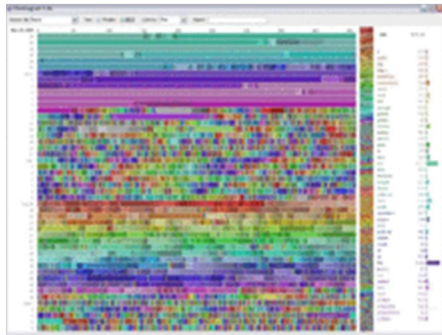


Big data

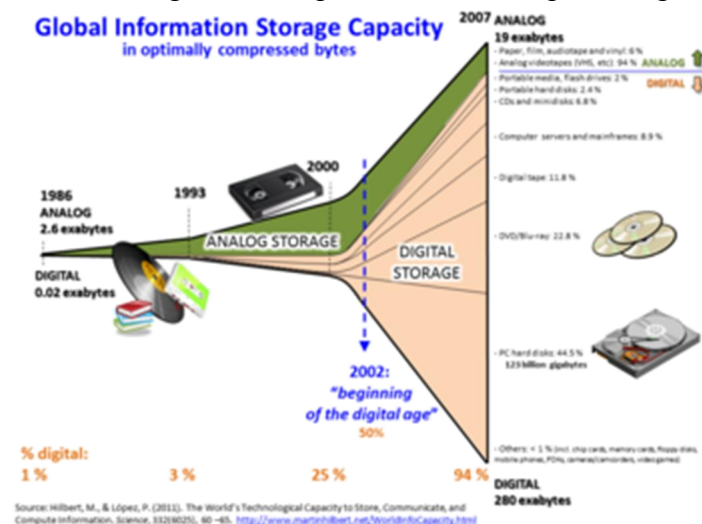
From Wikipedia, the free encyclopedia

Jump to: [navigation](#), [search](#)

This article is about large collections of data. For the graph database, see [Graph database](#). For the band, see [Big Data \(band\)](#).



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



Growth of and Digitization of Global Information Storage Capacity^[1]

Big data is a broad term for **data sets** so large or complex that traditional **data processing** applications are inadequate. Challenges include **analysis**, capture, **data curation**, search, **sharing**, **storage**, **transfer**, **visualization**, **querying** and **information privacy**. The term often refers simply to the use of **predictive analytics** or certain other advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on."^[2] Scientists, business executives, practitioners of medicine, advertising and **governments** alike regularly meet difficulties with large data sets in areas including **Internet search**, finance and **business informatics**. Scientists encounter

limitations in [e-Science](#) work, including [meteorology](#), [genomics](#),^[3] [connectomics](#), complex physics simulations, biology and environmental research.^[4]

Data sets are growing rapidly in part because they are increasingly gathered by cheap and numerous information-sensing [mobile devices](#), aerial ([remote sensing](#)), software logs, [cameras](#), microphones, [radio-frequency identification](#) (RFID) readers and [wireless sensor networks](#).^{[5][6]} The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;^[7] as of 2012, every day 2.5 [exabytes](#) (2.5×10^{18}) of data are created.^[8] One question for large enterprises is determining who should own big data initiatives that affect the entire organization.^[9]

[Relational database management systems](#) and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers".^[10] What is considered "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."^[11]

Contents

[\[hide\]](#)

- [1 Definition](#)
- [2 Characteristics](#)
- [3 Architecture](#)
- [4 Technologies](#)
- [5 Applications](#)
 - [5.1 Government](#)
 - [5.1.1 United States of America](#)
 - [5.1.2 India](#)
 - [5.1.3 United Kingdom](#)
 - [5.2 International development](#)
 - [5.3 Manufacturing](#)
 - [5.3.1 Cyber-physical models](#)
 - [5.4 Healthcare](#)
 - [5.5 Education](#)
 - [5.6 Media](#)
 - [5.6.1 Internet of Things \(IoT\)](#)
 - [5.6.2 Technology](#)
 - [5.7 Private sector](#)
 - [5.7.1 Retail](#)
 - [5.7.2 Retail banking](#)
 - [5.7.3 Real estate](#)
 - [5.8 Science](#)

- [5.8.1 Science and research](#)
 - [5.9 Sports](#)
- [6 Research activities](#)
 - [6.1 Sampling Big Data](#)
- [7 Critique](#)
 - [7.1 Critiques of the big data paradigm](#)
 - [7.2 Critiques of big data execution](#)
- [8 See also](#)
- [9 References](#)
- [10 Further reading](#)
- [11 External links](#)

Definition[[edit](#)]

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.^[12] Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data. 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.^[13]

In a 2001 research report^[14] and related lectures, [META Group](#) (now [Gartner](#)) analyst [Doug Laney](#) 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.^[15] 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".^[16] Additionally, a new V "Veracity" is added by some organizations to describe it,^[17] revisionism challenged by some industry authorities.^[18] The 3Vs have been expanded to other complementary characteristics of big data.^{[19][20]}

- [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^[21]
- [Digital footprint](#): big data is often a cost-free byproduct of digital interaction^[20]

The growing maturity of the concept more starkly delineates the difference between big data and [Business Intelligence](#).^[22]

- 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](#)^[23] to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density^[24] to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.^{[23][25]}

In a popular tutorial article published in IEEE Access Journal,^[26] the authors classified existing definitions of big data into three categories: Attribute Definition, Comparative Definition and Architectural Definition. The authors also presented a big-data technology map that illustrates its key technological evolutions.

Characteristics[\[edit\]](#)

Big data can be described by the following characteristics:^{[19][20]}

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)^{[27][28]}
- 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.^{[29][30]}

Architecture[\[edit\]](#)

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.^[31] and in 2008 acquired [ChoicePoint, Inc.](#)^[32] 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. Currently, HPCC and [Quantcast File System](#)^[33] are the only publicly available platforms capable of analyzing multiple exabytes of data.

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,^[34] so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open-source project named [Hadoop](#).^[35]

[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".^[36] 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.^[37]

Recent studies show 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.^[38]

Big Data Analytics for Manufacturing Applications can be based on a 5C architecture (connection, conversion, cyber, cognition, and configuration).^[39]

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.^[40]

Technologies[\[edit\]](#)

A 2011 [McKinsey Global Institute](#) report characterizes the main components and ecosystem of big data as follows.^[41]

- 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,^[42] such as [multilinear subspace learning](#).^[43] Additional technologies being applied to big data include massively parallel-processing ([MPP](#)) databases, [search-based applications](#), [data mining](#), [distributed file systems](#), [distributed databases](#), [cloud-based](#) infrastructure (applications, storage and computing resources) and the Internet.^[citation needed]

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](#).^[44]

[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](#).^[45]

The practitioners of big data analytics processes are generally hostile to slower shared storage,^[46] 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.^[47]

Applications[[edit](#)]



Bus wrapped with [SAP](#) Big data parked outside [IDF13](#).

Big data has increased the demand of information management specialists in 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.^[2]

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.^[2] Between 1990 and 2005, more than 1 billion people worldwide entered the middle class, which means more people become more literate, which in turn leads 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^[7] and predictions put the amount of internet traffic at 667 exabytes annually by 2014.^[2] According to one estimate, one third of the globally stored information is in the form of alphanumeric text and still image data,^[48] 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.^[49]

Government[[edit](#)]

The use and adoption of Big Data within governmental processes is beneficial and allows efficiencies in terms of cost, productivity, and innovation. That said, this process 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 the thought leading examples within the Governmental Big Data space.

United States of America[[edit](#)]

- 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.^[50] The initiative is composed of 84 different big data programs spread across six departments.^[51]
- Big data analysis played a large role in [Barack Obama](#)'s successful [2012 re-election campaign](#).^[52]
 - The [United States Federal Government](#) owns six of the ten most powerful [supercomputers](#) in the world.^[53]
 - The [Utah Data Center](#) is a data center currently being 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](#).^{[54][55][56]}

India[\[edit\]](#)

- Big data analysis helped in parts, responsible for the NDA to win [Indian General Election 2014](#).^[57]
- The Indian Government utilises numerous techniques to ascertain how the Indian electorate is responding to government action, as well as ideas for policy augmentation.

United Kingdom[\[edit\]](#)

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/most up-to-date drugs take some time to filter through to the general patient.^[citation needed]
- 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.^[citation needed]

International development[\[edit\]](#)

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](#).^{[58][59]} 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.^{[60][61][62]} 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.^[60]

Manufacturing[\[edit\]](#)

Based on TCS 2013 Global Trend Study, improvements in supply planning and product quality provide the greatest benefit of big data for manufacturing.^[63] 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.^[64] 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).^[citation needed]

Cyber-physical models[\[edit\]](#)

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.^[30]

Healthcare[\[edit\]](#)

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.^[65]

Education^[edit]

A [McKinsey Global Institute](#) study found a shortage 1.5 million highly trained data professionals and managers^[41] and a number of universities^[66] 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](#).^[67]

Media^[edit]

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 tap 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 consumers 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.^[68]

- Targeting of consumers (for advertising by marketers)
- Data-capture

Internet of Things (IoT)^[edit]

Main article: [Internet of Things](#)

Big Data and the IoT work in conjunction. From a media perspective, data is the key derivative of device inter connectivity and allows accurate targeting. The [Internet of Things](#), with the help of big data, therefore transforms the media industry, companies and even governments, opening up a new era of economic growth and competitiveness. The intersection of people, data and intelligent algorithms have far-reaching impacts on media efficiency. The wealth of data generated allows an elaborate layer on the present targeting mechanisms of the industry.

Technology^[edit]

- [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.^[69]

- [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.^[70]
- Facebook handles 50 billion photos from its user base.^[71]
- As of August 2012, [Google](#) was handling roughly 100 billion searches per month.^[72]
- [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.^[73]

Private sector[[edit](#)]

Retail[[edit](#)]

- [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](#).^[2]

Retail banking[[edit](#)]

- FICO Card Detection System protects accounts world-wide.^[74]
- The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates.^{[75][76]}

Real estate[[edit](#)]

- [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.^[77]

Science[[edit](#)]

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%^[78] of these streams, there are 100 collisions of interest per second.^{[79][80][81]}

- 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.^{[82][83]} It is considered one of the most ambitious scientific projects ever undertaken.^[citation needed]

Science and research[\[edit\]](#)

- 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. 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.^[2]
- 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](#).^[84]
- The [Nasa](#) Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.^[85]
- 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.^[citation needed]

Sports[\[edit\]](#)

Big data can be used to improve training and understanding competitors. Besides, it is possible to predict winners in a match using big data analytics.^[86] Future performance of players could be predicted as well. Thus, players' value and salary is determined by data collected throughout the season.^[87]

The movie [MoneyBall](#) demonstrates how big data could be used to scout players and also identify undervalued players.^[88]

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. Then, this data is transferred to team headquarters in United Kingdom through fiber optic cables that could carry data at the speed of light.^[89] 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.^[90]

Research activities[\[edit\]](#)

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.^[91]

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.^[92]

The initiative included a National Science Foundation "Expeditions in Computing" grant of \$10 million over 5 years to the AMPLab^[93] at the University of California, Berkeley.^[94] 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^[95] to fighting cancer.^[96]

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,^[97] 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.^[98] 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.^[99]

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](#).^[100]

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.^[101]

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.^[102]

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.^[103] 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.^[104] Often these APIs are provided for free.^[104] [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.^{[105][106][107]} 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](#)'.^[108] 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.^[109] Their analysis of [Google](#) search volume for 98 terms of varying financial relevance, published in [Scientific Reports](#),^[110] suggests that increases in search volume for financially relevant search terms tend to precede large losses in financial markets.^{[111][112][113][114][115][116][117][118]}

Big data sets come with algorithmic challenges that previously did not exist. Hence, there is a need to fundamentally change the processing ways.^[119]

Sampling Big Data[\[edit\]](#)

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.

There has been some work done in Sampling algorithms for Big Data. A theoretical formulation for sampling Twitter data has been developed.^[120]

Critique^[edit]

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.^[121]

Critiques of the big data paradigm^[edit]

"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".^[12] 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: focusing in particular on the notion that big data must always be contextualized in their social, economic, and political contexts.^[122] 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 analyzed, must be complemented by "big judgment," according to an article in the Harvard Business Review.^[123]

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".^[60] Fed by a large number of data on past experiences, algorithms can predict future development if the future is similar to the past. If the systems dynamics of the future change, the past can say little about the future. For this, it would be necessary to have a thorough understanding of the systems dynamic, which implies theory.^[124] As a response to this critique it has been suggested to combine big data approaches with computer simulations, such as [agent-based models](#)^[60] 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.^{[125][126]} 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.^[127] 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.^[citation needed] In the massive approaches it is the formulation of a relevant hypothesis to explain the data that is the limiting factor.^[citation needed] The search logic is reversed and the limits of induction ("Glory of Science and Philosophy scandal", [C. D. Broad](#), 1926) are to be considered.^[citation needed]

[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.^{[128][129][130]}

Critiques of big data execution[\[edit\]](#)

Big data has been called a "fad" in scientific research and its use was even made fun of as an absurd practice in a satirical example on "pig data".^[104] 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.^[131] 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.^[132] In the provocative article "Critical Questions for Big Data",^[133] 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".^[133] 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.^[134]

Big data analysis is often shallow compared to analysis of smaller data sets.^[135] In many big data projects, there is no large data analysis happening, but the challenge is the [extract, transform, load](#) part of data preprocessing.^[135]

Big data is a [buzzword](#) and a "vague term",^{[136][137]} but at the same time an "obsession"^[137] 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"^[138] 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.

See also [[edit](#)]



Information technology portal

For a list of companies, and tools, see also: [Category:Big data](#).

- Big memory
- Data defined storage
- Data lineage
- Data science
- Machine learning
- Statistics

References[[edit](#)]

1. [Jump up](#) ^ [Source](#)
2. ^ [Jump up to:](#) ^{a b c d e f} "[*Data, data everywhere*](#)". *The Economist*. 25 February 2010. Retrieved 9 December 2012.
3. [Jump up](#) ^ "[*Community cleverness required*](#)". *Nature* **455** (7209): 1. 4 September 2008.
doi:10.1038/455001a.
4. [Jump up](#) ^ Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011). "Challenges and Opportunities of Open Data in Ecology". *Science* **331** (6018): 703–5.
doi:10.1126/science.1197962. *PMID* [21311007](#).
5. [Jump up](#) ^ Hellerstein, Joe (9 November 2008). "[*Parallel Programming in the Age of Big Data*](#)". Gigaom Blog.
6. [Jump up](#) ^ Segaran, Toby; Hammerbacher, Jeff (2009). [*Beautiful Data: The Stories Behind Elegant Data Solutions*](#). O'Reilly Media. p. 257. *ISBN* [978-0-596-15711-1](#).
7. ^ [Jump up to:](#) ^{a b} Hilbert, Martin; López, Priscila (2011). "[*The World's Technological Capacity to Store, Communicate, and Compute Information*](#)". *Science* **332** (6025): 60–65.
doi:10.1126/science.1200970. *PMID* [21310967](#).

8. **Jump up** ^ ["IBM What is big data? – Bringing big data to the enterprise"](#). [www.ibm.com](#). Retrieved 2013-08-26.
9. **Jump up** ^ Oracle and FSN, ["Mastering Big Data: CFO Strategies to Transform Insight into Opportunity"](#). December 2012
10. **Jump up** ^ Jacobs, A. (6 July 2009). ["The Pathologies of Big Data"](#). *ACMQueue*.
11. **Jump up** ^ Magoulas, Roger; Lorica, Ben (February 2009). ["Introduction to Big Data"](#). Release 2.0 (Sebastopol CA: O'Reilly Media) (11).
12. ^ **Jump up to:** ^a ^b Snijders, C.; Matzat, U.; Reips, U.-D. (2012). ["Big Data': Big gaps of knowledge in the field of Internet"](#). *International Journal of Internet Science* 7: 1–5.
13. **Jump up** ^ 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.
14. **Jump up** ^ Laney, Douglas. ["3D Data Management: Controlling Data Volume, Velocity and Variety"](#) (PDF). Gartner. Retrieved 6 February 2001.
15. **Jump up** ^ Beyer, Mark. ["Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data"](#). Gartner. Archived from the original on 10 July 2011. Retrieved 13 July 2011.
16. **Jump up** ^ De Mauro, Andrea; Greco, Marco; Grimaldi, Michele (2015). ["What is big data? A consensual definition and a review of key research topics"](#). *AIP Conference Proceedings* 1644: 97–104. doi:10.1063/1.4907823.
17. **Jump up** ^ ["What is Big Data?"](#). Villanova University.
18. **Jump up** ^ Grimes, Seth. ["Big Data: Avoid 'Wanna V' Confusion"](#). *InformationWeek*. Retrieved 5 January 2016.
19. ^ **Jump up to:** ^a ^b Hilbert, Martin. ["Big Data for Development: A Review of Promises and Challenges. Development Policy Review."](#). martinhilbert.net. Retrieved 2015-10-07.
20. ^ **Jump up to:** ^a ^b ^c Hilbert, M. (2015). Digital Technology and Social Change [Open Online Course at the University of California] (freely available). <https://www.youtube.com/watch?v=XRVIh1h47sA&index=51&list=PLtjBSCvWCU3rNm46D3R85efM0hrzjuAIg> Retrieved from <https://canvas.instructure.com/courses/949415>
21. **Jump up** ^ Mayer-Schönberger, V., & Cukier, K. (2013). Big data: a revolution that will transform how we live, work and think. London: John Murray.
22. **Jump up** ^ <http://www.bigdataparis.com/presentation/mercredi/PDelort.pdf?PHPSESSID=tv7k70pcr3egpi2r6fi3qbjtj6#page=4>
23. ^ **Jump up to:** ^a ^b Billings S.A. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains". Wiley, 2013
24. **Jump up** ^ Delort P., Big data Paris 2013 <http://www.andsi.fr/tag/dsi-big-data/>
25. **Jump up** ^ Delort P., Big Data car Low-Density Data ? La faible densité en information comme facteur discriminant <http://lecercle.lesechos.fr/entrepreneur/tendances-innovation/221169222/big-data-low-density-data-faible-densite-information-com>
26. **Jump up** ^ Hu, Han; Wen, Yonggang; Chua, Tat-Seng; Li, Xuelong (2014). ["Towards scalable systems for big data analytics: a technology tutorial"](#). *IEEE Access* 2: 652–687. doi:10.1109/ACCESS.2014.2332453.
27. **Jump up** ^ Wu, D., Liu, X., Hebert, S., Gentzsch, W., Terpenney, 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.
28. **Jump up** ^ 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.
29. **Jump up** ^ Lee, Jay; Bagheri, Behrad; Kao, Hung-An (2014). ["Recent Advances and Trends of Cyber-Physical Systems and Big Data Analytics in Industrial Informatics"](#). *IEEE Int. Conference on Industrial Informatics (INDIN)* 2014.

30. ^ [Jump up to:](#) ^{a b} Lee, Jay; Lapira, Edzel; Bagheri, Behrad; Kao, Hung-an. *"Recent advances and trends in predictive manufacturing systems in big data environment"*. *Manufacturing Letters* **1** (1): 38–41. doi:10.1016/j.mfglet.2013.09.005.
31. [Jump up](#) ^ *"LexisNexis To Buy Seisint For \$775 Million"*. *Washington Post*. Retrieved 15 July 2004.
32. [Jump up](#) ^ *"LexisNexis Parent Set to Buy ChoicePoint"*. *Washington Post*. Retrieved 22 February 2008.
33. [Jump up](#) ^ *"Quantcast Opens Exabyte-Ready File System"*. *www.datanami.com*. Retrieved 1 October 2012.
34. [Jump up](#) ^ Bertolucci, Jeff *"Hadoop: From Experiment To Leading Big Data Platform"*, *"Information Week"*, 2013. Retrieved on 14 November 2013.
35. [Jump up](#) ^ Webster, John. *"MapReduce: Simplified Data Processing on Large Clusters"*, *"Search Storage"*, 2004. Retrieved on 25 March 2013.
36. [Jump up](#) ^ *"Big Data Solution Offering"*. MIKE2.0. Retrieved 8 December 2013.
37. [Jump up](#) ^ *"Big Data Definition"*. MIKE2.0. Retrieved 9 March 2013.
38. [Jump up](#) ^ Boja, C; Pocovnicu, A; Bătăgan, L. (2012). *"Distributed Parallel Architecture for Big Data"*. *Informatica Economica* **16** (2): 116–127.
39. [Jump up](#) ^ *Intelligent Maintenance System*
40. [Jump up](#) ^ http://www.heltech.com/sites/default/files/solving_key_businesschallenges_with_big_data_lake_0.pdf
41. ^ [Jump up to:](#) ^{a b} 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"*. McKinsey Global Institute. Retrieved January 16, 2016.
42. [Jump up](#) ^ *"Future Directions in Tensor-Based Computation and Modeling"* (PDF). May 2009.
43. [Jump up](#) ^ Lu, Haiping; Plataniotis, K.N.; Venetsanopoulos, A.N. (2011). *"A Survey of Multilinear Subspace Learning for Tensor Data"* (PDF). *Pattern Recognition* **44** (7): 1540–1551. doi:10.1016/j.patcog.2011.01.004.
44. [Jump up](#) ^ Monash, Curt (30 April 2009). *"eBay's two enormous data warehouses"*. Monash, Curt (6 October 2010). *"eBay followup – Greenplum out, Teradata > 10 petabytes, Hadoop has some value, and more"*.
45. [Jump up](#) ^ *"Resources on how Topological Data Analysis is used to analyze big data"*. Ayasdi.
46. [Jump up](#) ^ CNET News (1 April 2011). *"Storage area networks need not apply"*.
47. [Jump up](#) ^ *"How New Analytic Systems will Impact Storage"*. September 2011.
48. [Jump up](#) ^ *"What Is the Content of the World's Technologically Mediated Information and Communication Capacity: How Much Text, Image, Audio, and Video?"*, Martin Hilbert (2014), *The Information Society*; free access to the article through this link: http://www.martinhilbert.net/WhatsTheContent_Hilbert.pdf
49. [Jump up](#) ^ Rajpurohit, Anmol (11 July 2014). *"Interview: Amy Gershkoff, Director of Customer Analytics & Insights, eBay on How to Design Custom In-House BI Tools"*. 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."
50. [Jump up](#) ^ Kalil, Tom. *"Big Data is a Big Deal"*. White House. Retrieved 26 September 2012.
51. [Jump up](#) ^ Executive Office of the President (March 2012). *"Big Data Across the Federal Government"* (PDF). White House. Retrieved 26 September 2012.
52. [Jump up](#) ^ Lampitt, Andrew. *"The real story of how big data analytics helped Obama win"*. *Infoworld*. Retrieved 31 May 2014.
53. [Jump up](#) ^ Hoover, J. Nicholas. *"Government's 10 Most Powerful Supercomputers"*. *Information Week*. UBM. Retrieved 26 September 2012.
54. [Jump up](#) ^ Bamford, James (15 March 2012). *"The NSA Is Building the Country's Biggest Spy Center (Watch What You Say)"*. *Wired Magazine*. Retrieved 2013-03-18.
55. [Jump up](#) ^ *"Groundbreaking Ceremony Held for \$1.2 Billion Utah Data Center"*. National Security Agency Central Security Service. Retrieved 2013-03-18.

56. [Jump up](#) ^ Hill, Kashmir. *"Blueprints of NSA's Ridiculously Expensive Data Center in Utah Suggest It Holds Less Info Than Thought"*. *Forbes*. Retrieved 2013-10-31.
57. [Jump up](#) ^ *"News: Live Mint"*. Are Indian companies making enough sense of Big Data?. *Live Mint*. 23 June 2014. Retrieved 2014-11-22.
58. [Jump up](#) ^ UN GLocal Pulse (2012). *Big Data for Development: Opportunities and Challenges* (White p. by Letouzé, E.). New York: United Nations
59. [Jump up](#) ^ WEF (World Economic Forum), & Vital Wave Consulting. (2012). *Big Data, Big Impact: New Possibilities for International Development*. World Economic Forum. Retrieved 24 August 2012, from <http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development>
60. ^ [Jump up to: ^a ^b ^c ^d](#) *"Big Data for Development: From Information- to Knowledge Societies"*, Martin Hilbert (2013), SSRN Scholarly Paper No. ID 2205145). Rochester, NY: Social Science Research Network; <http://papers.ssrn.com/abstract=2205145>
61. [Jump up](#) ^ *"Elena Kvochko, Four Ways To talk About Big Data (Information Communication Technologies for Development Series)"*. *worldbank.org*. Retrieved 2012-05-30.
62. [Jump up](#) ^ *"Daniele Medri: Big Data & Business: An on-going revolution"*. *Statistics Views*. 21 October 2013.
63. [Jump up](#) ^ *"Manufacturing: Big Data Benefits and Challenges"*. *TCS Big Data Study*. Mumbai, India: *Tata Consultancy Services Limited*. Retrieved 2014-06-03.
64. [Jump up](#) ^ 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).
65. [Jump up](#) ^ *"Impending Challenges for the Use of Big Data"*. [doi:10.1016/j.ijrobp.2015.10.060](https://doi.org/10.1016/j.ijrobp.2015.10.060).
66. [Jump up](#) ^ *"Degrees in Big Data: Fad or Fast Track to Career Success"*. *Forbes*. Retrieved 2016-02-21.
67. [Jump up](#) ^ *"NY gets new bootcamp for data scientists: It's free, but harder to get into than Harvard"*. *Venture Beat*. Retrieved 2016-02-21.
68. [Jump up](#) ^ 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.
69. [Jump up](#) ^ Tay, Liz. *"Inside eBay's 90PB data warehouse"*. *ITNews*. Retrieved 2016-02-12.
70. [Jump up](#) ^ Layton, Julia. *"Amazon Technology"*. *Money.howstuffworks.com*. Retrieved 2013-03-05.
71. [Jump up](#) ^ *"Scaling Facebook to 500 Million Users and Beyond"*. *Facebook.com*. Retrieved 2013-07-21.
72. [Jump up](#) ^ *"Google Still Doing at Least 1 Trillion Searches Per Year"*. *Search Engine Land*. 16 January 2015. Retrieved 15 April 2015.
73. [Jump up](#) ^ Lamb, Charles. *"Oracle NoSQL Database Exceeds 1 Million Mixed YCSB Ops/Sec"*.
74. [Jump up](#) ^ *"FICO® Falcon® Fraud Manager"*. *Fico.com*. Retrieved 2013-07-21.
75. [Jump up](#) ^ *"eBay Study: How to Build Trust and Improve the Shopping Experience"*. *Knowwpcarey.com*. 8 May 2012. Retrieved 2015-12-20.
76. [Jump up](#) ^ *Leading Priorities for Big Data for Business and IT*. *eMarketer*. October 2013. Retrieved January 2014.
77. [Jump up](#) ^ Wingfield, Nick (12 March 2013). *"Predicting Commutes More Accurately for Would-Be Home Buyers – NYTimes.com"*. *Bits.blogs.nytimes.com*. Retrieved 2013-07-21.
78. [Jump up](#) ^ Alexandru, Dan. *"Prof"* (PDF). *cds.cern.ch*. CERN. Retrieved 24 March 2015.
79. [Jump up](#) ^ *"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."*. CERN-Brochure-2010-006-Eng. LHC Brochure, English version. CERN. Retrieved 20 January 2013.
80. [Jump up](#) ^ *"LHC Guide, English version. A collection of facts and figures about the Large Hadron Collider (LHC) in the form of questions and answers."*. CERN-Brochure-2008-001-Eng. LHC Guide, English version. CERN. Retrieved 20 January 2013.
81. [Jump up](#) ^ Brumfiel, Geoff (19 January 2011). *"High-energy physics: Down the petabyte highway"*. *Nature* **469**. pp. 282–83. [doi:10.1038/469282a](https://doi.org/10.1038/469282a).

82. [Jump up](#) ^ <http://www.zurich.ibm.com/pdf/astron/CeBIT%202013%20Background%20DOME.pdf>
83. [Jump up](#) ^ ["Future telescope array drives development of exabyte processing"](#). *Ars Technica*. Retrieved 15 April 2015.
84. [Jump up](#) ^ Delort P., OECD ICCP Technology Foresight Forum, 2012. http://www.oecd.org/sti/ieconomy/Session_3_Delort.pdf#page=6
85. [Jump up](#) ^ Webster, Phil. ["Supercomputing the Climate: NASA's Big Data Mission"](#). CSC World. Computer Sciences Corporation. Retrieved 2013-01-18.
86. [Jump up](#) ^ Admire Moyo. ["Data scientists predict Springbok defeat"](#). www.itweb.co.za. Retrieved 12 December 2015.
87. [Jump up](#) ^ Regina Pazvakavambwa. ["Predictive analytics, big data transform sports"](#). www.itweb.co.za. Retrieved 12 December 2015.
88. [Jump up](#) ^ Rich Miller. ["The Lessons of Moneyball for Big Data Analysis"](#). www.datecenterknowledge.com. Retrieved 12 December 2015.
89. [Jump up](#) ^ Dave Ryan. ["Sports: Where Big Data Finally Makes Sense"](#). www.huffingtonpost.com. Retrieved 12 December 2015.
90. [Jump up](#) ^ Frank Bi. ["How Formula One Teams Are Using Big Data To Get The Inside Edge"](#). www.forbes.com. Retrieved 12 December 2015.
91. [Jump up](#) ^ Siwach, Gautam; Esmailpour, Amir (March 2014). ["Encrypted Search & Cluster Formation in Big Data"](#) (PDF). *ASEE 2014 Zone I Conference*. University of Bridgeport, Bridgeport, Connecticut, US.
92. [Jump up](#) ^ ["Obama Administration Unveils "Big Data" Initiative; Announces \\$200 Million In New R&D Investments"](#) (PDF). The White House.
93. [Jump up](#) ^ ["AMPLab at the University of California, Berkeley"](#). [Amplab.cs.berkeley.edu](http://amplab.cs.berkeley.edu). Retrieved 2013-03-05.
94. [Jump up](#) ^ ["NSF Leads Federal Efforts in Big Data"](#). National Science Foundation (NSF). 29 March 2012.
95. [Jump up](#) ^ Timothy Hunter; Teodor Moldovan; Matei Zaharia; Justin Ma; Michael Franklin; Pieter Abbeel; Alexandre Bayen (October 2011). ["Scaling the Mobile Millennium System in the Cloud"](#).
96. [Jump up](#) ^ David Patterson (5 December 2011). ["Computer Scientists May Have What It Takes to Help Cure Cancer"](#). *The New York Times*.
97. [Jump up](#) ^ ["Secretary Chu Announces New Institute to Help Scientists Improve Massive Data Set Research on DOE Supercomputers"](#). ["energy.gov"](http://energy.gov).
98. [Jump up](#) ^ ["Governor Patrick announces new initiative to strengthen Massachusetts' position as a World leader in Big Data"](#). Commonwealth of Massachusetts.
99. [Jump up](#) ^ ["Big Data @ CSAIL"](#). Bigdata.csail.mit.edu. 22 February 2013. Retrieved 2013-03-05.
100. [Jump up](#) ^ ["Big Data Public Private Forum"](#). Cordis.europa.eu. 1 September 2012. Retrieved 2013-03-05.
101. [Jump up](#) ^ ["Alan Turing Institute to be set up to research big data"](#). *BBC News*. 19 March 2014. Retrieved 2014-03-19.
102. [Jump up](#) ^ ["Inspiration day at University of Waterloo, Stratford Campus"](#). <http://www.betakit.com/>. Retrieved 2014-02-28. External link in `|publisher=` ([help](#))
103. [Jump up](#) ^ Lee, Jay; Lapira, Edzel; Bagheri, Behrad; Kao, Hung-An (2013). ["Recent Advances and Trends in Predictive Manufacturing Systems in Big Data Environment"](#). *Manufacturing Letters* **1** (1): 38–41. doi:10.1016/j.mfglet.2013.09.005.
104. [Jump up to:](#) ^a ^b ^c Reips, Ulf-Dietrich; Matzat, Uwe (2014). ["Mining "Big Data" using Big Data Services"](#). *International Journal of Internet Science* **1** (1): 1–8.
105. [Jump up](#) ^ Preis, Tobias; Moat, Helen Susannah; Stanley, H. Eugene; Bishop, Steven R. (2012). ["Quantifying the Advantage of Looking Forward"](#). *Scientific Reports* **2**: 350. doi:10.1038/srep00350. PMC 3320057. PMID 22482034.
106. [Jump up](#) ^ Marks, Paul (5 April 2012). ["Online searches for future linked to economic success"](#). *New Scientist*. Retrieved 9 April 2012.
107. [Jump up](#) ^ Johnston, Casey (6 April 2012). ["Google Trends reveals clues about the mentality of richer nations"](#). *Ars Technica*. Retrieved 9 April 2012.

108. [Jump up](#) ^ Tobias Preis (24 May 2012). ["Supplementary Information: The Future Orientation Index is available for download"](#) (PDF). Retrieved 2012-05-24.
109. [Jump up](#) ^ Philip Ball (26 April 2013). ["Counting Google searches predicts market movements"](#). *Nature*. Retrieved 9 August 2013.
110. [Jump up](#) ^ Tobias Preis, Helen Susannah Moat and H. Eugene Stanley (2013). ["Quantifying Trading Behavior in Financial Markets Using Google Trends"](#). *Scientific Reports* **3**: 1684. doi:10.1038/srep01684. PMC 3635219. PMID 23619126.
111. [Jump up](#) ^ Nick Bilton (26 April 2013). ["Google Search Terms Can Predict Stock Market, Study Finds"](#). *New York Times*. Retrieved 9 August 2013.
112. [Jump up](#) ^ Christopher Matthews (26 April 2013). ["Trouble With Your Investment Portfolio? Google It!"](#). *TIME Magazine*. Retrieved 9 August 2013.
113. [Jump up](#) ^ Philip Ball (26 April 2013). ["Counting Google searches predicts market movements"](#). *Nature*. Retrieved 9 August 2013.
114. [Jump up](#) ^ Bernhard Warner (25 April 2013). ["'Big Data' Researchers Turn to Google to Beat the Markets"](#). *Bloomberg Businessweek*. Retrieved 9 August 2013.
115. [Jump up](#) ^ Hamish McRae (28 April 2013). ["Hamish McRae: Need a valuable handle on investor sentiment? Google it"](#). *The Independent* (London). Retrieved 9 August 2013.
116. [Jump up](#) ^ Richard Waters (25 April 2013). ["Google search proves to be new word in stock market prediction"](#). *Financial Times*. Retrieved 9 August 2013.
117. [Jump up](#) ^ David Leinweber (26 April 2013). ["Big Data Gets Bigger: Now Google Trends Can Predict The Market"](#). *Forbes*. Retrieved 9 August 2013.
118. [Jump up](#) ^ Jason Palmer (25 April 2013). ["Google searches predict market moves"](#). *BBC*. Retrieved 9 August 2013.
119. [Jump up](#) ^ E. Sejdić, "Adapt current tools for use with big data," *Nature*, vol. 507, no. 7492, pp. 306, Mar. 2014.
120. [Jump up](#) ^ Deepan Palguna, Vikas Joshi, Venkatesan Chakaravarthy, Ravi Kothari and L. V. Subramaniam (2015). Analysis of Sampling Algorithms for Twitter. [International Joint Conference on Artificial Intelligence](#).
121. [Jump up](#) ^ 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.
122. [Jump up](#) ^ Graham M. (9 March 2012). ["Big data and the end of theory?"](#). *The Guardian* (London).
123. [Jump up](#) ^ ["Good Data Won't Guarantee Good Decisions. Harvard Business Review"](#). Shah, Shvetank; Horne, Andrew; Capellá, Jaime;. *HBR.org*. Retrieved 8 September 2012.
124. [Jump up](#) ^ Anderson, C. (23 June 2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, (Science: Discoveries). http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
125. [Jump up](#) ^ Rauch, J. (2002). Seeing Around Corners. *The Atlantic*, (April), 35–48. <http://www.theatlantic.com/magazine/archive/2002/04/seeing-around-corners/302471/>
126. [Jump up](#) ^ Epstein, J. M., & Axtell, R. L. (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. A Bradford Book.
127. [Jump up](#) ^ Delort P., Big data in Biosciences, Big Data Paris, 2012 <http://www.bigdataparis.com/documents/Pierre-Delort-INSERM.pdf#page=5>
128. [Jump up](#) ^ Ohm, Paul. ["Don't Build a Database of Ruin"](#). *Harvard Business Review*.
129. [Jump up](#) ^ Darwin Bond-Graham, [Iron Cagebook – The Logical End of Facebook's Patents](#), [Counterpunch.org](#), 2013.12.03
130. [Jump up](#) ^ Darwin Bond-Graham, [Inside the Tech industry's Startup Conference](#), [Counterpunch.org](#), 2013.09.11
131. [Jump up](#) ^ danah boyd (29 April 2010). ["Privacy and Publicity in the Context of Big Data"](#). *WWW 2010 conference*. Retrieved 2011-04-18.
132. [Jump up](#) ^ Jones, MB; Schildhauer, MP; Reichman, OJ; Bowers, S (2006). ["The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere"](#) (PDF). *Annual Review of Ecology, Evolution, and Systematics* **37** (1): 519–544. doi:10.1146/annurev.ecolsys.37.091305.110031.

133. ^ [Jump up to:](#) ^a ^b Boyd, D.; Crawford, K. (2012). "Critical Questions for Big Data". *Information, Communication & Society* **15** (5): 662–679. doi:[10.1080/1369118X.2012.678878](#).
134. [Jump up](#) ^ [Failure to Launch: From Big Data to Big Decisions](#), Forte Wares.
135. ^ [Jump up to:](#) ^a ^b Gregory Piatetsky (12 August 2014). "[Interview: Michael Berthold, KNIME Founder, on Research, Creativity, Big Data, and Privacy, Part 2](#)". *KDnuggets*. Retrieved 2014-08-13.
136. [Jump up](#) ^ Pelt, Mason. "[\"Big Data\" is an over used buzzword and this Twitter bot proves it](#)". *siliconangle.com*. SiliconANGLE. Retrieved 4 November 2015.
137. ^ [Jump up to:](#) ^a ^b Harford, Tim (28 March 2014). "[Big data: are we making a big mistake?](#)". *Financial Times*. *Financial Times*. Retrieved 2014-04-07.
138. [Jump up](#) ^ Ioannidis, J. P. A. (2005). "[Why Most Published Research Findings Are False](#)". *PLoS Medicine* **2** (8): e124. doi:[10.1371/journal.pmed.0020124](#). PMC [1182327](#). PMID [16060722](#).


Further reading[\[edit\]](#)

- Sharma, Sugam; Tim, Udoyara S; Wong, Johnny; Gadia, Shashi; Sharma, Subhash (2014). "[A BRIEF REVIEW ON LEADING BIG DATA MODELS](#)". *Data Science Journal* **13**.
- [Big Data Computing and Clouds: Challenges, Solutions, and Future Directions](#). Marcos D. Assuncao, Rodrigo N. Calheiros, Silvia Bianchi, Marco A. S. Netto, Rajkumar Buyya. Technical Report CLOUDS-TR-2013-1, Cloud Computing and Distributed Systems Laboratory, The University of Melbourne, 17 December 2013.
- [Encrypted search & cluster formation in Big Data](#). Gautam Siwach, Dr. A. Esmailpour. American Society for Engineering Education, Conference at the University of Bridgeport, Bridgeport, Connecticut 3–5 April 2014.
- ["Big Data for Good"](#) (PDF). ODBMS.org. 5 June 2012. Retrieved 2013-11-12.
- ["The Rise of Industrial Big Data"](#). GE Intelligent Platforms. Retrieved 2013-11-12.
- Stark, John (2015). *Product Lifecycle Management: Vol 2. The Devil is in the Details. Appendix A: PLM and Big Data*. Springer. ISBN [9783319244341](#).
- [History of Big Data Timeline](#). A visual history of Big Data with links to supporting articles.
- Hu, Han; Wen, Yonggang; Chua, Tat-Seng; Li, Xuelong (2014). "[Towards scalable systems for big data analytics: a technology tutorial](#)". *IEEE Access* **2**: 652–687. doi:[10.1109/ACCESS.2014.2332453](#).

External links[\[edit\]](#)



Wikimedia Commons has media related to [Big data](#).

-  The dictionary definition of [big data](#) at Wiktionary

[\[show\]](#)

- [y](#)
- [t](#)
- [e](#)

Database models

[[show](#)]

- [y](#)
- [t](#)
- [e](#)

Database management systems

[[show](#)]

- [y](#)
- [t](#)
- [e](#)

Software engineering

Developmental

- [Agile](#)
- [EUP](#)
- [Executable UML](#)
- [Incremental model](#)
- [Iterative model](#)
- [Prototype model](#)
- [RAD](#)
- [UP](#)

Other

- [SPICE](#)
- [CMMI](#)
- [Data model](#)
- [ER model](#)

		<ul style="list-style-type: none"> • Function model • Information model • Metamodeling • Object model • Systems model • View model
	Languages	<ul style="list-style-type: none"> • IDEF • UML • SysML
Authority control		<ul style="list-style-type: none"> • GND: 4802620-7 • NDL: 001147262

Retrieved from "https://en.wikipedia.org/w/index.php?title=Big_data&oldid=706212630"
Categories:

- [Database management systems](#)
- [Big data](#)
- [Data management](#)
- [Distributed computing problems](#)
- [Technology forecasting](#)
- [Transaction processing](#)

Hidden categories:

- [CS1 errors: external links](#)
- [Articles containing potentially dated statements from 2012](#)
- [All articles containing potentially dated statements](#)
- [All articles with unsourced statements](#)
- [Articles with unsourced statements from September 2011](#)
- [Articles containing potentially dated statements from 2011](#)
- [Articles with unsourced statements from April 2015](#)
- [Articles with unsourced statements from October 2015](#)
- [Articles with unsourced statements from July 2015](#)
- [Commons category with local link different than on Wikidata](#)
- [Use dmy dates from December 2015](#)
- [Wikipedia articles with GND identifiers](#)

Navigation menu

Personal tools

- Not logged in
- [Talk](#)
- [Contributions](#)
- [Create account](#)
- [Log in](#)

Namespaces

- [Article](#)
- [Talk](#)

Variants

Views

- [Read](#)
- [Edit](#)
- [View history](#)

More

Search

Navigation

- [Main page](#)
- [Contents](#)
- [Featured content](#)
- [Current events](#)
- [Random article](#)
- [Donate to Wikipedia](#)
- [Wikipedia store](#)

Interaction

- [Help](#)
- [About Wikipedia](#)
- [Community portal](#)

- [Recent changes](#)
- [Contact page](#)

Tools

- [What links here](#)
- [Related changes](#)
- [Upload file](#)
- [Special pages](#)
- [Permanent link](#)
- [Page information](#)
- [Wikidata item](#)
- [Cite this page](#)

Print/export

- [Create a book](#)
- [Download as PDF](#)
- [Printable version](#)

Other projects

- [Wikimedia Commons](#)

Languages

- [العربية](#)
- [Bosanski](#)
- [Català](#)
- [Čeština](#)
- [Dansk](#)
- [Deutsch](#)
- [Español](#)
- [فارسی](#)
- [Français](#)
- [한국어](#)
- [Bahasa Indonesia](#)
- [Italiano](#)
- [עברית](#)
- [Latviešu](#)
- [Lietuvių](#)
- [Nederlands](#)
- [日本語](#)
- [Norsk bokmål](#)
- [O‘zbekcha/Ўзбекча](#)
- [Polski](#)

- [Português](#)
- [Română](#)
- [Русский](#)
- [සිංහල](#)
- [Српски / srpski](#)
- [Suomi](#)
- [Svenska](#)
- [தமிழ்](#)
- [Татарча/tatarça](#)
- [ᐅᓂᓂ](#)
- [Türkçe](#)
- [Українська](#)
- [Tiếng Việt](#)
- [中文](#)

[Edit links](#)

- This page was last modified on 22 February 2016, at 03:19.
- Text is available under the [Creative Commons Attribution-ShareAlike License](#); additional terms may apply. By using this site, you agree to the [Terms of Use](#) and [Privacy Policy](#). Wikipedia® is a registered trademark of the [Wikimedia Foundation, Inc.](#), a non-profit organization.
- [Privacy policy](#)
- [About Wikipedia](#)
- [Disclaimers](#)
- [Contact Wikipedia](#)
- [Developers](#)
- [Cookie statement](#)
- [Mobile view](#)

