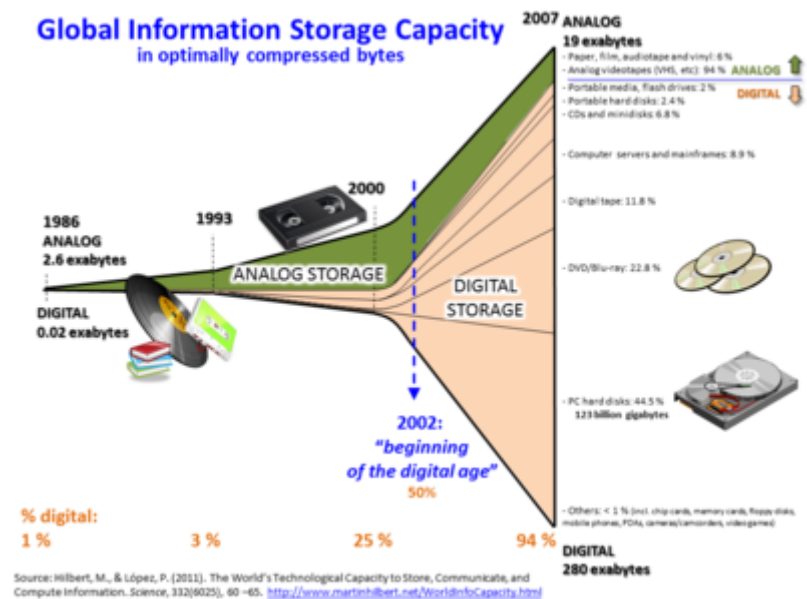


# Big data

**Big data** is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many fields (columns) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.<sup>[2]</sup> Big data analysis challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: *volume*, *variety*, and *velocity*.<sup>[3]</sup> The analysis of big data presents challenges in sampling, and thus previously allowing for only observations and sampling. Therefore, big data often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time and *value*.

Current usage of the term *big data* tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from big 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."<sup>[4]</sup> Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on".<sup>[5]</sup> Scientists, business executives, medical practitioners, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet searches, fintech, healthcare analytics, geographic information systems, urban informatics, and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics,<sup>[6]</sup> connectomics, complex physics simulations, biology, and environmental research.<sup>[7]</sup>

The size and number of available data sets have grown rapidly as data is collected by devices such as mobile devices, cheap and numerous information-sensing Internet of things devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.<sup>[8][9]</sup> The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;<sup>[10]</sup> as of 2012, every day 2.5 exabytes ( $2.5 \times 2^{60}$  bytes) of data are generated.<sup>[11]</sup> Based on an IDC report prediction, the global data volume was predicted to grow exponentially from 4.4 zettabytes to 44 zettabytes between 2013 and 2020. By 2025, IDC predicts there will be 163 zettabytes of data.<sup>[12]</sup> One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.<sup>[13]</sup>



Non-linear growth of digital global information-storage capacity and the waning of analog storage<sup>[1]</sup>

Relational database management systems and desktop statistical software packages used to visualize data often have difficulty processing and analyzing big data. The processing and analysis of big data may require "massively parallel software running on tens, hundreds, or even thousands of servers".<sup>[14]</sup> What qualifies as "big data" varies depending on the capabilities of those analyzing it and their tools. Furthermore, 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."<sup>[15]</sup>

## **Contents**

---

### **Definition**

Big data vs. business intelligence

### **Characteristics**

### **Architecture**

### **Technologies**

### **Applications**

Government

International development

Benefits

Challenges

Healthcare

Education

Media

Insurance

Internet of things (IoT)

Information technology

### **Case studies**

Government

China

India

Israel

United Kingdom

United States

Retail

Science

Sports

Technology

COVID-19

### **Research activities**

Sampling big data

### **Critique**

Critiques of the big data paradigm

Critiques of the "V" model

[Critiques of novelty](#)

[Critiques of big data execution](#)

[Critiques of big data policing and surveillance](#)

### **[In popular culture](#)**

[Books](#)

[Film](#)

### **[See also](#)**

### **[References](#)**

### **[Further reading](#)**

### **[External links](#)**

## **Definition**

---

The term *big data* has been in use since the 1990s, with some giving credit to [John Mashey](#) for popularizing the term.<sup>[16][17]</sup> 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.<sup>[18]</sup> Big data philosophy encompasses unstructured, semi-structured and structured data, however the main focus is on unstructured data.<sup>[19]</sup> Big data "size" is a constantly moving target; as of 2012 ranging from a few dozen terabytes to many [zettabytes](#) of data.<sup>[20]</sup> Big data requires a set of techniques and technologies with new forms of [integration](#) to reveal insights from [data-sets](#) that are diverse, complex, and of a massive scale.<sup>[21]</sup>

"Variety", "veracity", and various other "Vs" are added by some organizations to describe it, a revision challenged by some industry authorities.<sup>[22]</sup> The Vs of big data were often referred to as the "three Vs", "four Vs", and "five Vs". They represented the qualities of big data in volume, variety, velocity, [veracity](#), and value.<sup>[3]</sup> Variability is often included as an additional quality of big data.

A 2018 definition states "Big data is where parallel computing tools are needed to handle data", and notes, "This represents a distinct and clearly defined change in the computer science used, via parallel programming theories, and losses of some of the guarantees and capabilities made by [Codd's relational model](#)."<sup>[23]</sup>

In a comparative study of big datasets, [Kitchin](#) and McArdle found that none of the commonly considered characteristics of big data appear consistently across all of the analyzed cases.<sup>[24]</sup> For this reason, other studies identified the redefinition of power dynamics in knowledge discovery as the defining trait.<sup>[25]</sup> Instead of focusing on intrinsic characteristics of big data, this alternative perspective pushes forward a relational understanding of the object claiming that what matters is the way in which data is collected, stored, made available and analyzed.

## **Big data vs. business intelligence**

The growing maturity of the concept more starkly delineates the difference between "big data" and "[business intelligence](#)":<sup>[26]</sup>

- Business intelligence uses applied mathematics tools and [descriptive statistics](#) with data with high information density to measure things, detect trends, etc.

- Big data uses mathematical analysis, optimization, inductive statistics, and concepts from nonlinear system identification<sup>[27]</sup> to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density<sup>[28]</sup> to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.<sup>[27][29]</sup>

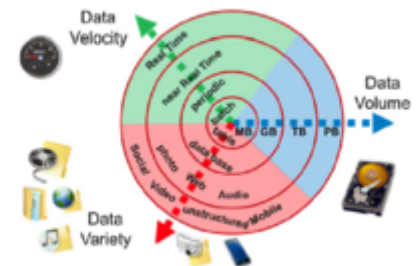
## Characteristics

---

Big data can be described by the following characteristics:

### Volume

The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not. The size of big data is usually larger than terabytes and petabytes.<sup>[30]</sup>



Shows the growth of big data's primary characteristics of volume, velocity, and variety

### Variety

The type and nature of the data. The earlier technologies like RDBMSs were capable to handle structured data efficiently and effectively. However, the change in type and nature from structured to semi-structured or unstructured challenged the existing tools and technologies. The big data technologies evolved with the prime intention to capture, store, and process the semi-structured and unstructured (variety) data generated with high speed (velocity), and huge in size (volume). Later, these tools and technologies were explored and used for handling structured data also but preferable for storage. Eventually, the processing of structured data was still kept as optional, either using big data or traditional RDBMSs. This helps in analyzing data towards effective usage of the hidden insights exposed from the data collected via social media, log files, sensors, etc. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.

### Velocity

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. Big data is often available in real-time. Compared to small data, big data is produced more continually. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing.<sup>[31]</sup>

### Veracity

The truthfulness or reliability of the data, which refers to the data quality and the data value.<sup>[32]</sup> Big data must not only be large in size, but also must be reliable in order to achieve value in the analysis of it. The data quality of captured data can vary greatly, affecting an accurate analysis.<sup>[33]</sup>

### Value

The worth in information that can be achieved by the processing and analysis of large datasets. Value also can be measured by an assessment of the other qualities of big data.<sup>[34]</sup> Value may also represent the profitability of information that is retrieved from the analysis of big data.

### Variability

The characteristic of the changing formats, structure, or sources of big data. Big data can include structured, unstructured, or combinations of structured and unstructured data. Big

data analysis may integrate raw data from multiple sources. The processing of raw data may also involve transformations of unstructured data to structured data.

Other possible characteristics of big data are:<sup>[35]</sup>

#### **Exhaustive**

Whether the entire system (i.e.,  $n=all$ ) is captured or recorded or not. Big data may or may not include all the available data from sources.

#### **Fine-grained and uniquely lexical**

Respectively, the proportion of specific data of each element per element collected and if the element and its characteristics are properly indexed or identified.

#### **Relational**

If the data collected contains common fields that would enable a conjoining, or meta-analysis, of different data sets.

#### **Extensional**

If new fields in each element of the data collected can be added or changed easily.

#### **Scalability**

If the size of the big data storage system can expand rapidly.

## **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 the largest database report.<sup>[36]</sup>

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 2.5 GB in 1991 so the definition of big data continuously evolves. 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 Risk Solutions) developed a C++-based distributed platform for data processing and querying known as the HPCC Systems platform. This system automatically partitions, distributes, stores and delivers structured, semi-structured, and unstructured data across multiple commodity servers. Users can write data processing pipelines and queries in a declarative dataflow programming language called ECL. Data analysts working in ECL are not required to define data schemas upfront and can rather focus on the particular problem at hand, reshaping data in the best possible manner as they develop the solution. In 2004, LexisNexis acquired Seisint Inc.<sup>[37]</sup> and their high-speed parallel processing platform and successfully used this platform to integrate the data systems of Choicepoint Inc. when they acquired that company in 2008.<sup>[38]</sup> In 2011, the HPCC systems platform was open-sourced under the Apache v2.0 License.

CERN and other physics experiments have collected big data sets for many decades, usually analyzed via high-throughput computing rather than the map-reduce architectures usually meant by the current "big data" movement.

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,<sup>[39]</sup> so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open-source project named "Hadoop".<sup>[40]</sup> Apache Spark was developed in 2012 in response to limitations in the MapReduce paradigm, as it adds the ability to set up many operations (not just map followed by reducing).

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".<sup>[41]</sup> 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.<sup>[42]</sup>

Studies in 2012 showed that a multiple-layer architecture was 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.<sup>[43]</sup>

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.<sup>[44][45]</sup>

## Technologies

---

A 2011 McKinsey Global Institute report characterizes the main components and ecosystem of big data as follows:<sup>[46]</sup>

- 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 OLAP data cubes or, mathematically, tensors. Array database systems have set out to provide storage and high-level query support on this data type. Additional technologies being applied to big data include efficient tensor-based computation,<sup>[47]</sup> such as multilinear subspace learning,<sup>[48]</sup> massively parallel-processing (MPP) databases, search-based applications, data mining,<sup>[49]</sup> distributed file systems, distributed cache (e.g., burst buffer and Memcached), distributed databases, cloud and HPC-based infrastructure (applications, storage and computing resources),<sup>[50]</sup> and the Internet. Although, many approaches and technologies have been developed, it still remains difficult to carry out machine learning with big data.<sup>[51]</sup>

Some 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.<sup>[52]</sup>

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".<sup>[53]</sup>

The practitioners of big data analytics processes are generally hostile to slower shared storage,<sup>[54]</sup> 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 direct-attached memory or disk is good—data on memory or disk at the other end of an FC SAN connection is not. The cost of an SAN at the scale needed for analytics applications is much higher than other storage techniques.

## 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.<sup>[5]</sup>



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.<sup>[5]</sup> Between 1990 and 2005, more than 1 billion people worldwide entered the middle class, which means more people became more literate, which in turn led 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<sup>[10]</sup> and predictions put the amount of internet traffic at 667 exabytes annually by 2014.<sup>[5]</sup> According to one estimate, one-third of the globally stored information is in the form of alphanumeric text and still image data,<sup>[55]</sup> 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 products for big data, experts promote the development of in-house custom-tailored systems if the company has sufficient technical capabilities.<sup>[56]</sup>

## Government

The use and adoption of big data within governmental processes allows efficiencies in terms of cost, productivity, and innovation,<sup>[57]</sup> 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. A common government organization that makes use of big data is the National Security Administration (NSA), which monitors the activities of the Internet constantly in search for potential patterns of suspicious or illegal activities their system may pick up.

Civil registration and vital statistics (CRVS) collects all certificates status from birth to death. CRVS is a source of big data for governments.

## 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.<sup>[58][59]</sup> 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.<sup>[60][61][62]</sup> Additionally, user-generated data offers new opportunities to give the unheard a voice.<sup>[63]</sup> 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.<sup>[60]</sup> The challenge of "big data for development"<sup>[60]</sup> is currently evolving toward the application of this data through machine learning, known as "artificial intelligence for development (AI4D)".<sup>[64]</sup>

## Benefits

A major practical application of big data for development has been "fighting poverty with data".<sup>[65]</sup> In 2015, Blumenstock and colleagues estimated predicted poverty and wealth from mobile phone metadata<sup>[66]</sup> and in 2016 Jean and colleagues combined satellite imagery and machine learning to predict poverty.<sup>[67]</sup> Using digital trace data to study the labor market and the digital economy in Latin America, Hilbert and colleagues<sup>[68][69]</sup> argue that digital trace data has several benefits such as:

- Thematic coverage: including areas that were previously difficult or impossible to measure
- Geographical coverage: our international sources provided sizable and comparable data for almost all countries, including many small countries that usually are not included in international inventories
- Level of detail: providing fine-grained data with many interrelated variables, and new aspects, like network connections
- Timeliness and timeseries: graphs can be produced within days of being collected

## Challenges

At the same time, working with digital trace data instead of traditional survey data does not eliminate the traditional challenges involved when working in the field of international quantitative analysis. Priorities change, but the basic discussions remain the same. Among the main challenges are:

- Representativeness. While traditional development statistics is mainly concerned with the representativeness of random survey samples, digital trace data is never a random sample.<sup>[70]</sup>
- Generalizability. While observational data always represents this source very well, it only represents what it represents, and nothing more. While it is tempting to generalize from specific observations of one platform to broader settings, this is often very deceptive.
- Harmonization. Digital trace data still requires international harmonization of indicators. It adds the challenge of so-called "data-fusion", the harmonization of different sources.
- Data overload. Analysts and institutions are not used to effectively deal with a large number of variables, which is efficiently done with interactive dashboards. Practitioners still lack a standard workflow that would allow researchers, users and policymakers to efficiently and effectively.<sup>[68]</sup>

## Healthcare



Big data analytics was used in healthcare 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.<sup>[71][72][73][74]</sup> 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.<sup>[75]</sup> "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.<sup>[76]</sup> While extensive information in healthcare is now electronic, it fits under the big data umbrella as most is unstructured and difficult to use.<sup>[77]</sup> The use of big data in healthcare has raised significant ethical challenges ranging from risks for individual rights, privacy and autonomy, to transparency and trust.<sup>[78]</sup>

Big data in health research is particularly promising in terms of exploratory biomedical research, as data-driven analysis can move forward more quickly than hypothesis-driven research.<sup>[79]</sup> Then, trends seen in data analysis can be tested in traditional, hypothesis-driven follow up biological research and eventually clinical research.

A related application sub-area, that heavily relies on big data, within the healthcare field is that of computer-aided diagnosis in medicine.<sup>[80]</sup> For instance, for epilepsy monitoring it is customary to create 5 to 10 GB of data daily.<sup>[81]</sup> Similarly, a single uncompressed image of breast tomosynthesis averages 450 MB of data.<sup>[82]</sup> These are just a few of the many examples where computer-aided diagnosis uses big data. For this reason, big data has been recognized as one of the seven key challenges that computer-aided diagnosis systems need to overcome in order to reach the next level of performance.<sup>[83]</sup>

## Education

A McKinsey Global Institute study found a shortage of 1.5 million highly trained data professionals and managers<sup>[46]</sup> and a number of universities<sup>[84]</sup> including University of Tennessee and UC Berkeley, have created masters programs to meet this demand. Private boot camps have also developed programs to meet that demand, including free programs like The Data Incubator or paid programs like General Assembly.<sup>[85]</sup> In the specific field of marketing, one of the problems stressed by Wedel and Kannan<sup>[86]</sup> is that marketing has several sub domains (e.g., advertising, promotions, product development, branding) that all use different types of data.

## Media

To understand how the media uses 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.<sup>[87]</sup>

- Targeting of consumers (for advertising by marketers)<sup>[88]</sup>
- 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.<sup>[89]</sup>

## Insurance

Health insurance providers are collecting data on social "determinants of health" such as food and TV consumption, marital status, clothing size, and purchasing habits, from which they make predictions on health costs, in order to spot health issues in their clients. It is controversial whether these predictions are currently being used for pricing.<sup>[90]</sup>

## 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. The IoT is also increasingly adopted as a means of gathering sensory data, and this sensory data has been used in medical,<sup>[91]</sup> manufacturing<sup>[92]</sup> and transportation<sup>[93]</sup> contexts.

Kevin Ashton, the digital innovation expert who is credited with coining the term,<sup>[94]</sup> defines the Internet of things in this quote: "If we had computers that knew everything there was to know about things—using data they gathered without any help from us—we would be able to track and count everything, and greatly reduce waste, loss, and cost. We would know when things needed replacing, repairing, or recalling, and whether they were fresh or past their best."

## 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).<sup>[95]</sup> By applying big data principles into the concepts of machine intelligence and deep computing, IT departments can predict potential issues and prevent them.<sup>[95]</sup> ITOA businesses offer platforms for systems management that bring data silos together and generate insights from the whole of the system rather than from isolated pockets of data.

## Case studies

---

### Government

#### China

- The Integrated Joint Operations Platform (IJOP, 一体化联合作战平台) is used by the government to monitor the population, particularly Uyghurs.<sup>[96]</sup> Biometrics, including DNA samples, are gathered through a program of free physicals.<sup>[97]</sup>
- By 2020, China plans to give all its citizens a personal "social credit" score based on how they behave.<sup>[98]</sup> The Social Credit System, now being piloted in a number of Chinese cities, is considered a form of mass surveillance which uses big data analysis technology.<sup>[99][100]</sup>

## India

- Big data analysis was tried out for the BJP to win the 2014 Indian General Election.<sup>[101]</sup>
- The Indian government uses numerous techniques to ascertain how the Indian electorate is responding to government action, as well as ideas for policy augmentation.

## Israel

- Personalized diabetic treatments can be created through GlucoMe's big data solution.<sup>[102]</sup>

## 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 and examine 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.<sup>[103]</sup>
- 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.<sup>[104]</sup>

## United States

- 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.<sup>[105]</sup> The initiative is composed of 84 different big data programs spread across six departments.<sup>[106]</sup>
- Big data analysis played a large role in Barack Obama's successful 2012 re-election campaign.<sup>[107]</sup>
- The United States Federal Government owns five of the ten most powerful supercomputers in the world.<sup>[108][109]</sup>
- 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.<sup>[110][111][112]</sup> This has posed security concerns regarding the anonymity of the data collected.<sup>[113]</sup>

## 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.<sup>[5]</sup>
- Windermere Real Estate uses location information 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.<sup>[114]</sup>
- FICO Card Detection System protects accounts worldwide.<sup>[115]</sup>

## 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%<sup>[116]</sup> of these streams, there are 1,000 collisions of interest per second.<sup>[117][118][119]</sup>
  - 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 ( $5 \times 10^{20}$ ) 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.<sup>[120][121]</sup> It is considered one of the most ambitious scientific projects ever undertaken.<sup>[122]</sup>
- 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.<sup>[5]</sup> 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.<sup>[5]</sup>
- 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.<sup>[123]</sup>
- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.<sup>[124][125]</sup>
- 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.<sup>[126][127]</sup>
- 23andme's DNA database contains the genetic information of over 1,000,000 people worldwide.<sup>[128]</sup> The company explores selling the "anonymous aggregated genetic data" to other researchers and pharmaceutical companies for research purposes if patients give their consent.<sup>[129][130][131][132][133]</sup> 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.<sup>[129]</sup> 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.<sup>[134]</sup>

- Computational fluid dynamics (CFD) and hydrodynamic turbulence research generate massive data sets. The Johns Hopkins Turbulence Databases (JHTDB (<http://turbulence.pha.jhu.edu>)) contains over 350 terabytes of spatiotemporal 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 JHTDB 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 cut out 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.<sup>[135]</sup> Future performance of players could be predicted as well. Thus, players' value and salary is determined by data collected throughout the season.<sup>[136]</sup>

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.<sup>[137]</sup> 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.<sup>[138]</sup>

## 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.<sup>[139]</sup>
- 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.<sup>[140]</sup>
- Facebook handles 50 billion photos from its user base.<sup>[141]</sup> As of June 2017, Facebook reached 2 billion monthly active users.<sup>[142]</sup>
- Google was handling roughly 100 billion searches per month as of August 2012.<sup>[143]</sup>

## COVID-19

During the COVID-19 pandemic, big data was raised as a way to minimise the impact of the disease. Significant applications of big data included minimising the spread of the virus, case identification and development of medical treatment.<sup>[144]</sup>

Governments used big data to track infected people to minimise spread. Early adopters included China, Taiwan, South Korea, and Israel.<sup>[145][146][147]</sup>

## Research activities

---

Encrypted search and cluster formation in big data were 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 Amir Esmailpour at the UNH Research Group investigated the key features of big data as the formation of clusters and their interconnections. They focused on the security of big data and the orientation of the term towards the presence of different types 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.<sup>[148]</sup>

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.<sup>[149]</sup>

The initiative included a National Science Foundation "Expeditions in Computing" grant of \$10 million over five years to the AMPLab<sup>[150]</sup> at the University of California, Berkeley.<sup>[151]</sup> 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<sup>[152]</sup> to fighting cancer.<sup>[153]</sup>

The White House Big Data Initiative also included a commitment by the Department of Energy to provide \$25 million in funding over five years to establish the Scalable Data Management, Analysis and Visualization (SDAV) Institute,<sup>[154]</sup> 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.<sup>[155]</sup> 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.<sup>[156]</sup>

The European Commission is funding the two-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.<sup>[157]</sup>

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 analyze large data sets.<sup>[158]</sup>

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.<sup>[159]</sup>

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.<sup>[160]</sup> Often these APIs are provided for free.<sup>[160]</sup> Tobias Preis et al. used Google Trends data to demonstrate that Internet users from countries with a higher per capita gross domestic products (GDPs) 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 behaviors and real-world economic indicators.<sup>[161][162][163]</sup> 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".<sup>[164]</sup> 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.

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.<sup>[165]</sup> Their analysis of Google search volume for 98 terms of varying financial relevance, published in *Scientific Reports*,<sup>[166]</sup> suggests that increases in search volume for financially relevant search terms tend to precede large losses in financial markets.<sup>[167][168][169][170][171][172][173]</sup>

Big data sets come with algorithmic challenges that previously did not exist. Hence, there is seen by some to be a need to fundamentally change the processing ways.<sup>[174]</sup>

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.<sup>[175]</sup> Regarding big data, such concepts of magnitude are relative. As it is stated "If the past is of any guidance, then today's big data most likely will not be considered as such in the near future."<sup>[80]</sup>

## Sampling big data

A research question that is asked about big data sets is whether it is necessary to look at the full data to draw certain conclusions about the properties of the data or if a sample is good enough. The name big data itself contains a term related to size and this is an important characteristic of big data. But sampling enables the selection of right data points from within the larger data set to estimate the characteristics of the whole population. 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 downtime 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 use 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.<sup>[176]</sup>

## 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.<sup>[177]</sup> 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."<sup>[18]</sup> 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:<sup>[178]</sup> focusing in particular on the notion that big data must always be contextualized in their social, economic, and political contexts.<sup>[179]</sup> 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*.<sup>[180]</sup>

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".<sup>[60]</sup> Fed by a large number of data on past experiences, algorithms can predict future development if the future is similar to the past.<sup>[181]</sup> If the system's 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.<sup>[181]</sup> As a response to this critique Alemany Oliver and Vayre suggest to use "abductive reasoning as a first step in the research process in order to bring context to consumers' digital traces and make new theories emerge".<sup>[182]</sup> Additionally, it has been suggested to combine big data approaches with computer simulations, such as agent-based models<sup>[60]</sup> 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.<sup>[183][184]</sup> Finally, the 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 (e.g. contingency tables) 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.<sup>[185]</sup> 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.<sup>[186][187]</sup> In the massive approaches it is the formulation of a relevant hypothesis to explain the data that is the limiting factor.<sup>[188]</sup> 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.<sup>[189]</sup> The misuse of big data in several cases by media, companies, and even the government has allowed for abolition of trust in almost every fundamental institution holding up society.<sup>[190]</sup>

Nayef Al-Rodhan argues that a new kind of social contract will be needed to protect individual liberties in the context of big data and giant corporations that own vast amounts of information, and that the use of big data should be monitored and better regulated at the national and international levels.<sup>[191]</sup> Barocas and Nissenbaum argue that one way of protecting individual users is by being informed about the types of information being collected, with whom it is shared, under what constraints and for what purposes.<sup>[192]</sup>

## Critiques of the "V" model

The "V" model of big data is concerning as it centers around computational scalability and lacks in a loss around the perceptibility and understandability of information. This led to the framework of cognitive big data, which characterizes big data applications according to:<sup>[193]</sup>

- Data completeness: understanding of the non-obvious from data
- Data correlation, causation, and predictability: causality as not essential requirement to achieve predictability



- Explainability and interpretability: humans desire to understand and accept what they understand, where algorithms do not cope with this
- Level of automated decision making: algorithms that support automated decision making and algorithmic self-learning

## Critiques of novelty

Large data sets have been analyzed by computing machines for well over a century, including the US census analytics performed by IBM's punch-card machines which computed statistics including means and variances of populations across the whole continent. In more recent decades, science experiments such as CERN have produced data on similar scales to current commercial "big data". However, science experiments have tended to analyze their data using specialized custom-built high-performance computing (super-computing) clusters and grids, rather than clouds of cheap commodity computers as in the current commercial wave, implying a difference in both culture and technology stack.

## Critiques of big data execution

Ulf-Dietrich Reips and Uwe Matzat wrote in 2014 that big data had become a "fad" in scientific research.<sup>[160]</sup> 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 handling the huge amounts of data.<sup>[194]</sup> This approach may lead to results that have a bias in one way or another.<sup>[195]</sup> 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.<sup>[196]</sup> In the provocative article "Critical Questions for Big Data",<sup>[197]</sup> 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".<sup>[197]</sup> Recent developments in BI domain, such as pro-active reporting especially target improvements in the usability of big data, through automated filtering of non-useful data and correlations.<sup>[198]</sup> Big structures are full of spurious correlations<sup>[199]</sup> either because of non-causal coincidences (law of truly large numbers), solely nature of big randomness<sup>[200]</sup> (Ramsey theory), or existence of non-included factors so the hope, of early experimenters to make large databases of numbers "speak for themselves" and revolutionize scientific method, is questioned.<sup>[201]</sup> Catherine Tucker has pointed to "hype" around big data, writing "By itself, big data is unlikely to be valuable." The article explains: "The many contexts where data is cheap relative to the cost of retaining talent to process it, suggests that processing skills are more important than data itself in creating value for a firm."<sup>[202]</sup>

Big data analysis is often shallow compared to analysis of smaller data sets.<sup>[203]</sup> In many big data projects, there is no large data analysis happening, but the challenge is the extract, transform, load part of data pre-processing.<sup>[203]</sup>

Big data is a buzzword and a "vague term",<sup>[204][205]</sup> but at the same time an "obsession"<sup>[205]</sup> 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; 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"<sup>[206]</sup> 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 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<sup>[207]</sup> with varying degrees of success.

## Critiques of big data policing and surveillance

Big data has been used in policing and surveillance by institutions like law enforcement and corporations.<sup>[208]</sup> Due to the less visible nature of data-based surveillance as compared to traditional methods of policing, objections to big data policing are less likely to arise. According to Sarah Brayne's *Big Data Surveillance: The Case of Policing*,<sup>[209]</sup> big data policing can reproduce existing societal inequalities in three ways:

- Placing suspected criminals under increased surveillance by using the justification of a mathematical and therefore unbiased algorithm
- Increasing the scope and number of people that are subject to law enforcement tracking and exacerbating existing racial overrepresentation in the criminal justice system
- Encouraging members of society to abandon interactions with institutions that would create a digital trace, thus creating obstacles to social inclusion

If these potential problems are not corrected or regulated, the effects of big data policing may continue to shape societal hierarchies. Conscientious usage of big data policing could prevent individual level biases from becoming institutional biases, Brayne also notes.

## In popular culture

---

### Books

- *Moneyball* is a non-fiction book that explores how the Oakland Athletics used statistical analysis to outperform teams with larger budgets. In 2011 a film adaptation starring Brad Pitt was released.

### Film

- In *Captain America: The Winter Soldier*, H.Y.D.R.A (disguised as S.H.I.E.L.D) develops helicarriers that use data to determine and eliminate threats over the globe.
- In *The Dark Knight*, Batman uses a sonar device that can spy on all of Gotham City. The data is gathered from the mobile phones of people within the city.

## See also

---

- Big data ethics
- Big memory
- Big Data Maturity Model
- Data curation

- Data defined storage
- Data lineage
- Data philanthropy
- Data science
- Datafication
- Document-oriented database
- In-memory processing
- List of big data companies
- Urban informatics
- Very large database
- XLDB

## References

---

1. Hilbert, Martin; López, Priscila (2011). "The World's Technological Capacity to Store, Communicate, and Compute Information" (<http://www.martinhilbert.net/WorldInfoCapacity.html>). *Science*. **332** (6025): 60–65. Bibcode:2011Sci...332...60H (<https://ui.adsabs.harvard.edu/abs/2011Sci...332...60H>). doi:10.1126/science.1200970 (<https://doi.org/10.1126%2Fscience.1200970>). PMID 21310967 (<https://pubmed.ncbi.nlm.nih.gov/21310967>). S2CID 206531385 (<https://api.semanticscholar.org/CorpusID:206531385>). Retrieved 13 April 2016.
2. Breur, Tom (July 2016). "Statistical Power Analysis and the contemporary "crisis" in social sciences" (<https://doi.org/10.1057%2Fs41270-016-0001-3>). *Journal of Marketing Analytics*. London, England: Palgrave Macmillan. **4** (2–3): 61–65. doi:10.1057/s41270-016-0001-3 (<https://doi.org/10.1057%2Fs41270-016-0001-3>). ISSN 2050-3318 (<https://www.worldcat.org/issn/2050-3318>).
3. "The 5 V's of big data" (<https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/>). *Watson Health Perspectives*. 17 September 2016. Retrieved 20 January 2021.
4. boyd, dana; Crawford, Kate (21 September 2011). "Six Provocations for Big Data" (<http://osf.io/nrjhn/>). *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>). S2CID 148610111 (<https://api.semanticscholar.org/CorpusID:148610111>).
5. "Data, data everywhere" (<http://www.economist.com/node/15557443>). *The Economist*. 25 February 2010. Retrieved 9 December 2012.
6. "Community cleverness required" (<https://doi.org/10.1038%2F455001a>). *Nature*. **455** (7209): 1. September 2008. Bibcode:2008Natur.455....1. (<https://ui.adsabs.harvard.edu/abs/2008Natur.455....1>). doi:10.1038/455001a (<https://doi.org/10.1038%2F455001a>). PMID 18769385 (<https://pubmed.ncbi.nlm.nih.gov/18769385>).
7. Reichman OJ, Jones MB, Schildhauer MP (February 2011). "Challenges and opportunities of open data in ecology" (<https://escholarship.org/uc/item/7627s45z>). *Science*. **331** (6018): 703–5. Bibcode:2011Sci...331..703R (<https://ui.adsabs.harvard.edu/abs/2011Sci...331..703R>). doi:10.1126/science.1197962 (<https://doi.org/10.1126%2Fscience.1197962>). PMID 21311007 (<https://pubmed.ncbi.nlm.nih.gov/21311007>). S2CID 22686503 (<https://api.semanticscholar.org/CorpusID:22686503>).
8. 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*.
9. 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.
10. Hilbert M, López P (April 2011). "The world's technological capacity to store, communicate, and compute information" (<http://www.uvm.edu/pdodds/files/papers/others/2011/hilbert2011a.pdf>) (PDF). *Science*. **332** (6025): 60–5. Bibcode:2011Sci...332...60H (<https://ui.adsabs.harvard.edu/abs/2011Sci...332...60H>). doi:10.1126/science.1200970 (<https://doi.org/10.1126%2Fscience.1200970>). PMID 21310967 (<https://pubmed.ncbi.nlm.nih.gov/21310967>). S2CID 206531385 (<https://api.semanticscholar.org/CorpusID:206531385>).

11. "IBM What is big data? – Bringing big data to the enterprise" (<http://www.ibm.com/big-data/us/en/>). *ibm.com*. Retrieved 26 August 2013.
12. Reinsel, David; Gantz, John; Rydning, John (13 April 2017). "Data Age 2025: The Evolution of Data to Life-Critical" (<https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>) (PDF). *seagate.com*. Framingham, MA, US: International Data Corporation. Retrieved 2 November 2017.
13. 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](http://www.fsn.co.uk/channel_bi_bpm_cpm/mastering_big_data_cfo_strategies_to_transform_insight_into_opportunity)) Archived ([https://web.archive.org/web/20130804062518/http://www.fsn.co.uk/channel\\_bi\\_bpm\\_cpm/mastering\\_big\\_data\\_cfo\\_strategies\\_to\\_transform\\_insight\\_into\\_opportunity](https://web.archive.org/web/20130804062518/http://www.fsn.co.uk/channel_bi_bpm_cpm/mastering_big_data_cfo_strategies_to_transform_insight_into_opportunity)) 4 August 2013 at the Wayback Machine, December 2012
14. Jacobs, A. (6 July 2009). "The Pathologies of Big Data" (<http://queue.acm.org/detail.cfm?id=1563874>). *ACMQueue*.
15. Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data" (<https://academics.uccs.edu/~ooluwada/courses/datamining/ExtraReading/BigData>). *Release 2.0*. Sebastopol CA: O'Reilly Media (11).
16. John R. Mashey (25 April 1998). "Big Data ... and the Next Wave of InfraStress" ([http://static.usenix.org/event/usenix99/invited\\_talks/mashey.pdf](http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf)) (PDF). *Slides from invited talk*. Usenix. Retrieved 28 September 2016.
17. 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/>). *The New York Times*. Retrieved 28 September 2016.
18. 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](http://www.ijis.net/ijis7_1/ijis7_1_editorial.html)). *International Journal of Internet Science*. 7: 1–5.
19. Dedić, N.; Stanier, C. (2017). "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery" (<http://eprints.staffs.ac.uk/3551/1/Towards%20Differentiating%20Business%20Intelligence%20Big%20Data%20Data%20Analytics%20and%20Knowledge%20Discovery.docx>). *Innovations in Enterprise Information Systems Management and Engineering*. Lecture Notes in Business Information Processing. Vol. 285. Berlin ; Heidelberg: Springer International Publishing. pp. 114–122. doi:10.1007/978-3-319-58801-8\_10 ([https://doi.org/10.1007%2F978-3-319-58801-8\\_10](https://doi.org/10.1007%2F978-3-319-58801-8_10)). ISBN 978-3-319-58800-1. ISSN 1865-1356 (<https://www.worldcat.org/issn/1865-1356>). OCLC 909580101 (<https://www.worldcat.org/oclc/909580101>).
20. Everts, Sarah (2016). "Information Overload" (<https://www.sciencehistory.org/distillations/magazine/information-overload>). *Distillations*. Vol. 2, no. 2. pp. 26–33. Retrieved 22 March 2018.
21. 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>).
22. 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.
23. Fox, Charles (25 March 2018). *Data Science for Transport* (<https://www.springer.com/us/book/9783319729527>). Springer Textbooks in Earth Sciences, Geography and Environment. Springer. ISBN 9783319729527.

24. Kitchin, Rob; McArdle, Gavin (2016). "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets". *Big Data & Society*. **3**: 1–10. doi:10.1177/2053951716631130 (<https://doi.org/10.1177%2F2053951716631130>). S2CID 55539845 (<https://api.semanticscholar.org/CorpusID:55539845>).
25. Balazka, Dominik; Rodighiero, Dario (2020). "Big Data and the Little Big Bang: An Epistemological (R)evolution" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931920>). *Frontiers in Big Data*. **3**: 31. doi:10.3389/fdata.2020.00031 (<https://doi.org/10.3389%2Fdata.2020.00031>). hdl:1721.1/128865 (<https://hdl.handle.net/1721.1%2F128865>). PMC 7931920 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931920>). PMID 33693404 (<https://pubmed.ncbi.nlm.nih.gov/33693404>).
26. "avec focalisation sur Big Data & Analytique" (<https://web.archive.org/web/20210225014647/https://www.bigdataparis.com/presentation/mercredi/PDelort.pdf?PHPSESSID=tv7k70pcr3egpi2r6fi3qbjtj6#page=4>) (PDF). *Bigdataparis.com*. Archived from the original (<http://www.bigdataparis.com/presentation/mercredi/PDelort.pdf?PHPSESSID=tv7k70pcr3egpi2r6fi3qbjtj6#page=4>) (PDF) on 25 February 2021. Retrieved 8 October 2017.
27. Billings S.A. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains". Wiley, 2013
28. "le Blog ANDSI » DSI Big Data" (<http://www.andsi.fr/tag/dsi-big-data/>). *Andsi.fr*. Retrieved 8 October 2017.
29. 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/entrepreneurlendances-innovation/221169222/big-data-low-density-data-faible-densite-information-com>). *Lesechos.fr*. Retrieved 8 October 2017.
30. Sagioglu, Seref (2013). "Big data: A review". *2013 International Conference on Collaboration Technologies and Systems (CTS)*: 42–47. doi:10.1109/CTS.2013.6567202 (<https://doi.org/10.1109%2FCTS.2013.6567202>). ISBN 978-1-4673-6404-1. S2CID 5724608 (<https://api.semanticscholar.org/CorpusID:5724608>).
31. Kitchin, Rob; McArdle, Gavin (17 February 2016). "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets" (<https://doi.org/10.1177%2F2053951716631130>). *Big Data & Society*. **3** (1): 205395171663113. doi:10.1177/2053951716631130 (<https://doi.org/10.1177%2F2053951716631130>).
32. Onay, Ceylan; Öztürk, Elif (2018). "A review of credit scoring research in the age of Big Data". *Journal of Financial Regulation and Compliance*. **26** (3): 382–405. doi:10.1108/JFRC-06-2017-0054 (<https://doi.org/10.1108%2FJFRC-06-2017-0054>). S2CID 158895306 (<https://api.semanticscholar.org/CorpusID:158895306>).
33. Big Data's Fourth V (<https://web.archive.org/web/20180731105912/https://spotlessdata.com/blog/big-datas-fourth-v>)
34. "Measuring the Business Value of Big Data | IBM Big Data & Analytics Hub" (<https://www.ibmbigdatahub.com/blog/measuring-business-value-big-data>). *www.ibmbigdatahub.com*. Retrieved 20 January 2021.
35. Kitchin, Rob; McArdle, Gavin (5 January 2016). "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets" (<https://doi.org/10.1177%2F2053951716631130>). *Big Data & Society*. **3** (1): 205395171663113. doi:10.1177/2053951716631130 (<https://doi.org/10.1177%2F2053951716631130>). ISSN 2053-9517 (<https://www.worldcat.org/issn/2053-9517>).
36. "Survey: Biggest Databases Approach 30 Terabytes" (<http://www.eweek.com/database/survey-biggest-databases-approach-30-terabytes>). *Eweek.com*. 8 November 2003. Retrieved 8 October 2017.
37. "LexisNexis To Buy Seisint For \$775 Million" (<https://www.washingtonpost.com/wp-dyn/articles/A50577-2004Jul14.html>). *The Washington Post*. Retrieved 15 July 2004.

38. The Washington Post (<https://www.washingtonpost.com/wp-dyn/content/article/2008/02/21/AR2008022100809.html>)
39. Bertolucci, Jeff "Hadoop: From Experiment To Leading Big Data Platform" (<http://www.informationweek.com/software/hadoop-from-experiment-to-leading-big-data-platform/d/d-id/1110491?>), "Information Week", 2013. Retrieved on 14 November 2013.
40. 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.
41. "Big Data Solution Offering" ([http://mike2.openmethodology.org/wiki/Big\\_Data\\_Solution\\_Offering](http://mike2.openmethodology.org/wiki/Big_Data_Solution_Offering)). MIKE2.0. Retrieved 8 December 2013.
42. "Big Data Definition" ([http://mike2.openmethodology.org/wiki/Big\\_Data\\_Definition](http://mike2.openmethodology.org/wiki/Big_Data_Definition)). MIKE2.0. Retrieved 9 March 2013.
43. Boja, C; Pocovnicu, A; Bătăgan, L. (2012). "Distributed Parallel Architecture for Big Data". *Informatica Economica*. **16** (2): 116–127.
44. "Solving Key Business Challenges With a Big Data Lake" ([http://www.hcltech.com/sites/default/files/solving\\_key\\_businesschallenges\\_with\\_big\\_data\\_lake\\_0.pdf](http://www.hcltech.com/sites/default/files/solving_key_businesschallenges_with_big_data_lake_0.pdf)) (PDF). *Hcltech.com*. August 2014. Retrieved 8 October 2017.
45. "Method for testing the fault tolerance of MapReduce frameworks" (<https://secplab.ppgia.pucpr.br/files/papers/2015-0.pdf>) (PDF). *Computer Networks*. 2015.
46. 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" ([https://www.mckinsey.com/~media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi\\_big\\_data\\_full\\_report.pdf](https://www.mckinsey.com/~media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi_big_data_full_report.pdf)) (PDF). McKinsey Global Institute. Retrieved 22 May 2021.
47. "Future Directions in Tensor-Based Computation and Modeling" (<http://www.cs.cornell.edu/cv/tenwork/finalreport.pdf>) (PDF). May 2009.
48. 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](http://www.dsp.utoronto.ca/~haiping/Publication/SurveyMSL_PR2011.pdf)) (PDF). *Pattern Recognition*. **44** (7): 1540–1551. Bibcode:2011PatRe..44.1540L (<https://ui.adsabs.harvard.edu/abs/2011PatRe..44.1540L>). doi:10.1016/j.patcog.2011.01.004 (<https://doi.org/10.1016%2Fj.patcog.2011.01.004>).
49. Pllana, Sabri; Janciak, Ivan; Brezany, Peter; Wöhrer, Alexander (2016). "A Survey of the State of the Art in Data Mining and Integration Query Languages". *2011 14th International Conference on Network-Based Information Systems. 2011 International Conference on Network-Based Information Systems (NBIS 2011)*. IEEE Computer Society. pp. 341–348. arXiv:1603.01113 (<https://arxiv.org/abs/1603.01113>). Bibcode:2016arXiv160301113P (<https://ui.adsabs.harvard.edu/abs/2016arXiv160301113P>). doi:10.1109/NBiS.2011.58 (<https://doi.org/10.1109%2FNBiS.2011.58>). ISBN 978-1-4577-0789-6. S2CID 9285984 (<https://api.semanticscholar.org/CorpusID:9285984>).
50. Wang, Yandong; Goldstone, Robin; Yu, Weikuan; Wang, Teng (October 2014). "Characterization and Optimization of Memory-Resident MapReduce on HPC Systems". *2014 IEEE 28th International Parallel and Distributed Processing Symposium*. IEEE. pp. 799–808. doi:10.1109/IPDPS.2014.87 (<https://doi.org/10.1109%2FIPDPS.2014.87>). ISBN 978-1-4799-3800-1. S2CID 11157612 (<https://api.semanticscholar.org/CorpusID:11157612>).

51. L'Heureux, A.; Grolinger, K.; Elyamany, H. F.; Capretz, M. A. M. (2017). "Machine Learning With Big Data: Challenges and Approaches" (<https://doi.org/10.1109%2FACCESS.2017.2696365>). *IEEE Access*. **5**: 7776–7797. doi:10.1109/ACCESS.2017.2696365 (<https://doi.org/10.1109%2FACCESS.2017.2696365>). ISSN 2169-3536 (<https://www.worldcat.org/issn/2169-3536>).
52. 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/>).
53. "Resources on how Topological Data Analysis is used to analyze big data" (<http://www.ayasdi.com/resources/>). Ayasdi.
54. CNET News (1 April 2011). "Storage area networks need not apply" ([http://news.cnet.com/8301-21546\\_3-20049693-10253464.html](http://news.cnet.com/8301-21546_3-20049693-10253464.html)).
55. Hilbert, Martin (2014). "What is the Content of the World's Technologically Mediated Information and Communication Capacity: How Much Text, Image, Audio, and Video?" (<http://escholarship.org/uc/item/87w5f6wb>). *The Information Society*. **30** (2): 127–143. doi:10.1080/01972243.2013.873748 (<https://doi.org/10.1080%2F01972243.2013.873748>). S2CID 45759014 (<https://api.semanticscholar.org/CorpusID:45759014>).
56. 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 14 July 2014. "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."
57. "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.
58. "White Paper: Big Data for Development: Opportunities & Challenges (2012) – United Nations Global Pulse" (<http://www.unglobalpulse.org/projects/BigDataforDevelopment>). *Unglobalpulse.org*. Retrieved 13 April 2016.
59. "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.
60. Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, 34(1), 135–174. <https://doi.org/10.1111/dpr.12142> free access: <https://www.martinhilbert.net/big-data-for-development/>
61. "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. 4 December 2012. Retrieved 30 May 2012.
62. "Daniele Medri: Big Data & Business: An on-going revolution" (<https://web.archive.org/web/20150617211645/http://www.statisticsviews.com/details/feature/5393251/Big-Data--Business-An-on-going-revolution.html>). Statistics Views. 21 October 2013. Archived from the original (<http://www.statisticsviews.com/details/feature/5393251/Big-Data--Business-An-on-going-revolution.html>) on 17 June 2015. Retrieved 21 June 2015.
63. 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*.

64. Mann, S., & Hilbert, M. (2020). AI4D: Artificial Intelligence for Development. *International Journal of Communication*, 14(0), 21. <https://www.martinhilbert.net/ai4d-artificial-intelligence-for-development/>
65. Blumenstock, J. E. (2016). Fighting poverty with data. *Science*, 353(6301), 753–754. <https://doi.org/10.1126/science.aah5217>
66. Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073–1076. <https://doi.org/10.1126/science.aac4420>
67. Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790–794. <https://doi.org/10.1126/science.aaf7894>
68. Hilbert, M., & Lu, K. (2020). The online job market trace in Latin America and the Caribbean (UN ECLAC LC/TS.2020/83; p. 79). United Nations Economic Commission for Latin America and the Caribbean. <https://www.cepal.org/en/publications/45892-online-job-market-trace-latin-america-and-caribbean>
69. UN ECLAC, (United Nations Economic Commission for Latin America and the Caribbean). (2020). Tracking the digital footprint in Latin America and the Caribbean: Lessons learned from using big data to assess the digital economy (Productive Development, Gender Affairs LC/TS.2020/12; Documentos de Proyecto). United Nations ECLAC. <https://repositorio.cepal.org/handle/11362/45484>
70. Banerjee, Amitav; Chaudhury, Suprakash (2010). "Statistics without tears: Populations and samples" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3105563>). *Industrial Psychiatry Journal*. **19** (1): 60–65. doi:10.4103/0972-6748.77642 (<https://doi.org/10.4103%2F0972-6748.77642>). ISSN 0972-6748 (<https://www.worldcat.org/issn/0972-6748>). PMC 3105563 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3105563>). PMID 21694795 (<https://pubmed.ncbi.nlm.nih.gov/21694795>).
71. Huser V, Cimino JJ (July 2016). "Impending Challenges for the Use of Big Data" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4860172>). *International Journal of Radiation Oncology, Biology, Physics*. **95** (3): 890–894. doi:10.1016/j.ijrobp.2015.10.060 (<https://doi.org/10.1016%2Fj.ijrobp.2015.10.060>). PMC 4860172 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4860172>). PMID 26797535 (<https://pubmed.ncbi.nlm.nih.gov/26797535>).
72. Sejdic, Ervin; Falk, Tiago H. (4 July 2018). *Signal Processing and Machine Learning for Biomedical Big Data*. Sejdíć, Ervin, Falk, Tiago H. [Place of publication not identified]. ISBN 9781351061216. OCLC 1044733829 (<https://www.worldcat.org/oclc/1044733829>).
73. Raghupathi W, Raghupathi V (December 2014). "Big data analytics in healthcare: promise and potential" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817>). *Health Information Science and Systems*. **2** (1): 3. doi:10.1186/2047-2501-2-3 (<https://doi.org/10.1186%2F2047-2501-2-3>). PMC 4341817 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817>). PMID 25825667 (<https://pubmed.ncbi.nlm.nih.gov/25825667>).
74. Viceconti M, Hunter P, Hose R (July 2015). "Big data, big knowledge: big data for personalized healthcare" (<http://eprints.whiterose.ac.uk/89104/1/pap%20JBHI%20BigData%20in%20VPH%20revision%20v2.pdf>) (PDF). *IEEE Journal of Biomedical and Health Informatics*. **19** (4): 1209–15. doi:10.1109/JBHI.2015.2406883 (<https://doi.org/10.1109%2FJBHI.2015.2406883>). PMID 26218867 (<https://pubmed.ncbi.nlm.nih.gov/26218867>). S2CID 14710821 (<https://api.semanticscholar.org/CorpusID:14710821>).
75. O'Donoghue, John; Herbert, John (1 October 2012). "Data Management Within mHealth Environments: Patient Sensors, Mobile Devices, and Databases". *Journal of Data and Information Quality*. **4** (1): 5:1–5:20. doi:10.1145/2378016.2378021 (<https://doi.org/10.1145%2F2378016.2378021>). S2CID 2318649 (<https://api.semanticscholar.org/CorpusID:2318649>).



76. Mirkes EM, Coats TJ, Levesley J, Gorban AN (August 2016). "Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes". *Computers in Biology and Medicine*. **75**: 203–16. arXiv:1604.00627 (<https://arxiv.org/abs/1604.00627>). Bibcode:2016arXiv160400627M (<https://ui.adsabs.harvard.edu/abs/2016arXiv160400627M>). doi:10.1016/j.combiomed.2016.06.004 (<https://doi.org/10.1016%2Fj.combiomed.2016.06.004>). PMID 27318570 (<https://pubmed.ncbi.nlm.nih.gov/27318570>). S2CID 5874067 (<https://api.semanticscholar.org/CorpusID:5874067>).
77. Murdoch TB, Detsky AS (April 2013). "The inevitable application of big data to health care". *JAMA*. **309** (13): 1351–2. doi:10.1001/jama.2013.393 (<https://doi.org/10.1001%2Fjama.2013.393>). PMID 23549579 (<https://pubmed.ncbi.nlm.nih.gov/23549579>).
78. Vayena E, Salathé M, Madoff LC, Brownstein JS (February 2015). "Ethical challenges of big data in public health" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321985>). *PLOS Computational Biology*. **11** (2): e1003904. Bibcode:2015PLSCB..11E3904V (<https://ui.adsabs.harvard.edu/abs/2015PLSCB..11E3904V>). doi:10.1371/journal.pcbi.1003904 (<https://doi.org/10.1371%2Fjournal.pcbi.1003904>). PMC 4321985 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321985>). PMID 25664461 (<https://pubmed.ncbi.nlm.nih.gov/25664461>).
79. Copeland, CS (July–August 2017). "Data Driving Discovery" ([http://claudiacopeland.com/uploads/3/5/5/6/35560346/hjno\\_data\\_driving\\_discovery\\_2pv.pdf](http://claudiacopeland.com/uploads/3/5/5/6/35560346/hjno_data_driving_discovery_2pv.pdf)) (PDF). *Healthcare Journal of New Orleans*: 22–27.
80. Yanase J, Triantaphyllou E (2019). "A Systematic Survey of Computer-Aided Diagnosis in Medicine: Past and Present Developments". *Expert Systems with Applications*. **138**: 112821. doi:10.1016/j.eswa.2019.112821 (<https://doi.org/10.1016%2Fj.eswa.2019.112821>). S2CID 199019309 (<https://api.semanticscholar.org/CorpusID:199019309>).
81. Dong X, Bahroos N, Sadhu E, Jackson T, Chukhman M, Johnson R, Boyd A, Hynes D (2013). "Leverage Hadoop framework for large scale clinical informatics applications". *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*. **2013**: 53. PMID 24303235 (<https://pubmed.ncbi.nlm.nih.gov/24303235>).
82. Clunie D (2013). "Breast tomosynthesis challenges digital imaging infrastructure" (<http://www.auntminnie.com/index.aspx?sec=prtf&sub=def&pag=dis&itemId=102872&printpage=true&fsec=ser&fsub=def>).
83. Yanase J, Triantaphyllou E (2019). "The Seven Key Challenges for the Future of Computer-Aided Diagnosis in Medicine". *International Journal of Medical Informatics*. **129**: 413–422. doi:10.1016/j.ijmedinf.2019.06.017 (<https://doi.org/10.1016%2Fj.ijmedinf.2019.06.017>). PMID 31445285 (<https://pubmed.ncbi.nlm.nih.gov/31445285>). S2CID 198287435 (<https://api.semanticscholar.org/CorpusID:198287435>).
84. "Degrees in Big Data: Fad or Fast Track to Career Success" (<https://www.forbes.com/sites/jmaureenhenderson/2013/07/30/degrees-in-big-data-fad-or-fast-track-to-career-success/>). *Forbes*. Retrieved 21 February 2016.
85. "NY gets new boot camp 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 21 February 2016.
86. Wedel, Michel; Kannan, PK (2016). "Marketing Analytics for Data-Rich Environments". *Journal of Marketing*. **80** (6): 97–121. doi:10.1509/jm.15.0413 (<https://doi.org/10.1509%2Fjm.15.0413>). S2CID 168410284 (<https://api.semanticscholar.org/CorpusID:168410284>).
87. 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.

88. "Why Digital Advertising Agencies Suck at Acquisition and are in Dire Need of an AI Assisted Upgrade" (<https://web.archive.org/web/20190212174722/https://ishti.org/2018/04/15/why-digital-advertising-agencies-suck-at-acquisition-and-are-in-dire-need-of-an-ai-assisted-upgrade/>). *Ishti.org*. 15 April 2018. Archived from the original (<https://ishti.org/2018/04/15/why-digital-advertising-agencies-suck-at-acquisition-and-are-in-dire-need-of-an-ai-assisted-upgrade/>) on 12 February 2019. Retrieved 15 April 2018.
89. "Big data and analytics: C4 and Genius Digital" (<https://www.abc.org/tech-advances/big-data-and-analytics-c4-and-genius-digital/1076.article>). *abc.org*. Retrieved 8 October 2017.
90. Marshall Allen (17 July 2018). "Health Insurers Are Vacuuming Up Details About You – And It Could Raise Your Rates" (<https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates>). *www.propublica.org*. Retrieved 21 July 2018.
91. "QuiO Named Innovation Champion of the Accenture HealthTech Innovation Challenge" (<http://www.businesswire.com/news/home/20170109006500/en/QuiO-Named-Innovation-Champion-Accenture-HealthTech-Innovation>). *Businesswire.com*. 10 January 2017. Retrieved 8 October 2017.
92. "A Software Platform for Operational Technology Innovation" ([https://www.predix.com/sites/default/files/IDC\\_OT\\_Final\\_whitepaper\\_249120.pdf](https://www.predix.com/sites/default/files/IDC_OT_Final_whitepaper_249120.pdf)) (PDF). *Predix.com*. Retrieved 8 October 2017.
93. Z. Jenipher Wang (March 2017). "Big Data Driven Smart Transportation: the Underlying Story of IoT Transformed Mobility" (<http://www.wiomax.com/big-data-driven-smart-transportation-the-underlying-big-story-of-smart-iot-transformed-mobility/>).
94. "That Internet Of Things Thing" (<http://www.rfidjournal.com/articles/view?4986>).
95. Solnik, Ray. "The Time Has Come: Analytics Delivers for IT Operations" (<http://www.datacenterjournal.com/time-analytics-delivers-operations/>). *Data Center Journal*. Retrieved 21 June 2016.
96. Josh Rogin (2 August 2018). "Ethnic cleansing makes a comeback – in China" (<https://web.archive.org/web/20190331161843/https://www.washingtonpost.com/opinions/global-opinion/s/ethnic-cleansing-makes-a-comeback--in-china/2018/08/02/>). No. Washington Post. Archived from the original (<https://www.washingtonpost.com/opinions/global-opinion/s/ethnic-cleansing-makes-a-comeback--in-china/2018/08/02/>) on 31 March 2019. Retrieved 4 August 2018. "Add to that the unprecedented security and surveillance state in Xinjiang, which includes all-encompassing monitoring based on identity cards, checkpoints, facial recognition and the collection of DNA from millions of individuals. The authorities feed all this data into an artificial-intelligence machine that rates people's loyalty to the Communist Party in order to control every aspect of their lives."
97. "China: Big Data Fuels Crackdown in Minority Region: Predictive Policing Program Flags Individuals for Investigations, Detentions" (<https://www.hrw.org/news/2018/02/26/china-big-data-fuels-crackdown-minority-region>). *hrw.org*. Human Rights Watch. 26 February 2018. Retrieved 4 August 2018.
98. "Discipline and Punish: The Birth of China's Social-Credit System" (<https://www.thenation.com/article/china-social-credit-system/>). *The Nation*. 23 January 2019.
99. "China's behavior monitoring system bars some from travel, purchasing property" (<https://www.cbsnews.com/news/china-social-credit-system-surveillance-cameras/>). *CBS News*. 24 April 2018.
100. "The complicated truth about China's social credit system" (<https://www.wired.co.uk/article/china-social-credit-system-explained>). *WIRED*. 21 January 2019.
101. "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 22 November 2014.

102. "Israeli startup uses big data, minimal hardware to treat diabetes" (<https://www.timesofisrael.com/israeli-startup-uses-big-data-minimal-hardware-to-treat-diabetes/>). *The Times of Israel*. Retrieved 28 February 2018.
103. Singh, Gurparkash, Duane Schulthess, Nigel Hughes, Bart Vannieuwenhuyse, and Dipak Kalra (2018). "Real world big data for clinical research and drug development". *Drug Discovery Today*. **23** (3): 652–660. doi:10.1016/j.drudis.2017.12.002 (<https://doi.org/10.1016%2Fj.drudis.2017.12.002>). PMID 29294362 (<https://pubmed.ncbi.nlm.nih.gov/29294362>).
104. "Recent advances delivered by Mobile Cloud Computing and Internet of Things for Big Data applications: a survey" (<https://www.researchgate.net/publication/297762848>). International Journal of Network Management. 11 March 2016. Retrieved 14 September 2016.
105. Kalil, Tom (29 March 2012). "Big Data is a Big Deal" (<https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal>). *whitehouse.gov*. Retrieved 26 September 2012 – via National Archives.
106. Executive Office of the President (March 2012). "Big Data Across the Federal Government" ([https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big\\_data\\_fact\\_sheet\\_final\\_1.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf)) (PDF). *Office of Science and Technology Policy*. Archived ([https://web.archive.org/web/20170121233257/https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big\\_data\\_fact\\_sheet\\_final\\_1.pdf](https://web.archive.org/web/20170121233257/https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf)) (PDF) from the original on 21 January 2017. Retrieved 26 September 2012 – via National Archives.
107. Lampitt, Andrew (14 February 2013). "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.
108. "November 2018 | TOP500 Supercomputer Sites" (<https://www.top500.org/lists/2018/11/>).
109. Hoover, J. Nicholas. "Government's 10 Most Powerful Supercomputers" (<http://www.informationweek.com/government/enterprise-applications/image-gallery-governments-10-most-powerful/224700271>). *Information Week*. UBM. Retrieved 26 September 2012.
110. 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](https://www.wired.com/threatlevel/2012/03/ff_nsadatacenter/all/1)). *Wired*. Retrieved 18 March 2013.
111. "Groundbreaking Ceremony Held for \$1.2 Billion Utah Data Center" ([https://web.archive.org/web/20130905055004/http://www.nsa.gov/public\\_info/press\\_room/2011/utah\\_groundbreaking\\_ceremony.shtml](https://web.archive.org/web/20130905055004/http://www.nsa.gov/public_info/press_room/2011/utah_groundbreaking_ceremony.shtml)). National Security Agency Central Security Service. Archived from the original ([http://www.nsa.gov/public\\_info/press\\_room/2011/utah\\_groundbreaking\\_ceremony.shtml](http://www.nsa.gov/public_info/press_room/2011/utah_groundbreaking_ceremony.shtml)) on 5 September 2013. Retrieved 18 March 2013.
112. Hill, Kashmir. "Blueprints of NSA's Ridiculously Expensive Data Center in Utah Suggest It Holds Less Info Than Thought" (<https://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 31 October 2013.
113. Smith, Gerry; Hallman, Ben (12 June 2013). "NSA Spying Controversy Highlights Embrace of Big Data" ([https://www.huffingtonpost.com/2013/06/12/nsa-big-data\\_n\\_3423482.html](https://www.huffingtonpost.com/2013/06/12/nsa-big-data_n_3423482.html)). *Huffington Post*. Retrieved 7 May 2018.
114. Wingfield, Nick (12 March 2013). "Predicting Commutes More Accurately for Would-Be Home Buyers" (<http://bits.blogs.nytimes.com/2013/03/12/predicting-commutes-more-accurately-for-would-be-home-buyers/>). *The New York Times*. Retrieved 21 July 2013.
115. "FICO® Falcon® Fraud Manager" (<http://www.fico.com/en/Products/DMAApps/Pages/FICO-Falcon-Fraud-Manager.aspx>). Fico.com. Retrieved 21 July 2013.
116. Alexandru, Dan. "Prof" (<https://cds.cern.ch/record/1504817/files/CERN-THESIS-2013-004.pdf>) (PDF). *cds.cern.ch*. CERN. Retrieved 24 March 2015.

117. "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.
118. "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.
119. Brumfiel, Geoff (19 January 2011). "High-energy physics: Down the petabyte highway" (<http://www.nature.com/news/2011/110119/full/469282a.html>). *Nature*. Vol. 469. pp. 282–83. Bibcode:2011Natur.469..282B (<https://ui.adsabs.harvard.edu/abs/2011Natur.469..282B>). doi:10.1038/469282a (<https://doi.org/10.1038%2F469282a>).
120. "IBM Research – Zurich" (<http://www.zurich.ibm.com/pdf/astron/CeBIT+2013+Background+DOME.pdf>) (PDF). *Zurich.ibm.com*. Retrieved 8 October 2017.
121. "Future telescope array drives development of Exabyte processing" (<https://arstechnica.com/science/2012/04/future-telescope-array-drives-development-of-exabyte-processing/>). *Ars Technica*. 2 April 2012. Retrieved 15 April 2015.
122. "Australia's bid for the Square Kilometre Array – an insider's perspective" (<https://theconversation.com/australias-bid-for-the-square-kilometre-array-an-insiders-perspective-4891>). *The Conversation*. 1 February 2012. Retrieved 27 September 2016.
123. "Delort P., OECD ICCP Technology Foresight Forum, 2012" ([http://www.oecd.org/sti/economy/Session\\_3\\_Delort.pdf#page=6](http://www.oecd.org/sti/economy/Session_3_Delort.pdf#page=6)) (PDF). *Oecd.org*. Retrieved 8 October 2017.
124. "NASA – NASA Goddard Introduces the NASA Center for Climate Simulation" (<http://www.nasa.gov/centers/goddard/news/releases/2010/10-051.html>). *Nasa.gov*. Retrieved 13 April 2016.
125. Webster, Phil. "Supercomputing the Climate: NASA's Big Data Mission" ([https://web.archive.org/web/20130104220150/http://www.csc.com/cscworld/publications/81769/81773-supercomputing\\_the\\_climate\\_nasa\\_s\\_big\\_data\\_mission](https://web.archive.org/web/20130104220150/http://www.csc.com/cscworld/publications/81769/81773-supercomputing_the_climate_nasa_s_big_data_mission)). *CSC World*. Computer Sciences Corporation. Archived from the original ([http://www.csc.com/cscworld/publications/81769/81773-supercomputing\\_the\\_climate\\_nasa\\_s\\_big\\_data\\_mission](http://www.csc.com/cscworld/publications/81769/81773-supercomputing_the_climate_nasa_s_big_data_mission)) on 4 January 2013. Retrieved 18 January 2013.
126. "These six great neuroscience ideas could make the leap from lab to market" (<https://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.
127. "DNASTack tackles massive, complex DNA datasets with Google Genomics" (<https://cloud.google.com/customers/dnastack/>). Google Cloud Platform. Retrieved 1 October 2016.
128. "23andMe – Ancestry" (<https://www.23andme.com/en-int/ancestry/>). *23andme.com*. Retrieved 29 December 2016.
129. Potenza, Alessandra (13 July 2016). "23andMe wants researchers to use its kits, in a bid to expand its collection of genetic data" (<https://www.theverge.com/2016/7/13/12166960/23andme-genetic-testing-database-genotyping-research>). *The Verge*. Retrieved 29 December 2016.
130. "This Startup Will Sequence Your DNA, So You Can Contribute To Medical Research" (<http://www.fastcompany.com/3066775/innovation-agents/this-startup-will-sequence-your-dna-so-you-can-contribute-to-medical-research>). *Fast Company*. 23 December 2016. Retrieved 29 December 2016.

131. 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.
132. Zaleski, Andrew (22 June 2016). "This biotech start-up is betting your genes will yield the next wonder drug" (<https://www.cnbc.com/2016/06/22/23andme-thinks-your-genes-are-the-key-to-blockbuster-drugs.html>). CNBC. Retrieved 29 December 2016.
133. 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.
134. "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.
135. Admire Moyo (23 October 2015). "Data scientists predict Springbok defeat" ([http://www.itweb.co.za/index.php?option=com\\_content&view=article&id=147241](http://www.itweb.co.za/index.php?option=com_content&view=article&id=147241)). *itweb.co.za*. Retrieved 12 December 2015.
136. Regina Pazvakavambwa (17 November 2015). "Predictive analytics, big data transform sports" ([http://www.itweb.co.za/index.php?option=com\\_content&view=article&id=147852](http://www.itweb.co.za/index.php?option=com_content&view=article&id=147852)). *itweb.co.za*. Retrieved 12 December 2015.
137. Dave Ryan (13 November 2015). "Sports: Where Big Data Finally Makes Sense" ([https://www.huffingtonpost.com/dave-ryan/sports-where-big-data-fin\\_b\\_8553884.html](https://www.huffingtonpost.com/dave-ryan/sports-where-big-data-fin_b_8553884.html)). *huffingtonpost.com*. Retrieved 12 December 2015.
138. Frank Bi. "How Formula One Teams Are Using Big Data To Get The Inside Edge" (<https://www.forbes.com/sites/frankbi/2014/11/13/how-formula-one-teams-are-using-big-data-to-get-the-inside-edge/>). *Forbes*. Retrieved 12 December 2015.
139. Tay, Liz. "Inside eBay's 90PB data warehouse" (<http://www.itnews.com.au/news/inside-ebay-8217s-90pb-data-warehouse-342615>). ITNews. Retrieved 12 February 2016.
140. Layton, Julia (25 January 2006). "Amazon Technology" (<http://money.howstuffworks.com/amazon1.htm>). *Money.howstuffworks.com*. Retrieved 5 March 2013.
141. "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 21 July 2013.
142. Constine, Josh (27 June 2017). "Facebook now has 2 billion monthly users... and responsibility" (<https://techcrunch.com/2017/06/27/facebook-2-billion-users/>). *TechCrunch*. Retrieved 3 September 2018.
143. "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.
144. Haleem, Abid; Javaid, Mohd; Khan, Ibrahim; Vaishya, Raju (2020). "Significant Applications of Big Data in COVID-19 Pandemic" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204193>). *Indian Journal of Orthopaedics*. **54** (4): 526–528. doi:10.1007/s43465-020-00129-z (<https://doi.org/10.1007/s43465-020-00129-z>). PMC 7204193 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204193>). PMID 32382166 (<https://pubmed.ncbi.nlm.nih.gov/32382166>).
145. Manancourt, Vincent (10 March 2020). "Coronavirus tests Europe's resolve on privacy" (<https://www.politico.eu/article/coronavirus-tests-europe-resolve-on-privacy-tracking-apps-germany-italy/>). *Politico*. Retrieved 30 October 2020.
146. Choudhury, Amit Roy (27 March 2020). "Gov in the Time of Corona" (<https://govinsider.asia/innovation/gov-in-the-time-of-corona/>). *Gov Insider*. Retrieved 30 October 2020.

147. Cellan-Jones, Rory (11 February 2020). "China launches coronavirus 'close contact detector' app" (<https://web.archive.org/web/20200228003957/https://www.bbc.com/news/technology-51439401>). *BBC*. Archived from the original (<https://www.bbc.com/news/technology-51439401>) on 28 February 2020. Retrieved 30 October 2020.
148. Siwach, Gautam; Esmailpour, Amir (March 2014). *Encrypted Search & Cluster Formation in Big Data* (<https://web.archive.org/web/20140809045242/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. Archived from the original (<http://asee-ne.org/proceedings/2014/Student%20Papers/210.pdf>) (PDF) on 9 August 2014. Retrieved 26 July 2014.
149. "Obama Administration Unveils "Big Data" Initiative:Announces \$200 Million in New R&D Investments" ([https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big\\_data\\_press\\_release\\_final\\_2.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf)) (PDF). *Office of Science and Technology Policy*. Archived ([https://web.archive.org/web/20170121233309/https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big\\_data\\_press\\_release\\_final\\_2.pdf](https://web.archive.org/web/20170121233309/https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf)) (PDF) from the original on 21 January 2017 – via *National Archives*.
150. "AMPLab at the University of California, Berkeley" (<http://amplab.cs.berkeley.edu>). *Amplab.cs.berkeley.edu*. Retrieved 5 March 2013.
151. "NSF Leads Federal Efforts in Big Data" ([https://www.nsf.gov/news/news\\_summ.jsp?cntn\\_id=123607&org=NSF&from=news](https://www.nsf.gov/news/news_summ.jsp?cntn_id=123607&org=NSF&from=news)). National Science Foundation (NSF). 29 March 2012.
152. 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/>).
153. 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>). *The New York Times*.
154. "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*.
155. Young, Shannon (30 May 2012). "Mass. governor, MIT announce big data initiative" ([http://archive.boston.com/news/local/massachusetts/articles/2012/05/30/mass\\_gov\\_and\\_mit\\_to\\_announce\\_data\\_initiative/](http://archive.boston.com/news/local/massachusetts/articles/2012/05/30/mass_gov_and_mit_to_announce_data_initiative/)). *Boston.com*. Retrieved 29 July 2021.
156. "Big Data @ CSAIL" (<http://bigdata.csail.mit.edu/>). *Bigdata.csail.mit.edu*. 22 February 2013. Retrieved 5 March 2013.
157. "Big Data Public Private Forum" (<https://cordis.europa.eu/project/id/318062>). *cordis.europa.eu*. 1 September 2012. Retrieved 16 March 2020.
158. "Alan Turing Institute to be set up to research big data" (<https://www.bbc.co.uk/news/technology-26651179>). *BBC News*. 19 March 2014. Retrieved 19 March 2014.
159. "Inspiration day at University of Waterloo, Stratford Campus" (<https://web.archive.org/web/20140226181442/http://www.betakit.com/event/inspiration-day-at-university-of-waterloo-stratford-campus/>). *betakit.com/*. Archived from the original (<http://www.betakit.com/event/inspiration-day-at-university-of-waterloo-stratford-campus/>) on 26 February 2014. Retrieved 28 February 2014.
160. Reips, Ulf-Dietrich; Matzat, Uwe (2014). "Mining "Big Data" using Big Data Services" ([http://www.ijis.net/ijis9\\_1/ijis9\\_1\\_editorial\\_pre.html](http://www.ijis.net/ijis9_1/ijis9_1_editorial_pre.html)). *International Journal of Internet Science*. **1** (1): 1–8.

161. Preis T, Moat HS, Stanley HE, Bishop SR (2012). "Quantifying the advantage of looking forward" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3320057>). *Scientific Reports*. **2**: 350. Bibcode:2012NatSR...2E.350P (<https://ui.adsabs.harvard.edu/abs/2012NatSR...2E.350P>). doi:10.1038/srep00350 (<https://doi.org/10.1038%2Fsrep00350>). PMC 3320057 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3320057>). PMID 22482034 (<https://pubmed.ncbi.nlm.nih.gov/22482034>).
162. Marks, Paul (5 April 2012). "Online searches for future linked to economic success" (<https://www.newscientist.com/article/dn21678-online-searches-for-future-linked-to-economic-success.html>). *New Scientist*. Retrieved 9 April 2012.
163. 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.
164. Tobias Preis (24 May 2012). "Supplementary Information: The Future Orientation Index is available for download" ([http://www.tobiaspreis.de/bigdata/future\\_orientation\\_index.pdf](http://www.tobiaspreis.de/bigdata/future_orientation_index.pdf)) (PDF). Retrieved 24 May 2012.
165. 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*. doi:10.1038/nature.2013.12879 (<https://doi.org/10.1038%2Fnature.2013.12879>). S2CID 167357427 (<https://api.semanticscholar.org/CorpusID:167357427>). Retrieved 9 August 2013.
166. Preis T, Moat HS, Stanley HE (2013). "Quantifying trading behavior in financial markets using Google Trends" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3635219>). *Scientific Reports*. **3**: 1684. Bibcode:2013NatSR...3E1684P (<https://ui.adsabs.harvard.edu/abs/2013NatSR...3E1684P>). doi:10.1038/srep01684 (<https://doi.org/10.1038%2Fsrep01684>). PMC 3635219 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3635219>). PMID 23619126 (<https://pubmed.ncbi.nlm.nih.gov/23619126>).
167. 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/>). *The New York Times*. Retrieved 9 August 2013.
168. 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*. Retrieved 9 August 2013.
169. 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*. doi:10.1038/nature.2013.12879 (<https://doi.org/10.1038%2Fnature.2013.12879>). S2CID 167357427 (<https://api.semanticscholar.org/CorpusID:167357427>). Retrieved 9 August 2013.
170. 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.
171. Hamish McRae (28 April 2013). "Hamish McRae: Need a valuable handle on investor sentiment? Google it" (<https://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.
172. Richard Waters (25 April 2013). "Google search proves to be new word in stock market prediction" (<https://www.ft.com/intl/cms/s/0/e5d959b8-acf2-11e2-b27f-00144feabdc0.html>). *Financial Times*. Retrieved 9 August 2013.
173. Jason Palmer (25 April 2013). "Google searches predict market moves" (<https://www.bbc.co.uk/news/science-environment-22293693>). *BBC*. Retrieved 9 August 2013.
174. E. Sejdić (March 2014). "Adapt current tools for use with big data". *Nature*. **507** (7492): 306.

175. Stanford. "MMDS. Workshop on Algorithms for Modern Massive Data Sets" (<https://web.stanford.edu/group/mmds/>).
176. Deepan Palguna; Vikas Joshi; Venkatesan Chakravarthy; Ravi Kothari & L. V. Subramaniam (2015). *Analysis of Sampling Algorithms for Twitter*. *International Joint Conference on Artificial Intelligence*.
177. Chris Kimble; Giannis Milolidakis (7 October 2015). "Big Data and Business Intelligence: Debunking the Myths". *Global Business and Organizational Excellence*. **35** (1): 23–34. arXiv:1511.03085 (<https://arxiv.org/abs/1511.03085>). doi:10.1002/JOE.21642 (<https://doi.org/10.1002%2FJOE.21642>). ISSN 1932-2054 (<https://www.worldcat.org/issn/1932-2054>). Wikidata Q56532925.
178. 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/](https://www.wired.com/science/discoveries/magazine/16-07/pb_theory/)). *Wired*.
179. 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.
180. Shah, Shvetank; Horne, Andrew; Capellá, Jaime (April 2012). "Good Data Won't Guarantee Good Decisions" (<http://hbr.org/2012/04/good-data-wont-guarantee-good-decisions/ar/1>). *Harvard Business Review*. Retrieved 8 September 2012.
181. Big Data requires Big Visions for Big Change. (<https://www.youtube.com/watch?v=UXef6yfJZAI>), Hilbert, M. (2014). London: TEDx UCL, x=independently organized TED talks
182. Alemany Oliver, Mathieu; Vayre, Jean-Sebastien (2015). "Big Data and the Future of Knowledge Production in Marketing Research: Ethics, Digital Traces, and Abductive Reasoning". *Journal of Marketing Analytics*. **3** (1): 5–13. doi:10.1057/jma.2015.1 (<https://doi.org/10.1057%2Fjma.2015.1>). S2CID 111360835 (<https://api.semanticscholar.org/CorpusID:111360835>).
183. Jonathan Rauch (1 April 2002). "Seeing Around Corners" (<https://www.theatlantic.com/magazine/archive/2002/04/seeing-around-corners/302471/>). *The Atlantic*.
184. Epstein, J. M., & Axtell, R. L. (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. A Bradford Book.
185. "Delort P., Big data in Biosciences, Big Data Paris, 2012" (<http://www.bigdataparis.com/documents/Pierre-Delort-INSERM.pdf#page=5>) (PDF). *Bigdataparis.com*. Retrieved 8 October 2017.
186. "Next-generation genomics: an integrative approach" (<https://www.cs.cmu.edu/~durand/03-71/2011/Literature/Next-Gen-Genomics-NRG-2010.pdf>) (PDF). *nature*. July 2010. Retrieved 18 October 2016.
187. "Big Data in Biosciences" (<https://www.researchgate.net/publication/283298499>). October 2015. Retrieved 18 October 2016.
188. "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.
189. Ohm, Paul (23 August 2012). "Don't Build a Database of Ruin" ([http://blogs.hbr.org/cs/2012/08/dont\\_build\\_a\\_database\\_of\\_ruin.html](http://blogs.hbr.org/cs/2012/08/dont_build_a_database_of_ruin.html)). *Harvard Business Review*.
190. Bond-Graham, Darwin (2018). "The Perspective on Big Data" (<https://www.theperspective.com/debates/the-perspective-on-big-data/>). *The Perspective*.
191. Al-Rodhan, Nayef (16 September 2014). "The Social Contract 2.0: Big Data and the Need to Guarantee Privacy and Civil Liberties – Harvard International Review" (<https://web.archive.org/web/20170413090835/http://hir.harvard.edu/the-social-contract-2-0-big-data-and-the-need-to-guarantee-privacy-and-civil-liberties/>). *Harvard International Review*. Archived from the original (<http://hir.harvard.edu/the-social-contract-2-0-big-data-and-the-need-to-guarantee-privacy-and-civil-liberties/>) on 13 April 2017. Retrieved 3 April 2017.



192. Barocas, Solon; Nissenbaum, Helen; Lane, Julia; Stodden, Victoria; Bender, Stefan; Nissenbaum, Helen (June 2014). *Big Data's End Run around Anonymity and Consent*. Cambridge University Press. pp. 44–75. doi:10.1017/cbo9781107590205.004 (<https://doi.org/10.1017%2Fcbo9781107590205.004>). ISBN 9781107067356. S2CID 152939392 (<https://api.semanticscholar.org/CorpusID:152939392>).
193. Lugmayr, Artur; Stockleben, Bjoern; Scheib, Christoph; Mailaparampil, Mathew; Mesia, Noora; Ranta, Hannu; Lab, Emmi (1 June 2016). "A Comprehensive Survey On Big-Data Research and Its Implications – What is Really 'New' in Big Data? – It's Cognitive Big Data!" (<https://www.researchgate.net/publication/304784955>).
194. 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 18 April 2011.
195. Katyal, Sonia K. (2019). "Artificial Intelligence, Advertising, and Disinformation" (<https://music.jhu.edu/article/745987>). *Advertising & Society Quarterly*. **20** (4). doi:10.1353/asr.2019.0026 (<https://doi.org/10.1353%2Fasr.2019.0026>). ISSN 2475-1790 (<https://www.worldcat.org/issn/2475-1790>). S2CID 213397212 (<https://api.semanticscholar.org/CorpusID:213397212>).
196. 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](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>).
197. 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>). hdl:10983/1320 (<https://hdl.handle.net/10983%2F1320>). S2CID 51843165 (<https://api.semanticscholar.org/CorpusID:51843165>).
198. Failure to Launch: From Big Data to Big Decisions ([http://www.fortewares.com/Administrator/userfiles/Banner/forte-ware--pro-active-reporting\\_EN.pdf](http://www.fortewares.com/Administrator/userfiles/Banner/forte-ware--pro-active-reporting_EN.pdf)) Archived ([https://web.archive.org/web/20161206145026/http://www.fortewares.com/Administrator/userfiles/Banner/forte-ware--pro-active-reporting\\_EN.pdf](https://web.archive.org/web/20161206145026/http://www.fortewares.com/Administrator/userfiles/Banner/forte-ware--pro-active-reporting_EN.pdf)) 6 December 2016 at the Wayback Machine, Forte Wares.
199. "15 Insane Things That Correlate with Each Other" (<https://www.tylervigen.com/spurious-correlations>).
200. Random structures & algorithms (<https://onlinelibrary.wiley.com/loi/10982418>)
201. Cristian S. Calude, Giuseppe Longo, (2016), The Deluge of Spurious Correlations in Big Data, *Foundations of Science*
202. Anja Lambrecht and Catherine Tucker (2016) "The 4 Mistakes Most Managers Make with Analytics," *Harvard Business Review*, July 12. <https://hbr.org/2016/07/the-4-mistakes-most-managers-make-with-analytics>
203. 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 13 August 2014.
204. Pelt, Mason (26 October 2015). "'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*. Retrieved 4 November 2015.
205. Harford, Tim (28 March 2014). "Big data: are we making a big mistake?" (<https://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html>). *Financial Times*. Retrieved 7 April 2014.

206. Ioannidis JP (August 2005). "Why most published research findings are false" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327>). *PLOS Medicine*. **2** (8): e124. doi:10.1371/journal.pmed.0020124 (<https://doi.org/10.1371%2Fjournal.pmed.0020124>). PMC 1182327 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327>). PMID 16060722 (<https://pubmed.ncbi.nlm.nih.gov/16060722>).
207. Lohr, Steve; Singer, Natasha (10 November 2016). "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 27 November 2016.
208. "How data-driven policing threatens human freedom" (<https://www.economist.com/open-future/2018/06/04/how-data-driven-policing-threatens-human-freedom>). *The Economist*. 4 June 2018. ISSN 0013-0613 (<https://www.worldcat.org/issn/0013-0613>). Retrieved 27 October 2019.
209. Brayne, Sarah (29 August 2017). "Big Data Surveillance: The Case of Policing". *American Sociological Review*. **82** (5): 977–1008. doi:10.1177/0003122417725865 (<https://doi.org/10.1177%2F0003122417725865>). S2CID 3609838 (<https://api.semanticscholar.org/CorpusID:3609838>).



## Further reading

---

- Peter Kinnaird; Inbal Talgam-Cohen, eds. (2012). "Big Data" (<http://dl.acm.org/citation.cfm?id=2331042>). *XRDS: Crossroads, The ACM Magazine for Students*. Vol. 19, no. 1. Association for Computing Machinery. ISSN 1528-4980 (<https://www.worldcat.org/issn/1528-4980>). OCLC 779657714 (<https://www.worldcat.org/oclc/779657714>).
- Jure Leskovec; Anand Rajaraman; Jeffrey D. Ullman (2014). *Mining of massive datasets* (<http://mmds.org/>). Cambridge University Press. ISBN 9781107077232. OCLC 888463433 (<https://www.worldcat.org/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 (<https://www.worldcat.org/oclc/828620988>).
- Press, Gil (9 May 2013). "A Very Short History of Big Data" (<https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data>). *forbes.com*. Jersey City, NJ. Retrieved 17 September 2016.
- Stephens-Davidowitz, Seth (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. Dey Street Books. ISBN 978-0062390851.
- "Big Data: The Management Revolution" (<https://hbr.org/2012/10/big-data-the-management-revolution>). *Harvard Business Review*. October 2012.
- O'Neil, Cathy (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books. ISBN 978-0553418835.

## External links

---

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

Retrieved from "[https://en.wikipedia.org/w/index.php?title=Big\\_data&oldid=1070183707](https://en.wikipedia.org/w/index.php?title=Big_data&oldid=1070183707)"

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.