A Very Short History Of Big Data

MAY 9, 2013 @ 09:45 AM **130,411** VIEWS

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

FOLLOW ON FORBES (281)

Opinions expressed by Forbes Contributors are their own.

FULL BIO

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.

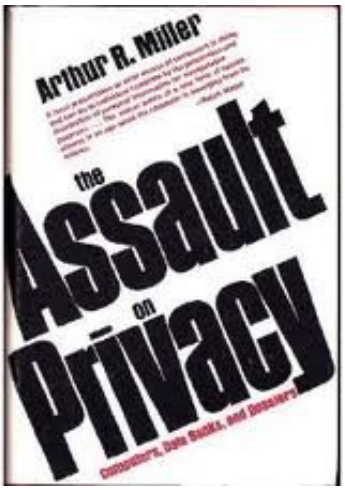
Last Update: December 21, 2013

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."

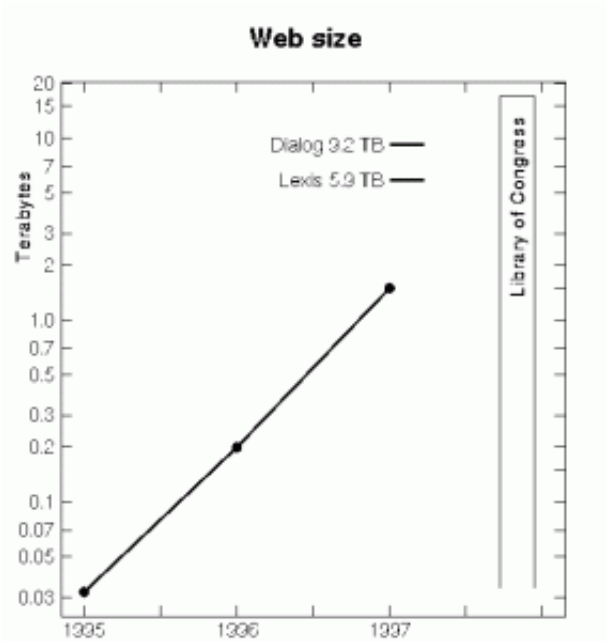
September 1990 Peter J. Denning publishes "Saving All the Bits" (PDF) in *American Scientist*. Says Denning: "The imperative [for scientists] to save all the bits forces us into an impossible situation: The rate and volume of information flow overwhelm our networks, storage devices and retrieval systems, 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."



Source: Michael Lesk

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."

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.

PAGE 1 / 2 Continue >