of information in 2003.

May 2012 danah boyd and Kate Crawford publish "Critical Questions for Big Data" in *Information, Communications, and Society*. They define big data as "a cultural, technological, and scholarly phenomenon that rests on the interplay of: (1) Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets. (2) Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims. (3) Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy."

An earlier version of this timeline was published on WhatsTheBigData.com

See also A Very Short History of Data Science and A Very Short History of Information Technology

Follow me on Twitter @GilPress or Facebook or Google+
