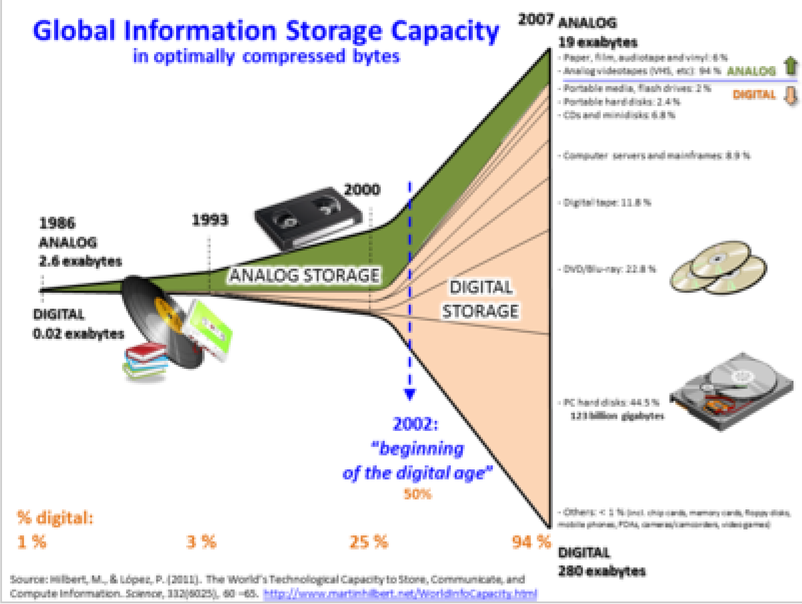
Big data

From Wikipedia, the free encyclopedia



Growth of and digitization of global information-storage capacity[1]

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem."[2] Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on."[3] Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, fintech, urban informatics, and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics,[4] connectomics, complex physics simulations, biology and environmental research.[5]

Data sets grow rapidly - in part because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.[6][7] The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;[8] as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data are generated.[9] One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.[10]

Relational database management systems and desktop statistics- and visualization-packages often have difficulty handling big data. The work may require "massively parallel software running on tens, hundreds, or even thousands of servers".[11] What counts as "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."[12]

Contents

Definition


Visualization created by IBM of daily Wikipedia edits . At multiple terabytes in size, the text and images of Wikipedia are an example of big data.

The term has been in use since the 1990s, with some giving credit to John Mashey for coining or at least making it popular.[13][14] Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.[15] Big Data philosophy encompasses unstructured, semi-structured and structured data, however the main focus is on unstructured data.[16] Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data.[17] Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale.[18]

In a 2001 research report[19] and related lectures, META Group (now Gartner) defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data.[20] In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." Gartner's definition of the 3Vs is still widely used, and in agreement with a consensual definition that states that "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value".[21] Additionally, a new V "Veracity" is added by some organizations to describe it,[22] revisionism challenged by some industry authorities.[23] The 3Vs have been expanded to other complementary characteristics of big data:[24][25]

- Volume: big data doesn't sample; it just observes and tracks what happens
- Velocity: big data is often available in real-time
- Variety: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion
- Machine learning: big data often doesn't ask why and simply detects patterns[26]
- Digital footprint: big data is often a cost-free byproduct of digital interaction[25][27]

The growing maturity of the concept more starkly delineates the difference between big data and Business Intelligence:[28]

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends, etc..
- Big data uses inductive statistics and concepts from nonlinear system identification[29] to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density[30] to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.[29][31]

Characteristics

Big data can be described by the following characteristics:[24][25]

Volume
: The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

Variety
: The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

Velocity
: In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

Variability
: Inconsistency of the data set can hamper processes to handle and manage it.

Veracity
: The quality of captured data can vary greatly, affecting accurate analysis.

Factory work and Cyber-physical systems may have a 6C system:

- Connection (sensor and networks)
- Cloud (computing and data on demand)[32][33]
- Cyber (model and memory)
- Content/context (meaning and correlation)
- Community (sharing and collaboration)
- Customization (personalization and value)

Data must be processed with advanced tools (analytics and algorithms) to reveal meaningful information. For example, to manage a factory one must consider both visible and invisible issues with various components. Information generation algorithms must detect and address invisible issues such as machine degradation, component wear, etc. on the factory floor.[34][35]

Architecture

Big data repositories have existed in many forms, often built by corporations with a special need. Commercial vendors historically offered parallel database management systems for big data beginning in the 1990s. For many years, WinterCorp published a largest database report.[36]

Teradata Corporation in 1984 marketed the parallel processing DBC 1012 system. Teradata systems were the first to store and analyze 1 terabyte of data in 1992. Hard disk drives were 2.5GB in 1991 so the definition of big data continuously evolves according to Kyders Law. Teradata installed the first petabyte class RDBMS based system in 2007. As of 2017, there are a few dozen petabyte class Teradata relational databases installed, the largest of which exceeds 50 PB. Systems up until 2008 were 100% structured relational data. Since then, Teradata has added unstructured data types including XML, JSON, and Avro.

In 2000, Seisint Inc. (now LexisNexis Group) developed a C++-based distributed file-sharing framework for data storage and query. The system stores and distributes structured, semi-structured, and unstructured data across multiple servers. Users can build queries in a C++ dialect called ECL. ECL uses an "apply schema on read" method to infer the structure of stored data when it is queried, instead of when it is stored. In 2004, LexisNexis acquired Seisint Inc.[37] and in 2008 acquired ChoicePoint, Inc.[38] and their high-speed parallel processing platform. The two platforms were merged into HPCC (or High-Performance Computing Cluster) Systems and in 2011, HPCC was open-sourced under the Apache v2.0 License. Quantcast File System was available about the same time.[39]

In 2004, Google published a paper on a process called MapReduce that uses a similar architecture. The MapReduce concept provides a parallel processing model, and an associated implementation was released to process huge amounts of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step). The framework was very successful,[40] so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open-source project named Hadoop.[41]

MIKE2.0 is an open approach to information management that acknowledges the need for revisions due to big data implications identified in an article titled "Big Data Solution Offering".[42] The methodology addresses handling big data in terms of useful permutations of data sources, complexity in interrelationships, and difficulty in deleting (or modifying) individual records.[43]

2012 studies showed that a multiple-layer architecture is one option to address the issues that big data presents. A distributed parallel architecture distributes data across multiple servers; these parallel execution environments can dramatically improve data processing speeds. This type of architecture inserts data into a parallel DBMS, which implements the use of MapReduce and Hadoop frameworks. This type of framework looks to make the processing power transparent to the end user by using a front-end application server.[44]

Big data analytics for manufacturing applications is marketed as a 5C architecture (connection, conversion, cyber, cognition, and configuration).[45]

The data lake allows an organization to shift its focus from centralized control to a shared model to respond to the changing dynamics of information management. This enables quick segregation of data into the data lake, thereby reducing the overhead time.[46][47]

Technologies

A 2011 McKinsey Global Institute report characterizes the main components and ecosystem of big data as follows:[48]

- Techniques for analyzing data, such as A/B testing, machine learning and natural language processing

- Big data technologies, like business intelligence, cloud computing and databases
- Visualization, such as charts, graphs and other displays of the data

Multidimensional big data can also be represented as tensors, which can be more efficiently handled by tensor-based computation,[49] such as multilinear subspace learning.[50] Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data mining,[51] distributed file systems, distributed databases, cloud and HPC-based infrastructure (applications, storage and computing resources)[52] and the Internet.

Some but not all MPP relational databases have the ability to store and manage petabytes of data. Implicit is the ability to load, monitor, back up, and optimize the use of the large data tables in the RDBMS.[53]

DARPA's Topological Data Analysis program seeks the fundamental structure of massive data sets and in 2008 the technology went public with the launch of a company called Ayasdi.[54]

The practitioners of big data analytics processes are generally hostile to slower shared storage,[55] preferring direct-attached storage (DAS) in its various forms from solid state drive (Ssd) to high capacity SATA disk buried inside parallel processing nodes. The perception of shared storage architectures—Storage area network (SAN) and Network-attached storage (NAS) —is that they are relatively slow, complex, and expensive. These qualities are not consistent with big data analytics systems that thrive on system performance, commodity infrastructure, and low cost.

Real or near-real time information delivery is one of the defining characteristics of big data analytics. Latency is therefore avoided whenever and wherever possible. Data in memory is good—data on spinning disk at the other end of a FC SAN connection is not. The cost of a SAN at the scale needed for analytics applications is very much higher than other storage techniques.

There are advantages as well as disadvantages to shared storage in big data analytics, but big data analytics practitioners as of 2011 did not favour it.[56]

Applications

Big data has increased the demand of information management specialists so much so that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms specializing in data management and analytics. In 2010, this industry was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.[3]

Bus wrapped with SAP Big data parked outside IDF13.

Developed economies increasingly use data-intensive technologies. There are 4.6 billion mobile-phone subscriptions worldwide, and between 1 billion and 2 billion people accessing the internet.[3] Between 1990 and 2005, more than 1 billion people worldwide entered the middle class, which means more people became more literate, which in turn lead to information growth. The world's effective capacity to exchange information through telecommunication networks was 281 petabytes in 1986, 471 petabytes in 1993, 2.2 exabytes in 2000, 65 exabytes in 2007[8] and predictions put the amount of internet traffic at 667 exabytes annually by 2014.[3] According to one estimate, one third of the globally stored information is in the form of alphanumeric text and still image data,[57] which is the format most useful for most big data applications. This also shows the potential of yet unused data (i.e. in the form of video and audio content).

While many vendors offer off-the-shelf solutions for big data, experts recommend the development of in-house solutions custom-tailored to solve the company's problem at hand if the company has sufficient technical capabilities.[58]

Government

The use and adoption of big data within governmental processes allows efficiencies in terms of cost, productivity, and innovation,[59] but does not come without its flaws. Data analysis often requires multiple parts of government (central and local) to work in collaboration and create new and innovative processes to deliver the desired outcome. Below are some examples of initiatives the governmental big data space.

United States of America

- In 2012, the Obama administration announced the Big Data Research and Development Initiative, to explore how big data could be used to address important problems faced by the government.[60] The initiative is composed of 84 different big data programs spread across six departments.[61]
- Big data analysis played a large role in Barack Obama's successful 2012 re-election campaign.[62]
- The United States Federal Government owns six of the ten most powerful supercomputers in the world.[63]
- The Utah Data Center has been constructed by the United States National Security Agency. When finished, the facility will be able to handle a large amount of information collected by the NSA over the Internet. The exact amount of storage space is unknown, but more recent sources claim it will be on the order of a few exabytes.[64][65][66]

India

- Big data analysis was tried out for the BJP to win the Indian General Election 2014.[67]
- The Indian government utilizes numerous techniques to ascertain how the Indian electorate is responding to government action, as well as ideas for policy augmentation.

United Kingdom

Examples of uses of big data in public services:

- Data on prescription drugs: by connecting origin, location and the time of each prescription, a research unit was able to exemplify the considerable delay between the release of any given drug, and a UK-wide adaptation of the National Institute for Health and Care Excellence guidelines. This suggests that new or most up-to-date drugs take some time to filter through to the general patient.[68]
- Joining up data: a local authority blended data about services, such as road gritting rotas, with services for people at risk, such as 'meals on wheels'. The connection of data allowed the local authority to avoid any weather-related delay.[69]

International development

Research on the effective usage of information and communication technologies for development (also known as ICT4D) suggests that big data technology can make important contributions but also present unique challenges to International development.[70][71] Advancements in big data analysis offer cost-effective opportunities to improve decision-making in critical development areas such as health care, employment, economic productivity, crime, security, and natural disaster and resource management.[72][73][74] Additionally, user-generated data offers new opportunities to give the unheard a voice.[75] However, longstanding challenges for developing regions such as inadequate technological infrastructure and economic and human resource scarcity exacerbate existing concerns with big data such as privacy, imperfect methodology, and interoperability issues.[72]

Manufacturing

Based on TCS 2013 Global Trend Study, improvements in supply planning and product quality provide the greatest benefit of big data for manufacturing. Big data provides an infrastructure for transparency in manufacturing industry, which is the ability to unravel uncertainties such as inconsistent component performance and availability. Predictive manufacturing as an applicable approach toward near-zero downtime and transparency requires vast amount of data and advanced prediction tools for a systematic process of data into useful information.[76] A conceptual framework of predictive manufacturing begins with data acquisition where different type of sensory data is available to acquire such as acoustics, vibration, pressure, current, voltage and controller data. Vast amount of sensory data in addition to historical data construct the big data in manufacturing. The generated big data acts as the input into predictive tools and preventive strategies such as Prognostics and Health Management (PHM).[77][78]

Cyber-physical models

Current PHM implementations mostly use data during the actual usage while analytical algorithms can perform more accurately when more information throughout the machine's lifecycle, such as system configuration, physical knowledge and working principles, are included. There is a need to systematically integrate, manage and analyze machinery or process data during different stages of machine life cycle to handle data/information more efficiently and further achieve better transparency of machine health condition for manufacturing industry.

With such motivation a cyber-physical (coupled) model scheme has been developed. The coupled model is a digital twin of the real machine that operates in the cloud platform and simulates the health condition with an integrated knowledge from both data driven analytical algorithms as well as other available physical knowledge. It can also be described as a 5S systematic approach consisting of sensing, storage, synchronization, synthesis and service. The coupled model first constructs a digital image from the early design stage. System information and physical knowledge are logged during product design, based on which a simulation model is built as a reference for future analysis. Initial parameters may be statistically generalized and they can be tuned using data from testing or the manufacturing process using parameter estimation. After that step, the simulation model can be considered a mirrored image of the real machine—able to continuously record and track machine condition during the later utilization stage. Finally, with the increased connectivity offered by cloud computing technology, the coupled model also provides better accessibility of machine condition for factory managers in cases where physical access to actual equipment or machine data is limited.[35]

Healthcare

Big data analytics has helped healthcare improve by providing personalized medicine and prescriptive analytics, clinical risk intervention and predictive analytics, waste and care variability reduction, automated external and internal reporting of patient data, standardized medical terms and patient registries and fragmented point solutions.[79] Some areas of improvement are more aspirational than actually implemented. The level of data generated within healthcare systems is not trivial. With the added adoption of mHealth, eHealth and wearable technologies the volume of data will continue to increase. This includes electronic health record data, imaging data, patient generated data, sensor data, and other forms of difficult to process data. There is now an even greater need for such environments to pay greater attention to data and information quality.[80] "Big data very often means `dirty data' and the fraction of data inaccuracies increases with data volume growth." Human inspection at the big data scale is impossible and there is a desperate need in health service for intelligent tools for accuracy and believability control and handling of information missed.[81] While extensive information in healthcare is now electronic, it fits under the big data umbrella as most is unstructured and difficult to use.[82]

Education

A McKinsey Global Institute study found a shortage of 1.5 million highly trained data professionals and managers[48] and a number of universities[83] including University of Tennessee and UC Berkeley, have created masters programs to meet this demand. Private bootcamps have also developed programs to meet that demand, including free programs like The Data Incubator or paid programs like General Assembly.[84]

Media

To understand how the media utilises big data, it is first necessary to provide some context into the mechanism used for media process. It has been suggested by Nick Couldry and Joseph Turow that practitioners in Media and Advertising approach big data as many actionable points of information about millions of individuals. The industry appears to be moving away from the traditional approach of using specific media environments such as newspapers, magazines, or television shows and instead taps into consumers with technologies that reach targeted people at optimal times in optimal locations. The ultimate aim is to serve, or convey, a message or content that is (statistically speaking) in line with the consumer's mindset. For example, publishing environments are increasingly tailoring messages (advertisements) and content (articles) to appeal to consumers that have been exclusively gleaned through various data-mining activities.[85]

- Targeting of consumers (for advertising by marketers)
- Data-capture
- Data journalism: publishers and journalists use big data tools to provide unique and innovative insights and infographics.

Channel 4, the British public-service television broadcaster, is a leader in the field of big data and data analysis.[86]

Internet of Things (IoT)

Big data and the IoT work in conjunction. Data extracted from IoT devices provides a mapping of device inter-connectivity. Such mappings have been used by the media industry, companies and governments to more accurately target their audience and increase media efficiency. IoT is also increasingly adopted as a means of gathering sensory data, and this sensory data has been used in medical [87] and manufacturing [88] contexts.

Technology

- eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising.[89]
- Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.[90]
- Facebook handles 50 billion photos from its user base.[91]
- Google was handling roughly 100 billion searches per month as of August 2012.[92]
- Oracle NoSQL Database has been tested to past the 1M ops/sec mark with 8 shards and proceeded to hit 1.2M ops/sec with 10 shards.[93]

Information Technology

Especially since 2015, big data has come to prominence within Business Operations as a tool to help employees work more efficiently and streamline the collection and distribution of Information Technology (IT). The use of big data to resolve IT and data collection issues within an enterprise is called IT Operations Analytics (ITOA).[94] By applying big data principles into the concepts of machine intelligence and deep computing, IT departments can predict potential issues and move to provide solutions before the problems even happen.[94] In this time, ITOA businesses were also beginning to play a major role in systems management by offering platforms that brought individual data silos together and generated insights from the whole of the system rather than from isolated pockets of data.

Retail

- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data—the equivalent of 167 times the information contained in all the books in the US Library of Congress.[3]

Retail banking

- FICO Card Detection System protects accounts worldwide.[95]
- The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates.[96][97]

Real estate

- Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.[98]

Science

The Large Hadron Collider experiments represent about 150 million sensors delivering data 40 million times per second. There are nearly 600 million collisions per second. After filtering and refraining from recording more than 99.99995%[99] of these streams, there are 100 collisions of interest per second.[100][101][102]

- As a result, only working with less than 0.001% of the sensor stream data, the data flow from all four LHC experiments represents 25 petabytes annual rate before replication (as of 2012). This becomes nearly 200 petabytes after replication.
- If all sensor data were recorded in LHC, the data flow would be extremely hard to work with. The data flow would exceed 150 million petabytes annual rate, or nearly 500 exabytes per day, before replication. To put the number in perspective, this is equivalent to 500 quintillion (5×10^{20}) bytes per day, almost 200 times more than all the other sources combined in the world.

The Square Kilometre Array is a radio telescope built of thousands of antennas. It is expected to be operational by 2024. Collectively, these antennas are expected to gather 14 exabytes and store one petabyte per day.[103][104] It is considered one of the most ambitious scientific projects ever undertaken.[105]

Science and research

- When the Sloan Digital Sky Survey (SDSS) began to collect astronomical data in 2000, it amassed more in its first few weeks than all data collected in the history of astronomy previously. Continuing at a rate of about 200 GB per night, SDSS has amassed more than 140 terabytes of information.[3] When the Large Synoptic Survey Telescope, successor to SDSS, comes online in 2020, its designers expect it to acquire that amount of data every five days.[3]
- Decoding the human genome originally took 10 years to process, now it can be achieved in less than a day. The DNA sequencers have divided the sequencing cost by 10,000 in the last ten years, which is 100

times cheaper than the reduction in cost predicted by Moore's Law.[106]

- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.[107][108]
- Google's DNAStack compiles and organizes DNA samples of genetic data from around the world to identify diseases and other medical defects. These fast and exact calculations eliminate any 'friction points,' or human errors that could be made by one of the numerous science and biology experts working with the DNA. DNAStack, a part of Google Genomics, allows scientists to use the vast sample of resources from Google's search server to scale social experiments that would usually take years, instantly.[109][110]
- 23andme's DNA database contains genetic information of over 1,000,000 people worldwide.[111] The company explores selling the "anonymous aggregated genetic data" to other researchers and pharmaceutical companies for research purposes if patients give their consent.[112][113][114][115][116] Ahmad Hariri, professor of psychology and neuroscience at Duke University who has been using 23andMe in his research since 2009 states that the most important aspect of the company's new service is that it makes genetic research accessible and relatively cheap for scientists.[112] A study that identified 15 genome sites linked to depression in 23andMe's database lead to a surge in demands to access the repository with 23andMe fielding nearly 20 requests to access the depression data in the two weeks after publication of the paper.[117]
- Computational Fluid Dynamics (CFD) and hydrodynamic turbulence research generate massive datasets. The Johns Hopkins Turbulence Databases (JHTDB) contain over 350 terabytes of spatio-temporal fields from Direct Numerical simulations of various turbulent flows. Such data have been difficult to share using traditional methods such as downloading flat simulation output files. The data within JHDTB can be accessed using "virtual sensors" with various access modes ranging from direct web-browser queries, access through Matlab, Python, Fortran and C programs executing on clients' platforms, to cutout services to download raw data. The data have been used in over 150 scientific publications.

Sports

Big data can be used to improve training and understanding competitors, using sport sensors. It is also possible to predict winners in a match using big data analytics.[118] Future performance of players could be predicted as well. Thus, players' value and salary is determined by data collected throughout the season.[119]

The movie *MoneyBall* demonstrates how big data could be used to scout players and also identify undervalued players.[120]

In Formula One races, race cars with hundreds of sensors generate terabytes of data. These sensors collect data points from tire pressure to fuel burn efficiency.[121] Based on the data, engineers and data analysts decide whether adjustments should be made in order to win a race. Besides, using big data, race teams try to predict the time they will finish the race beforehand, based on simulations using data collected over the season.[122]

Research activities

Encrypted search and cluster formation in big data was demonstrated in March 2014 at the American Society of Engineering Education. Gautam Siwach engaged at *Tackling the challenges of Big Data* by MIT Computer Science and Artificial Intelligence Laboratory and Dr. Amir Esmailpour at UNH Research Group investigated the key features of big data as formation of clusters and their interconnections. They focused on the security of big data and the actual orientation of the term towards the presence of different type of data in an encrypted form at cloud interface by providing the raw definitions and real time examples within the technology. Moreover, they proposed an approach for identifying the encoding technique to advance towards an expedited search over encrypted text leading to the security enhancements in big data.[123]

In March 2012, The White House announced a national "Big Data Initiative" that consisted of six Federal departments and agencies committing more than $200 million to big data research projects.[124]

The initiative included a National Science Foundation "Expeditions in Computing" grant of $10 million over 5 years to the AMPLab[125] at the University of California, Berkeley.[126] The AMPLab also received funds from DARPA, and over a dozen industrial sponsors and uses big data to attack a wide range of problems from predicting traffic congestion[127] to fighting cancer.[128]

The White House Big Data Initiative also included a commitment by the Department of Energy to provide $25 million in funding over 5 years to establish the Scalable Data Management, Analysis and Visualization (SDAV) Institute,[129] led by the Energy Department's Lawrence Berkeley National Laboratory. The SDAV Institute aims to bring together the expertise of six national laboratories and seven universities to develop new tools to help scientists manage and visualize data on the Department's supercomputers.

The U.S. state of Massachusetts announced the Massachusetts Big Data Initiative in May 2012, which provides funding from the state government and private companies to a variety of research institutions.[130] The Massachusetts Institute of Technology hosts the Intel Science and Technology Center for Big Data in the MIT Computer Science and Artificial Intelligence Laboratory, combining government, corporate, and institutional funding and research efforts.[131]

The European Commission is funding the 2-year-long Big Data Public Private Forum through their Seventh Framework Program to engage companies, academics and other stakeholders in discussing big data issues. The project aims to define a strategy in terms of research and innovation to guide supporting actions from the European Commission in the successful implementation of the big data economy. Outcomes of this project will be used as input for Horizon 2020, their next framework program.[132]

The British government announced in March 2014 the founding of the Alan Turing Institute, named after the computer pioneer and code-breaker, which will focus on new ways to collect and analyse large data sets.[133]

At the University of Waterloo Stratford Campus Canadian Open Data Experience (CODE) Inspiration Day, participants demonstrated how using data visualization can increase the understanding and appeal of big data sets and communicate their story to the world.[134]

To make manufacturing more competitive in the United States (and globe), there is a need to integrate more American ingenuity and innovation into manufacturing ; Therefore, National Science Foundation has granted the Industry University cooperative research center for Intelligent Maintenance Systems (IMS) at university of Cincinnati to focus on developing advanced predictive tools and techniques to be applicable in a big data environment.[135] In May 2013, IMS Center held an industry advisory board meeting focusing on big data where presenters from various industrial companies discussed their concerns, issues and future goals in big data environment.

Computational social sciences – Anyone can use Application Programming Interfaces (APIs) provided by big data holders, such as Google and Twitter, to do research in the social and behavioral sciences.[136] Often these APIs are provided for free.[136] Tobias Preis *et al.* used Google Trends data to demonstrate that Internet users from countries with a higher per capita gross domestic product (GDP) are more likely to search for information about the future than information about the past. The findings suggest there may be a link between online behaviour and real-world economic indicators.[137][138][139] The authors of the study examined Google queries logs made by ratio of the volume of searches for the coming year ('2011') to the volume of searches for the previous year ('2009'), which they call the 'future orientation index'.[140] They compared the future orientation index to the per capita GDP of each country, and found a strong tendency for countries where Google users inquire more about the future to have a higher GDP. The results hint that there may potentially be a relationship between the economic success of a country and the information-seeking behavior of its citizens captured in big data.

Tobias Preis and his colleagues Helen Susannah Moat and H. Eugene Stanley introduced a method to identify online precursors for stock market moves, using trading strategies based on search volume data provided by Google Trends.[141] Their analysis of Google search volume for 98 terms of varying financial relevance, published in *Scientific Reports*,[142] suggests that increases in search volume for financially relevant search terms tend to precede large losses in financial markets.[143][144][145][146][147][148][149][150]

Big data sets come with algorithmic challenges that previously did not exist. Hence, there is a need to fundamentally change the processing ways.[151]

The Workshops on Algorithms for Modern Massive Data Sets (MMDS) bring together computer scientists, statisticians, mathematicians, and data analysis practitioners to discuss algorithmic challenges of big data.[152]

Sampling big data

An important research question that can be asked about big data sets is whether you need to look at the full data to draw certain conclusions about the properties of the data or is a sample good enough. The name big data itself contains a term related to size and this is an important characteristic of big data. But Sampling (statistics) enables the selection of right data points from within the larger data set to estimate the characteristics of the whole population. For example, there are about 600 million tweets produced every day. Is it necessary to look at all of them to determine the topics that are discussed during the day? Is it necessary to look at all the tweets to determine the sentiment on each of the topics? In manufacturing different types of sensory data such as acoustics, vibration, pressure, current, voltage and controller data are available at short time intervals. To predict down-time it may not be necessary to look at all the data but a sample may be sufficient. Big Data can be broken down by various data point categories such as demographic, psychographic, behavioral, and transactional data. With large sets of data points, marketers are able to create and utilize more customized segments of consumers for more strategic targeting.

There has been some work done in Sampling algorithms for big data. A theoretical formulation for sampling Twitter data has been developed.[153]

Critique

Critiques of the big data paradigm come in two flavors, those that question the implications of the approach itself, and those that question the way it is currently done.[154] One approach to this criticism is the field of Critical data studies.

Critiques of the big data paradigm

"A crucial problem is that we do not know much about the underlying empirical micro-processes that lead to the emergence of the[se] typical network characteristics of Big Data".[15] In their critique, Snijders, Matzat, and Reips point out that often very strong assumptions are made about mathematical properties that may not at all reflect what is really going on at the level of micro-processes. Mark Graham has leveled broad critiques at Chris Anderson's assertion that big data will spell the end of theory:[155] focusing in particular on the notion that big data must always be contextualized in their social, economic, and political contexts.[156] Even as companies invest eight- and nine-figure sums to derive insight from information streaming in from suppliers and customers, less than 40% of employees have sufficiently mature processes and skills to do so. To overcome this insight deficit, big data, no matter how comprehensive or well analysed, must be complemented by "big judgment," according to an article in the Harvard Business Review.[157]

Much in the same line, it has been pointed out that the decisions based on the analysis of big data are inevitably "informed by the world as it was in the past, or, at best, as it currently is".[72] Fed by a large number of data on past experiences, algorithms can predict future development if the future is similar to the past.[158] If the systems dynamics of the future change (if it is not a stationary process), the past can say little about the future. In order to make predictions in changing environments, it would be necessary to have a thorough understanding of the systems dynamic, which requires theory.[158] As a response to this critique it has been suggested to combine big data approaches with computer simulations, such as agent-based models[72] and Complex Systems. Agent-based models are increasingly getting better in predicting the outcome of social complexities of even unknown future scenarios through computer simulations that are based on a collection of mutually interdependent algorithms.[159][160] In addition, use of multivariate methods that probe for the latent structure of the data, such as factor analysis and cluster analysis, have proven useful as analytic approaches that go well beyond the bi-variate approaches (cross-tabs) typically employed with smaller data sets.

In health and biology, conventional scientific approaches are based on experimentation. For these approaches, the limiting factor is the relevant data that can confirm or refute the initial hypothesis.[161] A new postulate is accepted now in biosciences: the information provided by the data in huge volumes (omics) without prior hypothesis is complementary and sometimes necessary to conventional approaches based on experimentation.[162][163] In the massive approaches it is the formulation of a relevant hypothesis to explain the data that is the limiting factor.[164] The search logic is reversed and the limits of induction ("Glory of Science and Philosophy scandal", C. D. Broad, 1926) are to be considered.

Privacy advocates are concerned about the threat to privacy represented by increasing storage and integration of personally identifiable information; expert panels have released various policy recommendations to conform practice to expectations of privacy.[165][166][167]

Nayef Al-Rodhan argues that a new kind of social contract will be needed to protect individual liberties in a context of Big Data and giant corporations that own vast amounts of information. The use of Big Data should be monitored and better regulated at the national and international levels.[168]

Critiques of big data execution


Danah Boyd

Ulf-Dietrich Reips and Uwe Matzat wrote in 2014 that big data had become a "fad" in scientific research.[136] Researcher Danah Boyd has raised concerns about the use of big data in science neglecting principles such as choosing a representative sample by being too concerned about actually handling the huge amounts of data.[169] This approach may lead to results bias in one way or another. Integration across heterogeneous data resources—some that might be considered big data and others not—presents formidable logistical as well as analytical challenges, but many researchers argue that such integrations are likely to represent the most promising new frontiers in science.[170] In the provocative article "Critical Questions for Big Data",[171] the authors title big data a part of mythology: "large data sets offer a higher form of intelligence and knowledge [...], with the aura of truth, objectivity, and accuracy". Users of big data are often "lost in the sheer volume of numbers", and "working with Big Data is still subjective, and what it quantifies does not necessarily have a closer claim on objective truth".[171] Recent developments in BI domain, such as pro-active reporting especially target improvements in usability of big data, through automated filtering of non-useful data and correlations.[172]

Big data analysis is often shallow compared to analysis of smaller data sets.[173] In many big data projects, there is no large data analysis happening, but the challenge is the extract, transform, load part of data preprocessing.[173]

Big data is a buzzword and a "vague term",[174][175] but at the same time an "obsession"[175] with entrepreneurs, consultants, scientists and the media. Big data showcases such as Google Flu Trends failed to deliver good predictions in recent years, overstating the flu outbreaks by a factor of two. Similarly, Academy awards and election predictions solely based on Twitter were more often off than on target. Big data often poses the same challenges as small data; and adding more data does not solve problems of bias, but may emphasize other problems. In particular data sources such as Twitter are not representative of the overall population, and results drawn from such sources may then lead to wrong conclusions. Google Translate—which is based on big data statistical analysis of text—does a good job at translating web pages. However, results from specialized domains may be dramatically skewed. On the other hand, big data may also introduce new problems, such as the multiple comparisons problem: simultaneously testing a large set of hypotheses is likely to produce many false results that mistakenly appear significant. Ioannidis argued that "most published research findings are false"[176] due to essentially the same effect: when many scientific teams and researchers each perform many experiments (i.e. process a big amount of scientific data; although not with big data technology), the likelihood of a "significant" result being actually false grows fast – even more so, when only positive results are published. Furthermore, big data analytics results are only as good as the model on which they are predicated. In an example, big data took part in attempting to predict the results of the 2016 U.S. Presidential Election[177] with varying degrees of success. Forbes predicted "If you believe in *Big Data* analytics, it's time to begin planning for a Hillary Clinton presidency and all that entails.".[178]

See also

- Big memory
- List of big data companies
- Datafication
- Data defined storage
- Data journalism
- Data lineage
- Data philanthropy
- Data science
- Statistics
- Surveillance capitalism
- Small data
- Urban informatics
- Big Data Maturity Model

References

1. "The World's Technological Capacity to Store, Communicate, and Compute Information" (http://www.martinhilbert.net/WorldInfoCapacity.html). *MartinHilbert.net*. Retrieved 13 April 2016.
2. boyd, dana; Crawford, Kate (September 21, 2011). "Six Provocations for Big Data". *Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. doi:10.2139/ssrn.1926431 (https://doi.org/10.2139%2Fssrn.1926431).
3. "Data, data everywhere" (http://www.economist.com/node/15557443). *The Economist*. 25 February 2010. Retrieved 9 December 2012.
4. "Community cleverness required" (http://www.nature.com/nature/journal/v455/n7209/full/455001a.html). *Nature*. **455** (7209): 1. 4 September 2008. PMID 18769385 (https://www.ncbi.nlm.nih.gov/pubmed/18769385). doi:10.1038/455001a (https://doi.org/10.1038%2F455001a).
5. Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011). "Challenges and Opportunities of Open Data in Ecology". *Science*. **331** (6018): 703–5. PMID 21311007 (https://www.ncbi.nlm.nih.gov/pubmed/21311007). doi:10.1126/science.1197962 (https://doi.org/10.1126%2Fscience.1197962).
6. Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data" (http://gigaom.com/2008/11/09/mapreduce-leads-the-way-for-parallel-programming/). *Gigaom Blog*.
7. Segaran, Toby; Hammerbacher, Jeff (2009). *Beautiful Data: The Stories Behind Elegant Data Solutions* (https://books.google.com/books?id=zxNglqU1FKgC). O'Reilly Media. p. 257. ISBN 978-0-596-15711-1.

8. Hilbert, Martin; López, Priscila (2011). "The World's Technological Capacity to Store, Communicate, and Compute Information" (http://martinhilbert.net/WorldInfoCapacity.html). *Science*. **332** (6025): 60–65. PMID 21310967 (https://www.ncbi.nlm.nih.gov/pubmed/21310967). doi:10.1126/science.1200970 (https://doi.org/10.1126%2Fscience.1200970).
9. "IBM What is big data? – Bringing big data to the enterprise" (http://www.ibm.com/big-data/us/en/). www.ibm.com. Retrieved 2013-08-26.
10. Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity" (http://www.fsn.co.uk/channel_bi_bpm_cpm/mastering_big_data_cfo_strategies_to_transform_insight_into_opportunity#.UO2Ac-TTuys), December 2012
11. Jacobs, A. (6 July 2009). "The Pathologies of Big Data" (http://queue.acm.org/detail.cfm?id=1563874). *ACMQueue*.
12. Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data" (http://radar.oreilly.com/r2/release2-0-11.html). *Release 2.0*. Sebastopol CA: O'Reilly Media (**11**).
13. John R. Mashey (25 April 1998). "Big Data ... and the Next Wave of InfraStress" (http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf) (PDF). *Slides from invited talk*. Usenix. Retrieved 28 September 2016.
14. Steve Lohr (1 February 2013). "The Origins of 'Big Data': An Etymological Detective Story" (http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/). New York Times. Retrieved 28 September 2016.
15. Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet" (http://www.ijis.net/ijis7_1/ijis7_1_editorial.html). *International Journal of Internet Science*. **7**: 1–5.
16. Dedić, N.; Stanier, C. (2017). "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery". Vol. 285. Berlin ; Heidelberg: Springer International Publishing. ISSN 1865-1356 (https://www.worldcat.org/issn/1865-1356). OCLC 909580101 (https://www.worldcat.org/oclc/909580101).
17. Everts, Sarah (2016). "Information Overload" (https://www.chemheritage.org/distillations/magazine/information-overload). *Distillations*. **2** (2): 26–33. Retrieved 17 February 2017.
18. Ibrahim; Targio Hashem, Abaker; Yaqoob, Ibrar; Badrul Anuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "big data" on cloud computing: Review and open research issues". *Information Systems*. **47**: 98–115. doi:10.1016/j.is.2014.07.006 (https://doi.org/10.1016%2Fj.is.2014.07.006).
19. Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety" (http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf) (PDF). Gartner. Retrieved 6 February 2001.
20. Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data" (http://www.gartner.com/it/page.jsp?id=1731916). Gartner. Archived (https://web.archive.org/web/20110710043533/http://www.gartner.com/it/page.jsp?id=1731916) from the original on 10 July 2011. Retrieved 13 July 2011.
21. De Mauro, Andrea; Greco, Marco; Grimaldi, Michele (2016). "A Formal definition of Big Data based on its essential Features" (http://www.emeraldinsight.com/doi/abs/10.1108/LR-06-2015-0061). *Library Review*. **65**: 122–135. doi:10.1108/LR-06-2015-0061 (https://doi.org/10.1108%2FLR-06-2015-0061).
22. "What is Big Data?" (http://www.villanovau.com/university-online-programs/what-is-big-data/). Villanova University.
23. Grimes, Seth. "Big Data: Avoid 'Wanna V' Confusion" (http://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077?). InformationWeek. Retrieved 5 January 2016.
24. Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review." (http://www.martinhilbert.net/big-data-for-development). *martinhilbert.net*. Retrieved 2015-10-07.
25. *DT&SC 7-3: What is Big Data?* (https://www.youtube.com/watch?v=XRVIh1h47sA&index=51&list=PLtjBSCvWCU3Nm46D3R85efM0hrzjuAIg). 12 August 2015 – via YouTube.
26. Mayer-Schönberger, V., & Cukier, K. (2013). Big data: a revolution that will transform how we live, work and think. London: John Murray.
27. "Digital Technology & Social Change" (https://canvas.instructure.com/courses/949415).
28. http://www.bigdataparis.com/presentation/mercredi/PDelort.pdf?PHPSESSID=tv7k70pcr3egpi2r6fi3qbjtj6#page=4
29. Billings S.A. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains". Wiley, 2013
30. "le Blog ANDSI » DSI Big Data" (http://www.andsi.fr/tag/dsi-big-data/).
31. Les Echos (3 April 2013). "Les Echos – Big Data car Low-Density Data ? La faible densité en information comme facteur discriminant – Archives" (http://lecercle.lesechos.fr/entrepreneur/tendances-innovation/221169222/big-data-low-density-data-faible-densite-information-com). *lesechos.fr*.
32. Wu, D., Liu. X., Hebert, S., Gentzsch, W., Terpenny, J. (2015). Performance Evaluation of Cloud-Based High Performance Computing for Finite Element Analysis. Proceedings of the ASME 2015 International Design Engineering Technical Conference & Computers and Information in Engineering Conference (IDETC/CIE2015), Boston, Massachusetts, U.S.
33. Wu, D.; Rosen, D.W.; Wang, L.; Schaefer, D. (2015). "Cloud-Based Design and Manufacturing: A New Paradigm in Digital Manufacturing and Design Innovation". *Computer-Aided Design*. **59** (1): 1–14. doi:10.1016/j.cad.2014.07.006 (https://doi.org/10.1016%2Fj.cad.2014.07.006).

34. Lee, Jay; Bagheri, Behrad; Kao, Hung-An (2014). "Recent Advances and Trends of Cyber-Physical Systems and Big Data Analytics in Industrial Informatics" (https://www.researchgate.net/profile/Behrad_Bagheri/publication/266375284_Recent_Advances_and_Trends_of_Cyber-Physical_Systems_and_Big_Data_Analytics_in_Industrial_Informatics/links/542dc0100cf27e39fa948a7d?origin=publication_detail). *IEEE Int. Conference on Industrial Informatics (INDIN) 2014*.
35. Lee, Jay; Lapira, Edzel; Bagheri, Behrad; Kao, Hung-an. "Recent advances and trends in predictive manufacturing systems in big data environment" (http://www.sciencedirect.com/science/article/pii/S2213846313000114). *Manufacturing Letters* **1** (1): 38–41. doi:10.1016/j.mfglet.2013.09.005 (https://doi.org/10.1016%2Fj.mfglet.2013.09.005).
36. http://www.eweek.com/database/survey-biggest-databases-approach-30-terabytes Survey: Biggest Databases Approach 30 Terabytes, 2003
37. "LexisNexis To Buy Seisint For $775 Million" (http://www.washingtonpost.com/wp-dyn/articles/A50577-2004Jul14.html). Washington Post. Retrieved 15 July 2004.
38. "LexisNexis Parent Set to Buy ChoicePoint" (http://www.washingtonpost.com/wp-dyn/content/article/2008/02/21/AR2008022100809.html). Washington Post. Retrieved 22 February 2008.
39. "Quantcast Opens Exabyte-Ready File System" (http://www.datanami.com/2012/10/01/quantcast_opens_exabyte_ready_file_system/). www.datanami.com. Retrieved 1 October 2012.
40. Bertolucci, Jeff "Hadoop: From Experiment To Leading Big Data Platform" (http://www.informationweek.com/big-data/news/software-platforms/hadoop-from-experiment-to-leading-big-d/240157176), "Information Week", 2013. Retrieved on 14 November 2013.
41. Webster, John. "MapReduce: Simplified Data Processing on Large Clusters" (http://research.google.com/archive/mapreduce-osdi04.pdf), "Search Storage", 2004. Retrieved on 25 March 2013.
42. "Big Data Solution Offering" (http://mike2.openmethodology.org/wiki/Big_Data_Solution_Offering). MIKE2.0. Retrieved 8 December 2013.
43. "Big Data Definition" (http://mike2.openmethodology.org/wiki/Big_Data_Definition). MIKE2.0. Retrieved 9 March 2013.
44. Boja, C; Pocovnicu, A; Bătăgan, L. (2012). "Distributed Parallel Architecture for Big Data". *Informatica Economica* **16** (2): 116–127.
45. "IMS_CPS — IMS Center" (http://www.imscenter.net/cyber-physical-platform). Retrieved 16 June 2016.
46. http://www.hcltech.com/sites/default/files/solving_key_businesschallenges_with_big_data_lake_0.pdf
47. "Method for testing the fault tolerance of MapReduce frameworks" (https://secplab.ppgia.pucpr.br/files/papers/2015-0.pdf) (PDF). Computer Networks. 2015.
48. Manyika, James; Chui, Michael; Bughin, Jaques; Brown, Brad; Dobbs, Richard; Roxburgh, Charles; Byers, Angela Hung (May 2011). "Big Data: The next frontier for innovation, competition, and productivity" (http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation). McKinsey Global Institute. Retrieved January 16, 2016.
49. "Future Directions in Tensor-Based Computation and Modeling" (http://www.cs.cornell.edu/cv/tenwork/finalreport.pdf) (PDF). May 2009.
50. Lu, Haiping; Plataniotis, K.N.; Venetsanopoulos, A.N. (2011). "A Survey of Multilinear Subspace Learning for Tensor Data" (http://www.dsp.utoronto.ca/~haiping/Publication/SurveyMSL_PR2011.pdf) (PDF). *Pattern Recognition*. **44** (7): 1540–1551. doi:10.1016/j.patcog.2011.01.004 (https://doi.org/10.1016%2Fj.patcog.2011.01.004).
51. Pllana, Sabri; Janciak, Ivan; Brezany, Peter; Wöhrer, Alexander. "A Survey of the State of the Art in Data Mining and Integration Query Languages" (http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6041580). *2011 International Conference on Network-Based Information Systems (NBIS 2011)*. IEEE Computer Society. Retrieved 2 April 2016.
52. "Characterization and Optimization of Memory-Resident MapReduce on HPC Systems" (http://ieeexplore.ieee.org/document/6877311/) (PDF). IEEE. October 2014.
53. Monash, Curt (30 April 2009). "eBay's two enormous data warehouses" (http://www.dbms2.com/2009/04/30/ebays-two-enormous-data-warehouses/).
Monash, Curt (6 October 2010). "eBay followup – Greenplum out, Teradata > 10 petabytes, Hadoop has some value, and more" (http://www.dbms2.com/2010/10/06/ebay-followup-greenplum-out-teradata-10-petabytes-hadoop-has-some-value-and-more/).
54. "Resources on how Topological Data Analysis is used to analyze big data" (http://www.ayasdi.com/resources/). Ayasdi.
55. CNET News (1 April 2011). "Storage area networks need not apply" (http://news.cnet.com/8301-21546_3-20049693-10253464.html).
56. "How New Analytic Systems will Impact Storage" (http://www.evaluatorgroup.com/document/big-data-how-new-analytic-systems-will-impact-storage-2/). September 2011.
57. "An Error Occurred Setting Your User Cookie" (http://www.tandfonline.com/doi/abs/10.1080/01972243.2013.873748). *The Information Society*. **30**: 127–143. doi:10.1080/01972243.2013.873748 (https://doi.org/10.1080%2F01972243.2013.873748).

58. Rajpurohit, Anmol (11 July 2014). "Interview: Amy Gershkoff, Director of Customer Analytics & Insights, eBay on How to Design Custom In-House BI Tools" (http://www.kdnuggets.com/2014/07/interview-amy-gershkoff-ebay-in-house-BI-tools.html). *KDnuggets*. Retrieved 2014-07-14. "Dr. Amy Gershkoff: "Generally, I find that off-the-shelf business intelligence tools do not meet the needs of clients who want to derive custom insights from their data. Therefore, for medium-to-large organizations with access to strong technical talent, I usually recommend building custom, in-house solutions.""
59. "The Government and big data: Use, problems and potential" (http://www.computerworld.com/article/2472667/government-it/the-government-and-big-data--use--problems-and-potential.html). Computerworld. 21 March 2012. Retrieved 12 September 2016.
60. Kalil, Tom. "Big Data is a Big Deal" (http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal). White House. Retrieved 26 September 2012.
61. Executive Office of the President (March 2012). "Big Data Across the Federal Government" (http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf) (PDF). White House. Retrieved 26 September 2012.
62. Lampitt, Andrew. "The real story of how big data analytics helped Obama win" (http://www.infoworld.com/d/big-data/the-real-story-of-how-big-data-analytics-helped-obama-win-212862). *Infoworld*. Retrieved 31 May 2014.
63. Hoover, J. Nicholas. "Government's 10 Most Powerful Supercomputers" (http://www.informationweek.com/government/enterprise-applications/image-gallery-governments-10-most-powerf/224700271). *Information Week*. UBM. Retrieved 26 September 2012.
64. Bamford, James (15 March 2012). "The NSA Is Building the Country's Biggest Spy Center (Watch What You Say)" (https://www.wired.com/threatlevel/2012/03/ff_nsadatacenter/all/1). *Wired Magazine*. Retrieved 2013-03-18.
65. "Groundbreaking Ceremony Held for $1.2 Billion Utah Data Center" (http://www.nsa.gov/public_info/press_room/2011/utah_groundbreaking_ceremony.shtml). National Security Agency Central Security Service. Retrieved 2013-03-18.
66. Hill, Kashmir. "TBlueprints of NSA's Ridiculously Expensive Data Center in Utah Suggest It Holds Less Info Than Thought" (http://www.forbes.com/sites/kashmirhill/2013/07/24/blueprints-of-nsa-data-center-in-utah-suggest-its-storage-capacity-is-less-impressive-than-thought/). *Forbes*. Retrieved 2013-10-31.
67. "News: Live Mint" (http://www.livemint.com/Industry/bUQo8xQ3gStSAy5II9lxoK/Are-Indian-companies-making-enough-sense-of-Big-Data.html). *Are Indian companies making enough sense of Big Data?*. Live Mint. 23 June 2014. Retrieved 2014-11-22.
68. "Survey on Big Data Using Data Mining" (https://www.ijedr.org/papers/IJEDR1504022.pdf) (PDF). International Journal of Engineering Development and Research. 2015. Retrieved 14 September 2016.
69. "Recent advances delivered by Mobile Cloud Computing and Internet of Things for Big Data applications: a survey" (https://www.researchgate.net/publication/297762848_Recent_advances_delivered_by_mobile_cloud_computing_and_Internet_of_Things_for_Big_data_applications_A_Survey). International Journal of Network Management. 11 March 2016. Retrieved 14 September 2016.
70. "White Paper: Big Data for Development: Opportunities & Challenges (2012) – United Nations Global Pulse" (http://www.unglobalpulse.org/projects/BigDataforDevelopment). Retrieved 13 April 2016.
71. "WEF (World Economic Forum), & Vital Wave Consulting. (2012). Big Data, Big Impact: New Possibilities for International Development" (http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development). *World Economic Forum*. Retrieved 24 August 2012.
72. "Big Data for Development: From Information- to Knowledge Societies". SSRN 2205145 (https://ssrn.com/abstract=2205145).
73. "Elena Kvochko, Four Ways To talk About Big Data (Information Communication Technologies for Development Series)" (http://blogs.worldbank.org/ic4d/four-ways-to-talk-about-big-data/). worldbank.org. Retrieved 2012-05-30.
74. "Daniele Medri: Big Data & Business: An on-going revolution" (http://www.statisticsviews.com/details/feature/5393251/Big-Data--Business-An-on-going-revolution.html). Statistics Views. 21 October 2013.
75. Tobias Knobloch and Julia Manske (11 January 2016). "Responsible use of data" (http://www.dandc.eu/en/article/opportunities-and-risks-user-generated-and-automatically-compiled-data). *D+C, Development and Cooperation*.
76. Lee, Jay; Wu, F.; Zhao, W.; Ghaffari, M.; Liao, L (January 2013). "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications". *Mechanical Systems and Signal Processing*. **42** (1).
77. "Tutorials" (https://www.phmsociety.org/events/conference/phm/europe/16/tutorials). PHM Society. Retrieved 27 September 2016.
78. "Prognostic and Health Management Technology for MOCVD Equipment" (https://www.itri.org.tw/eng/Content/MSGPic01/contents.aspx?&SiteID=1&MmmID=620651706136357202&CatID=620653256103620163&MSID=654532365564567545). Industrial Technology Research Institute. Retrieved 27 September 2016.
79. "Impending Challenges for the Use of Big Data". *International Journal of Radiation Oncology*Biology*Physics*. doi:10.1016/j.ijrobp.2015.10.060 (https://doi.org/10.1016%2Fj.ijrobp.2015.10.060).
80. O'Donoghue, John; Herbert, John (1 October 2012). "Data Management Within mHealth Environments: Patient Sensors, Mobile Devices, and Databases" (http://doi.acm.org/10.1145/2378016.2378021). *Journal of Data and Information Quality*. **4** (1): 5:1–5:20. doi:10.1145/2378016.2378021 (https://doi.org/10.1145%2F2378016.2378021). Retrieved 16 June 2016 – via ACM Digital Library.

81. Mirkes, E.M.; Coats, T.J.; Levesley, J.; Gorban, A.N. (2016). "Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes" (https://www.researchgate.net/publication/300400110_Handling_missing_data_in_large_healthcare_dataset_A_case_study_of_unknown_trauma_outcomes). *Computers in Biology and Medicine*. **75**: 203–216. doi:10.1016/j.compbiomed.2016.06.004 (https://doi.org/10.1016%2Fj.compbiomed.2016.06.004).
82. Murdoch, Travis B.; Detsky, Allan S. (2013-04-03). "The Inevitable Application of Big Data to Health Care" (http://jamanetwork.com/journals/jama/article-abstract/1674245). *JAMA*. **309** (13): 1351. ISSN 0098-7484 (https://www.worldcat.org/issn/0098-7484). doi:10.1001/jama.2013.393 (https://doi.org/10.1001%2Fjama.2013.393).
83. "Degrees in Big Data: Fad or Fast Track to Career Success" (http://www.forbes.com/sites/jmaureenhenderson/2013/07/30/degrees-in-big-data-fad-or-fast-track-to-career-success/?). *Forbes*. Retrieved 2016-02-21.
84. "NY gets new bootcamp for data scientists: It's free, but harder to get into than Harvard" (http://venturebeat.com/2014/04/15/ny-gets-new-bootcamp-for-data-scientists-its-free-but-harder-to-get-into-than-harvard/). *Venture Beat*. Retrieved 2016-02-21.
85. Couldry, Nick; Turow, Joseph (2014). "Advertising, Big Data, and the Clearance of the Public Realm: Marketers' New Approaches to the Content Subsidy". *International Journal of Communication*. **8**: 1710–1726.
86. Big data and analytics: The pioneering experiences of C4 and Genius Digital (https://www.ibc.org/tech-advances/big-data-and-analytics-c4-and-genius-digital/1076.article)
87. http://www.businesswire.com/news/home/20170109006500/en/QuiO-Named-Innovation-Champion-Accenture-HealthTech-Innovation
88. https://www.predix.com/sites/default/files/IDC_OT_Final_whitepaper_249120.pdf
89. Tay, Liz. "Inside eBay's 90PB data warehouse" (http://www.itnews.com.au/news/inside-ebay8217s-90pb-data-warehouse-342615). ITNews. Retrieved 2016-02-12.
90. Layton, Julia. "Amazon Technology" (http://money.howstuffworks.com/amazon1.htm). Money.howstuffworks.com. Retrieved 2013-03-05.
91. "Scaling Facebook to 500 Million Users and Beyond" (https://www.facebook.com/notes/facebook-engineering/scaling-facebook-to-500-million-users-and-beyond/409881258919). Facebook.com. Retrieved 2013-07-21.
92. "Google Still Doing at Least 1 Trillion Searches Per Year" (http://searchengineland.com/google-1-trillion-searches-per-year-212940). *Search Engine Land*. 16 January 2015. Retrieved 15 April 2015.
93. Lamb, Charles. "Oracle NoSQL Database Exceeds 1 Million Mixed YCSB Ops/Sec" (https://blogs.oracle.com/charlesLamb/entry/oracle_nosql_database_exceeds_1).
94. Solnik, Ray. "The Time Has Come: Analytics Delivers for IT Operations" (http://www.datacenterjournal.com/time-analytics-delivers-operations/). *Data Center Journal*. Retrieved June 21, 2016.
95. "FICO® Falcon® Fraud Manager" (http://www.fico.com/en/Products/DMApps/Pages/FICO-Falcon-Fraud-Manager.aspx). Fico.com. Retrieved 2013-07-21.
96. "eBay Study: How to Build Trust and Improve the Shopping Experience" (http://research.wpcarey.asu.edu/managing-it/ebay-study-how-to-build-trust-and-improve-the-shopping-experience/). Knowwpcarey.com. 8 May 2012. Retrieved 2015-12-20.
97. Leading Priorities for Big Data for Business and IT (http://www.statista.com/statistics/280444/global-leading-priorities-for-big-data-according-to-business-and-it-executives/). eMarketer. October 2013. Retrieved January 2014.
98. Wingfield, Nick (12 March 2013). "Predicting Commutes More Accurately for Would-Be Home Buyers – NYTimes.com" (http://bits.blogs.nytimes.com/2013/03/12/predicting-commutes-more-accurately-for-would-be-home-buyers/). Bits.blogs.nytimes.com. Retrieved 2013-07-21.
99. Alexandru, Dan. "Prof" (https://cds.cern.ch/record/1504817/files/CERN-THESIS-2013-004.pdf) (PDF). *cds.cern.ch*. CERN. Retrieved 24 March 2015.
100. "LHC Brochure, English version. A presentation of the largest and the most powerful particle accelerator in the world, the Large Hadron Collider (LHC), which started up in 2008. Its role, characteristics, technologies, etc. are explained for the general public." (http://cds.cern.ch/record/1278169?ln=en). *CERN-Brochure-2010-006-Eng. LHC Brochure, English version.* CERN. Retrieved 20 January 2013.
101. "LHC Guide, English version. A collection of facts and figures about the Large Hadron Collider (LHC) in the form of questions and answers." (http://cds.cern.ch/record/1092437?ln=en). *CERN-Brochure-2008-001-Eng. LHC Guide, English version.* CERN. Retrieved 20 January 2013.
102. Brumfiel, Geoff (19 January 2011). "High-energy physics: Down the petabyte highway" (http://www.nature.com/news/2011/110119/full/469282a.html). *Nature*. **469**. pp. 282–83. doi:10.1038/469282a (https://doi.org/10.1038%2F469282a).
103. http://www.zurich.ibm.com/pdf/astron/CeBIT%202013%20Background%20DOME.pdf
104. "Future telescope array drives development of exabyte processing" (https://arstechnica.com/science/2012/04/future-telescope-array-drives-development-of-exabyte-processing/). *Ars Technica*. Retrieved 15 April 2015.
105. "Australia's bid for the Square Kilometre Array – an insider's perspective" (http://theconversation.com/australias-bid-for-the-square-kilometre-array-an-insiders-perspective-4891). *The Conversation*. 1 February 2012. Retrieved 27 September 2016.
106. Delort P., OECD ICCP Technology Foresight Forum, 2012. (http://www.oecd.org/sti/ieconomy/Session_3_Delort.pdf#page=6)

107. "NASA – NASA Goddard Introduces the NASA Center for Climate Simulation" (http://www.nasa.gov/centers/goddard/news/releases/2010/10-051.html). Retrieved 13 April 2016.
108. Webster, Phil. "Supercomputing the Climate: NASA's Big Data Mission" (http://www.csc.com/cscworld/publications/81769/81773-supercomputing_the_climate_nasa_s_big_data_mission). *CSC World*. Computer Sciences Corporation. Retrieved 2013-01-18.
109. "These six great neuroscience ideas could make the leap from lab to market" (http://www.theglobeandmail.com/life/health-and-fitness/health/these-six-great-neuroscience-ideas-could-make-the-leap-from-lab-to-market/article21681731/). The Globe and Mail. 20 November 2014. Retrieved 1 October 2016.
110. "DNAstack tackles massive, complex DNA datasets with Google Genomics" (https://cloud.google.com/customers/dnastack/). Google Cloud Platform. Retrieved 1 October 2016.
111. "23andMe - Ancestry" (https://www.23andme.com/en-int/ancestry/). *23andme.com*. Retrieved 29 December 2016.
112. Potenza, Alessandra (13 July 2016). "23andMe wants researchers to use its kits, in a bid to expand its collection of genetic data" (http://www.theverge.com/2016/7/13/12166960/23andme-genetic-testing-database-genotyping-research). The Verge. Retrieved 29 December 2016.
113. "This Startup Will Sequence Your DNA, So You Can Contribute To Medical Research" (https://www.fastcompany.com/3066775/innovation-agents/this-startup-will-sequence-your-dna-so-you-can-contribute-to-medical-resea). Fast Company. 23 December 2016. Retrieved 29 December 2016.
114. Seife, Charles. "23andMe Is Terrifying, but Not for the Reasons the FDA Thinks" (https://www.scientificamerican.com/article/23andme-is-terrifying-but-not-for-the-reasons-the-fda-thinks/). Scientific American. Retrieved 29 December 2016.
115. Zaleski, Andrew (22 June 2016). "This biotech start-up is betting your genes will yield the next wonder drug" (http://www.cnbc.com/2016/06/22/23andme-thinks-your-genes-are-the-key-to-blockbuster-drugs.html). CNBC. Retrieved 29 December 2016.
116. Regalado, Antonio. "How 23andMe turned your DNA into a $1 billion drug discovery machine" (https://www.technologyreview.com/s/601506/23andme-sells-data-for-drug-search/). MIT Technology Review. Retrieved 29 December 2016.
117. "23andMe reports jump in requests for data in wake of Pfizer depression study | FierceBiotech" (http://www.fiercebiotech.com/it/23andme-reports-jump-requests-for-data-wake-pfizer-depression-study). *fiercebiotech.com*. Retrieved 29 December 2016.
118. Admire Moyo. "Data scientists predict Springbok defeat" (http://www.itweb.co.za/index.php?option=com_content&view=article&id=147241). *www.itweb.co.za*. Retrieved 12 December 2015.
119. Regina Pazvakavambwa. "Predictive analytics, big data transform sports" (http://www.itweb.co.za/index.php?option=com_content&view=article&id=147852). *www.itweb.co.za*. Retrieved 12 December 2015.
120. Rich Miller. "The Lessons of Moneyball for Big Data Analysis" (http://www.datacenterknowledge.com/archives/2010/09/23/the-lessons-of-moneyball-for-big-data-analysis/). *www.datecenterknowledge.com*. Retrieved 12 December 2015.
121. Dave Ryan. "Sports: Where Big Data Finally Makes Sense" (http://www.huffingtonpost.com/dave-ryan/sports-where-big-data-fin_b_8553884.html). *www.huffingtonpost.com*. Retrieved 12 December 2015.
122. Frank Bi. "How Formula One Teams Are Using Big Data To Get The Inside Edge" (http://www.forbes.com/sites/frankbi/2014/11/13/how-formula-one-teams-are-using-big-data-to-get-the-inside-edge/). *www.forbes.com*. Retrieved 12 December 2015.
123. Siwach, Gautam; Esmailpour, Amir (March 2014). *Encrypted Search & Cluster Formation in Big Data* (http://asee-ne.org/proceedings/2014/Student%20Papers/210.pdf) (PDF). ASEE 2014 Zone I Conference (http://ubconferences.org/). University of Bridgeport, Bridgeport, Connecticut, US.
124. "Obama Administration Unveils "Big Data" Initiative:Announces $200 Million In New R&D Investments" (http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf) (PDF). The White House.
125. "AMPLab at the University of California, Berkeley" (http://amplab.cs.berkeley.edu). Amplab.cs.berkeley.edu. Retrieved 2013-03-05.
126. "NSF Leads Federal Efforts in Big Data" (http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607&org=NSF&from=news). National Science Foundation (NSF). 29 March 2012.
127. Timothy Hunter; Teodor Moldovan; Matei Zaharia; Justin Ma; Michael Franklin; Pieter Abbeel; Alexandre Bayen (October 2011). *Scaling the Mobile Millennium System in the Cloud* (https://amplab.cs.berkeley.edu/publication/scaling-the-mobile-millennium-system-in-the-cloud-2/).
128. David Patterson (5 December 2011). "Computer Scientists May Have What It Takes to Help Cure Cancer" (https://www.nytimes.com/2011/12/06/science/david-patterson-enlist-computer-scientists-in-cancer-fight.html?_r=0). The New York Times.
129. "Secretary Chu Announces New Institute to Help Scientists Improve Massive Data Set Research on DOE Supercomputers" (http://energy.gov/articles/secretary-chu-announces-new-institute-help-scientists-improve-massive-data-set-research-doe). "energy.gov".
130. "Governor Patrick announces new initiative to strengthen Massachusetts' position as a World leader in Big Data" (http://www.mass.gov/governor/pressoffice/pressreleases/2012/2012530-governor-announces-big-data-initiative.html). Commonwealth of Massachusetts.
131. "Big Data @ CSAIL" (http://bigdata.csail.mit.edu/). Bigdata.csail.mit.edu. 22 February 2013. Retrieved 2013-03-05.

132. "Big Data Public Private Forum" (http://cordis.europa.eu/search/index.cfm?fuseaction=proj.document&PJ_RCN=13267529). Cordis.europa.eu. 1 September 2012. Retrieved 2013-03-05.
133. "Alan Turing Institute to be set up to research big data" (http://www.bbc.co.uk/news/technology-26651179). BBC News. 19 March 2014. Retrieved 2014-03-19.
134. "Inspiration day at University of Waterloo, Stratford Campus" (http://www.betakit.com/event/inspiration-day-at-university-of-waterloo-stratford-campus/). betakit.com/. Retrieved 2014-02-28.
135. Lee, Jay; Lapira, Edzel; Bagheri, Behrad; Kao, Hung-An (2013). "Recent Advances and Trends in Predictive Manufacturing Systems in Big Data Environment" (http://www.sciencedirect.com/science/article/pii/S2213846313000114). *Manufacturing Letters*. **1** (1): 38–41. doi:10.1016/j.mfglet.2013.09.005 (https://doi.org/10.1016%2Fj.mfglet.2013.09.005).
136. Reips, Ulf-Dietrich; Matzat, Uwe (2014). "Mining "Big Data" using Big Data Services" (http://www.ijis.net/ijis9_1/ijis9_1_editorial_pre.html). *International Journal of Internet Science*. **1** (1): 1–8.
137. Preis, Tobias; Moat,, Helen Susannah; Stanley, H. Eugene; Bishop, Steven R. (2012). "Quantifying the Advantage of Looking Forward" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3320057). *Scientific Reports*. **2**: 350. PMC 3320057 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3320057). PMID 22482034 (https://www.ncbi.nlm.nih.gov/pubmed/22482034). doi:10.1038/srep00350 (https://doi.org/10.1038%2Fsrep00350).
138. Marks, Paul (5 April 2012). "Online searches for future linked to economic success" (http://www.newscientist.com/article/dn21678-online-searches-for-future-linked-to-economic-success.html). *New Scientist*. Retrieved 9 April 2012.
139. Johnston, Casey (6 April 2012). "Google Trends reveals clues about the mentality of richer nations" (https://arstechnica.com/gadgets/news/2012/04/google-trends-reveals-clues-about-the-mentality-of-richer-nations.ars). *Ars Technica*. Retrieved 9 April 2012.
140. Tobias Preis (24 May 2012). "Supplementary Information: The Future Orientation Index is available for download" (http://www.tobiaspreis.de/bigdata/future_orientation_index.pdf) (PDF). Retrieved 2012-05-24.
141. Philip Ball (26 April 2013). "Counting Google searches predicts market movements" (http://www.nature.com/news/counting-google-searches-predicts-market-movements-1.12879). *Nature*. Retrieved 9 August 2013.
142. Tobias Preis, Helen Susannah Moat and H. Eugene Stanley (2013). "Quantifying Trading Behavior in Financial Markets Using Google Trends" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3635219). *Scientific Reports*. **3**: 1684. PMC 3635219 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3635219). PMID 23619126 (https://www.ncbi.nlm.nih.gov/pubmed/23619126). doi:10.1038/srep01684 (https://doi.org/10.1038%2Fsrep01684).
143. Nick Bilton (26 April 2013). "Google Search Terms Can Predict Stock Market, Study Finds" (http://bits.blogs.nytimes.com/2013/04/26/google-search-terms-can-predict-stock-market-study-finds/). *New York Times*. Retrieved 9 August 2013.
144. Christopher Matthews (26 April 2013). "Trouble With Your Investment Portfolio? Google It!" (http://business.time.com/2013/04/26/trouble-with-your-investment-portfolio-google-it/). *TIME Magazine*. Retrieved 9 August 2013.
145. Philip Ball (26 April 2013). "Counting Google searches predicts market movements" (http://www.nature.com/news/counting-google-searches-predicts-market-movements-1.12879). *Nature*. Retrieved 9 August 2013.
146. Bernhard Warner (25 April 2013). "'Big Data' Researchers Turn to Google to Beat the Markets" (http://www.businessweek.com/articles/2013-04-25/big-data-researchers-turn-to-google-to-beat-the-markets). *Bloomberg Businessweek*. Retrieved 9 August 2013.
147. Hamish McRae (28 April 2013). "Hamish McRae: Need a valuable handle on investor sentiment? Google it" (http://www.independent.co.uk/news/business/comment/hamish-mcrae/hamish-mcrae-need-a-valuable-handle-on-investor-sentiment-google-it-8590991.html). *The Independent*. London. Retrieved 9 August 2013.
148. Richard Waters (25 April 2013). "Google search proves to be new word in stock market prediction" (http://www.ft.com/intl/cms/s/0/e5d959b8-acf2-11e2-b27f-00144feabdc0.html). *Financial Times*. Retrieved 9 August 2013.
149. David Leinweber (26 April 2013). "Big Data Gets Bigger: Now Google Trends Can Predict The Market" (http://www.forbes.com/sites/davidleinweber/2013/04/26/big-data-gets-bigger-now-google-trends-can-predict-the-market/). *Forbes*. Retrieved 9 August 2013.
150. Jason Palmer (25 April 2013). "Google searches predict market moves" (http://www.bbc.co.uk/news/science-environment-22293693). *BBC*. Retrieved 9 August 2013.
151. E. Sejdić, "Adapt current tools for use with big data," *Nature*, vol. vol. 507, no. 7492, pp. 306, Mar. 2014.
152. Stanford. "MMDS. Workshop on Algorithms for Modern Massive Data Sets" (http://web.stanford.edu/group/mmds/).
153. Deepan Palguna; Vikas Joshi; Venkatesan Chakaravarthy; Ravi Kothari & L. V. Subramaniam (2015). *Analysis of Sampling Algorithms for Twitter*. *International Joint Conference on Artificial Intelligence*.
154. Kimble, C.; Milolidakis, G. (2015). "Big Data and Business Intelligence: Debunking the Myths". *Global Business and Organizational Excellence*. **35** (1): 23–34. doi:10.1002/joe.21642 (https://doi.org/10.1002%2Fjoe.21642).
155. Chris Anderson (23 June 2008). "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" (https://www.wired.com/science/discoveries/magazine/16-07/pb_theory). *WIRED*.
156. Graham M. (9 March 2012). "Big data and the end of theory?" (https://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory). *The Guardian*. London.
157. "Good Data Won't Guarantee Good Decisions. Harvard Business Review" (http://hbr.org/2012/04/good-data-wont-guarantee-good-decisions/ar/1). *Shah, Shvetank; Horne, Andrew; Capellá, Jaime;*. HBR.org. Retrieved 8 September 2012.

158. Big Data requires Big Visions for Big Change. (https://www.youtube.com/watch?v=UXef6yfJZAI) Hilbert, M. (2014). London: TEDxUCL, x=independently organized TED talks
159. Jonathan Rauch (1 April 2002). "Seeing Around Corners" (https://www.theatlantic.com/magazine/archive/2002/04/seeing-around-corners/302471/). *The Atlantic*.
160. Epstein, J. M., & Axtell, R. L. (1996). Growing Artificial Societies: Social Science from the Bottom Up. A Bradford Book.
161. Delort P., Big data in Biosciences, Big Data Paris, 2012 (http://www.bigdataparis.com/documents/Pierre-Delort-INSERM.pdf#page=5)
162. "Next-generation genomics: an integrative approach" (http://www.cs.cmu.edu/~durand/03-711/2011/Literature/Next-Gen-Genomics-NRG-2010.pdf) (PDF). nature. July 2010. Retrieved 18 October 2016.
163. "BIG DATA IN BIOSCIENCES" (https://www.researchgate.net/publication/283298499_BIG_DATA_IN_BIOSCIENCES). ResearchGate. October 2015. Retrieved 18 October 2016.
164. "Big data: are we making a big mistake?" (https://next.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0). Financial Times. 28 March 2014. Retrieved 20 October 2016.
165. Ohm, Paul. "Don't Build a Database of Ruin" (http://blogs.hbr.org/cs/2012/08/dont_build_a_database_of_ruin.html). Harvard Business Review.
166. Darwin Bond-Graham, *Iron Cagebook – The Logical End of Facebook's Patents (http://www.counterpunch.org/2013/12/03/iron-cagebook/)*, Counterpunch.org, 2013.12.03
167. Darwin Bond-Graham, *Inside the Tech industry's Startup Conference (http://www.counterpunch.org/2013/09/11/inside-the-tech-industrys-startup-conference/)*, Counterpunch.org, 2013.09.11
168. Al-Rodhan, Nayef (2014-09-16). "The Social Contract 2.0: Big Data and the Need to Guarantee Privacy and Civil Liberties - Harvard International Review" (http://hir.harvard.edu/the-social-contract-2-0-big-data-and-the-need-to-guarantee-privacy-and-civil-liberties/). *Harvard International Review*. Retrieved 2017-04-03.
169. danah boyd (29 April 2010). "Privacy and Publicity in the Context of Big Data" (http://www.danah.org/papers/talks/2010/WWW2010.html). *WWW 2010 conference*. Retrieved 2011-04-18.
170. Jones, MB; Schildhauer, MP; Reichman, OJ; Bowers, S (2006). "The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere" (http://www.pnamp.org/sites/default/files/Jones2006_AREES.pdf) (PDF). *Annual Review of Ecology, Evolution, and Systematics*. **37** (1): 519–544. doi:10.1146/annurev.ecolsys.37.091305.110031 (https://doi.org/10.1146%2Fannurev.ecolsys.37.091305.110031).
171. Boyd, D.; Crawford, K. (2012). "Critical Questions for Big Data". *Information, Communication & Society*. **15** (5): 662–679. doi:10.1080/1369118X.2012.678878 (https://doi.org/10.1080%2F1369118X.2012.678878).
172. Failure to Launch: From Big Data to Big Decisions (http://www.fortewares.com/Administrator/userfiles/Banner/forte-wares--pro-active-reporting_EN.pdf), Forte Wares.
173. Gregory Piatetsky (12 August 2014). "Interview: Michael Berthold, KNIME Founder, on Research, Creativity, Big Data, and Privacy, Part 2" (http://www.kdnuggets.com/2014/08/interview-michael-berthold-knime-research-big-data-privacy-part2.html). KDnuggets. Retrieved 2014-08-13.
174. Pelt, Mason. " "Big Data" is an over used buzzword and this Twitter bot proves it" (http://siliconangle.com/blog/2015/10/26/big-data-is-an-over-used-buzzword-and-this-twitter-bot-proves-it/). *siliconangle.com*. SiliconANGLE. Retrieved 4 November 2015.
175. Harford, Tim (28 March 2014). "Big data: are we making a big mistake?" (http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html). *Financial Times*. Financial Times. Retrieved 2014-04-07.
176. Ioannidis, J. P. A. (2005). "Why Most Published Research Findings Are False" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327). *PLoS Medicine*. **2** (8): e124. PMC 1182327 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327). PMID 16060722 (https://www.ncbi.nlm.nih.gov/pubmed/16060722). doi:10.1371/journal.pmed.0020124 (https://doi.org/10.1371%2Fjournal.pmed.0020124).
177. Lohr, Steve; Singer, Natasha (2016-11-10). "How Data Failed Us in Calling an Election" (https://www.nytimes.com/2016/11/10/technology/the-data-said-clinton-would-win-why-you-shouldnt-have-believed-it.html). *The New York Times*. ISSN 0362-4331 (https://www.worldcat.org/issn/0362-4331). Retrieved 2016-11-27.
178. Markman, Jon. "Big Data And The 2016 Election" (http://www.forbes.com/sites/jonmarkman/2016/08/08/big-data-and-the-2016-election/#4802f20846d7). *Forbes*. Retrieved 2016-11-27.

Further reading

- Peter Kinnaird, Inbal Talgam-Cohen, eds. (2012). "Big Data". XRDS: Crossroads, The ACM Magazine for Students. No. 19 (1). Association for Computing Machinery. ISSN 1528-4980. OCLC 779657714.
- Jure Leskovec; Anand Rajaraman; Jeffrey D. Ullman (2014). *Mining of massive datasets*. Cambridge University Press. ISBN 9781107077232. OCLC 888463433.
- Viktor Mayer-Schönberger; Kenneth Cukier (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt. ISBN 9781299903029. OCLC 828620988.

- Press, Gil (2013-05-09). "A Very Short History Of Big Data". *forbes.com*. Jersey City, NJ: Forbes Magazine. Retrieved 2016-09-17.

External links

- Media related to Big data at Wikimedia Commons
- The dictionary definition of big data at Wiktionary

Categories: Big data | Data management | Distributed computing problems | Technology forecasting | Transaction processing

---