Big data

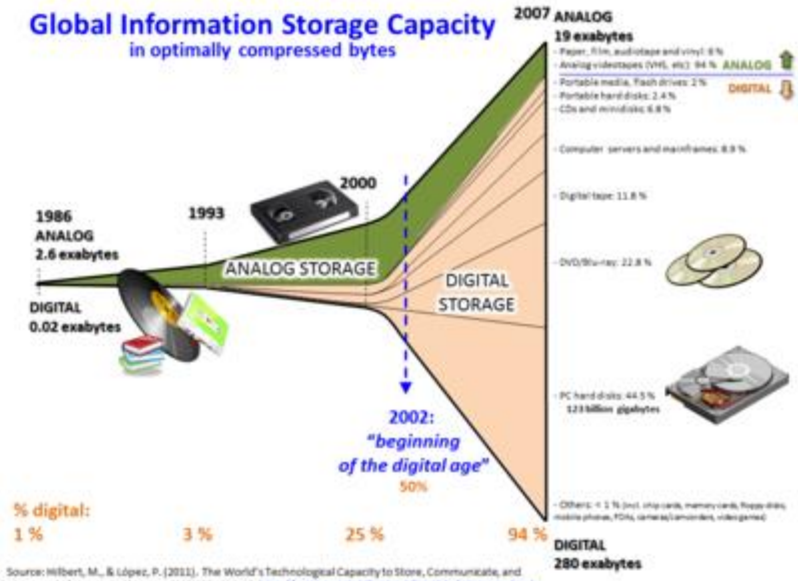
62 languages

- Article
- Talk
- Read
- Edit
- View history

Tools

From Wikipedia, the free encyclopedia

This article is about large collections of data. For the band, see Big Data (band). For the practice of buying and selling of personal and consumer data, see Surveillance capitalism.



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

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

Big data primarily refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many entries (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.[2] Though used sometimes loosely partly due to a lack of formal definition, the best interpretation is that it is a large body of information that cannot be comprehended when used in small amounts only.[3]

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*.[4] The analysis of big data presents challenges in sampling, and thus previously allowing for only observations and sampling. Thus a fourth concept, *veracity,* refers to the quality or insightfulness of the data. Without sufficient investment in expertise for big data veracity, the volume and variety of data can produce costs and risks that exceed an organization's capacity to create and capture *value* from big data.[5]

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."[6] Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on".[7] 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,[8] connectomics, complex physics simulations, biology, and environmental research.[9]

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) equipment, software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.[10][11] The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;[12] as of 2012, every day 2.5 exabytes ($2.17×2^{60}$ bytes) of data are generated.[13] 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.[14] According to IDC, global spending on big data and business analytics (BDA) solutions is estimated to reach $215.7 billion in 2021.[15][16] While Statista report, the global big data market is forecasted to grow to $103 billion by 2027.[17] In 2011 McKinsey & Company reported, if US healthcare were to use big data creatively and effectively to drive efficiency and quality, the sector could create more than $300 billion in value every year.[18] In the developed economies of Europe, government administrators could save more than €100 billion ($149 billion) in operational efficiency improvements alone by using big data.[18] And users of services enabled by personal-location data could capture $600 billion in consumer surplus.[18] One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.[19]

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".[20] 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."[21]

Definition[edit]

The term *big data* has been in use since the 1990s, with some giving credit to John Mashey for popularizing the term.[22][23] 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.[24][*page needed*] Big data philosophy encompasses unstructured, semi-structured and structured data; however, the main focus is on unstructured data.[25] Big data "size" is a constantly moving target; as of 2012 ranging from a few dozen terabytes to many zettabytes of data.[26] 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.[27]

"Variety", "veracity", and various other "Vs" are added by some organizations to describe it, a revision challenged by some industry authorities.[28] 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.[4] 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."[29]

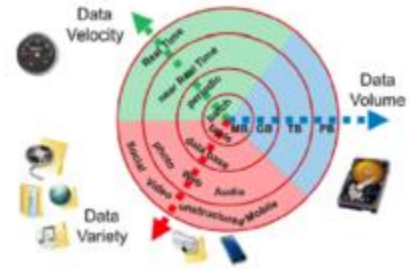
In a comparative study of big datasets, Kitchin and McArdle found that none of the commonly considered characteristics of big data appear consistently across all of the analyzed cases.[30] For this reason, other studies identified the redefinition of power dynamics in knowledge discovery as the defining trait.[31] Instead of focusing on the 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[edit]

The growing maturity of the concept more starkly delineates the difference between "big data" and "business intelligence":[32]

- 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[33] to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density[34] to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.[33][35][*promotional source?*]

Characteristics[edit]



This image shows the growth of big data's primary characteristics of volume, velocity, and variety.

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.[36]

Variety

The type and nature of the data. 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. 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.[37]

Veracity

The truthfulness or reliability of the data, which refers to the data quality and the data value.[38] 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.[39]

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.[40] 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:[41]

Exhaustive

Whether the entire system (i.e., =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[edit]

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.[42][*promotional source?*]

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.[43] 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.[44] 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,[45] so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open-source project named "Hadoop".[46] Apache Spark was developed in 2012 in response to limitations in the MapReduce paradigm, as it adds in-memory processing and 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".[47] 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.[48]

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.[49]

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.[50][51]

Technologies[edit]

A 2011 McKinsey Global Institute report characterizes the main components and ecosystem of big data as follows:[52]

- 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,[53] such as multilinear subspace learning,[54] massively parallel-processing (MPP) databases, search-based applications, data mining,[55] distributed file systems, distributed cache (e.g., burst buffer and Memcached), distributed databases, cloud and HPC-based infrastructure (applications, storage and computing resources),[56] and the Internet.[*citation needed*] Although, many approaches and technologies have been developed, it still remains difficult to carry out machine learning with big data.[57]

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.[58][promotional source?]

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".[59][third-party source needed]

The practitioners of big data analytics processes are generally hostile to slower shared storage,[60] 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[edit]

Bus wrapped with SAP big data parked outside IDF13

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.[7]

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.[7] 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[12] and predictions put the amount of internet traffic at 667 exabytes annually by 2014.[7] According to one estimate, one-third of the globally stored information is in the form of alphanumeric text and still image data,[61] 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).

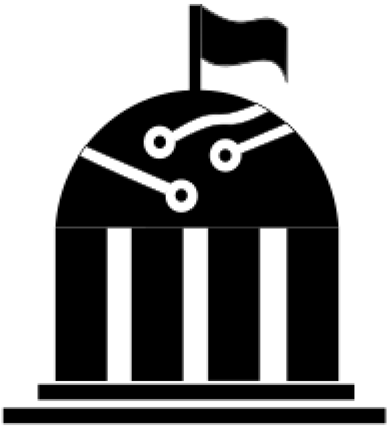
While many vendors offer off-the-shelf products for big data, experts promote the development of in-house custom-tailored systems if the company has sufficient technical capabilities.[62]

Government[edit]

See also: Government by algorithm

This section **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources in this section. Unsourced material may be challenged and removed.

Find sources: "Big data" – news · newspapers · books · scholar · JSTOR *(September 2023)* *(Learn how and when to remove this template message)*



The use and adoption of big data within governmental processes allows efficiencies in terms of cost, productivity, and innovation,[63] 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[edit]

Research on the effective usage of information and communication technologies for development (also known as "ICT4D") suggests that big data technology can make important contributions but also present unique challenges to international development.[64][65] 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.[66][page needed][67][68] Additionally, user-generated data offers new opportunities to give the unheard a voice.[69] 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.[66][page needed] The challenge of "big data for development"[66][page needed] is currently evolving toward the application of this data through machine learning, known as "artificial intelligence for development (AI4D).[70]

Benefits[edit]

A major practical application of big data for development has been "fighting poverty with data".[71] In 2015, Blumenstock and colleagues estimated predicted poverty and wealth from mobile phone metadata [72] and in 2016 Jean and colleagues combined satellite imagery and machine learning to predict poverty.[73] Using digital trace data to study the labor market and the digital economy in Latin America, Hilbert and colleagues [74][75] argue that digital trace data has several benefits such as:

- Thematic coverage: including areas that were previously difficult or impossible to measure
- Geographical coverage: providing 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[edit]

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.[76]
- 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 deal with data.[74]

Finance[edit]

Big Data is being rapidly adopted in Finance to 1) speed up processing and 2) deliver better, more informed inferences, both internally and to the clients of the financial institutions.[77] The financial applications of Big Data range from investing decisions and trading (processing volumes of available price data, limit order books, economic data and more, all at the same time), portfolio management (optimizing over an increasingly large array of financial instruments, potentially selected from different asset classes), risk management (credit rating based on extended information), and any other aspect where the data inputs are large.[78]

Healthcare[edit]

Big data analytics has been used in healthcare in 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.[79][80][81][82] 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.[83] "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.[84] While extensive information in healthcare is now electronic, it fits under the big data umbrella as most is unstructured and difficult to use.[85] 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.[86]

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.[87] 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.[88][page needed] For instance, for epilepsy monitoring it is customary to create 5 to 10 GB of data daily.[89] Similarly, a single uncompressed image of breast tomosynthesis averages 450 MB of data.[90] 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.[91]

Education[edit]

A McKinsey Global Institute study found a shortage of 1.5 million highly trained data professionals and managers[52] and a number of universities[92][better source needed] 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.[93] In the specific field of marketing, one of the problems stressed by Wedel and Kannan[94] is that marketing has several sub domains (e.g., advertising, promotions, product development, branding) that all use different types of data.

Media[edit]

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.[95]

- Targeting of consumers (for advertising by marketers)[96]
- 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.[97]

Insurance[edit]

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.[98]

Internet of things (IoT)[edit]

Main article: Internet of things

Further information: Edge computing

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,[99] manufacturing[100] and transportation[101] contexts.

Kevin Ashton, the digital innovation expert who is credited with coining the term,[102] 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[edit]

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).[103] By applying big data principles into the concepts of machine intelligence and deep computing, IT departments can predict potential issues and prevent them.[103] 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.

Survey science[edit]

Compared to survey-based data collection, big data has low cost per data point, applies analysis techniques via machine learning and data mining, and includes diverse and new data sources, e.g., registers, social media, apps, and other forms digital data. Since 2018, survey scientists have started to examine how big data and survey science can complement each other to allow researchers and practitioners to improve the production of statistics and its quality. To date, there have been three Big Data Meets Survey Science

(BigSurv) conferences in 2018, 2020 (virtual), 2023, and as of 2023 one conference forthcoming in 2025,[104] a special issue in the *Social Science Computer Review*,[105] a special issue in *Journal of the Royal Statistical Society*,[106] and a special issue in *EP J Data Science*,[107] and a book called *Big Data Meets Social Sciences*[108] edited by Craig Hill and five other Fellows of the American Statistical Association. In 2021, the founding members of BigSurv received the Warren J. Mitofsky Innovators Award from the American Association for Public Opinion Research.[109]

Marketing[edit]

Big data is notable in marketing due to the constant “datafication”[110] of everyday consumers of the internet, in which all forms of data are tracked. The datafication of consumers can be defined as quantifying many of or all human behaviors for the purpose of marketing.[111] The increasingly digital world of rapid datafication makes this idea relevant to marketing because the amount of data constantly grows exponentially. It is predicted to increase from 44 to 163 zettabytes within the span of five years.[112] The size of big data can often be difficult to navigate for marketers.[113] As a result, adopters of big data may find themselves at a disadvantage. Algorithmic findings can be difficult to achieve with such large datasets.[114] Big data in marketing is a highly lucrative tool that can be used for large corporations, its value being as a result of the possibility of predicting significant trends, interests, or statistical outcomes in a consumer-based manner.[115]

There are three significant factors in the use of big data in marketing:

1. Big data provides customer behavior pattern spotting for marketers, since all human actions are being quantified into readable numbers for marketers

to analyze and use for their research.[116]

2. Real-time market responsiveness is important for marketers because of the ability to shift marketing efforts and correct to current trends, which is helpful in maintaining relevance to consumers. This can supply corporations with the information necessary to predict the wants and needs of consumers in advance.[117]
3. Data-driven market ambidexterity are being highly fueled by big data.[118] New models and algorithms are being developed to make significant predictions about certain economic and social situations.[119]

Case studies[edit]

Government[edit]

China[edit]

- 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.[120] Biometrics, including DNA samples, are gathered through a program of free physicals.[121]
- By 2020, China plans to give all its citizens a personal "social credit" score based on how they behave.[122] 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.[123][124]

India[edit]

- Big data analysis was tried out for the BJP to win the 2014 Indian General Election.[125]
- 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[edit]

- Personalized diabetic treatments can be created through GlucoMe's big data solution.[126]

United Kingdom[edit]

Examples of uses of big data in public services:

- Data on prescription drugs: by connecting origin, location and the time of each prescription, a research unit was able to exemplify 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.[*citation needed*][127]
- 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.[128]

United States[edit]

- In 2012, the Obama administration announced the Big Data Research and Development Initiative, to explore how big data could be used to address important problems faced by the government.[129] The initiative is composed of 84 different big data programs spread across six departments.[130]

- Big data analysis played a large role in Barack Obama's successful 2012 re-election campaign.[131]
- The United States Federal Government owns five of the ten most powerful supercomputers in the world.[132][133]
- 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.[134][135][136] This has posed security concerns regarding the anonymity of the data collected.[137]

Retail[edit]

- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data—the equivalent of 167 times the information contained in all the books in the US Library of Congress.[7]
- 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.[138]
- FICO Card Detection System protects accounts worldwide.[139]

Science[edit]

- The Large Hadron Collider experiments represent about 150 million sensors delivering data

40 million times per second. There are nearly 600 million collisions per second. After filtering and refraining from recording more than 99.99995%[140] of these streams, there are 1,000 collisions of interest per second.[141][142][143]

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

- The Square Kilometre Array is a radio telescope built of thousands of antennas. It is expected to be operational by 2024. Collectively, these antennas are expected to gather 14 exabytes and store one petabyte per day.[144][145] It is considered one of the most ambitious scientific projects ever undertaken.[146]
- 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.[7] 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.[7]

- 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 less expensive than the reduction in cost predicted by Moore's law.[147]
- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.[148][149]
- 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.[150][151]
- 23andme's DNA database contains the genetic information of over 1,000,000 people worldwide.[152] The company explores selling the "anonymous aggregated genetic data" to other researchers and pharmaceutical companies for

research purposes if patients give their consent.[153][154][155][156][157] 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.[153] 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.[158]

- Computational fluid dynamics (CFD) and hydrodynamic turbulence research generate massive data sets. The Johns Hopkins Turbulence Databases (JHTDB) 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[edit]

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.[159] Future performance of players could be predicted as well.[160] Thus, players' value and salary is determined by data collected throughout the season.[161]

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.[162] 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.[163]

Technology[edit]

- As of 2013, 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.[164]
- 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.[165]
- Facebook handles 50 billion photos from its user base.[166] As of June 2017, Facebook reached 2 billion monthly active users.[167]
- Google was handling roughly 100 billion searches per month as of August 2012.[168]

COVID-19[edit]

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.[169]

Governments used big data to track infected people to minimise spread. Early adopters included China, Taiwan, South Korea, and Israel.[170][171][172]

Research activities[edit]

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.[173]

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.[174]

The initiative included a National Science Foundation "Expeditions in Computing" grant of $10 million over five years to the AMPLab[175] at the University of California, Berkeley.[176] 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[177] to fighting cancer.[178]

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,[179] 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.[180] 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.[181]

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.[182]

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.[183]

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.[184]

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.[185] Often these APIs are provided for free.[185] 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.[186][187][188] 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".[189] 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.[190] Their analysis of Google search volume for 98 terms of varying financial relevance, published in *Scientific Reports*,[191] suggests that increases in search volume for financially relevant search terms tend to precede large losses in financial markets.[192][193][194][195][196][197][198]

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.[199]

Sampling big data[edit]

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 is 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.

Critique[edit]

Critiques of the big data paradigm come in two flavors: those that question the implications of the approach itself, and those that question the way it is currently done.[200] One approach to this criticism is the field of critical data studies.

Critiques of the big data paradigm[edit]

"A crucial problem is that we do not know much about the underlying empirical micro-processes that lead to the emergence of the[se] typical network characteristics of Big Data."[24][*page needed*] 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:[201] focusing in particular on the notion that big data must always be contextualized in their social, economic, and political contexts.[202] 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*.[203]

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".[66][*page needed*] Fed by a large number of data on past experiences, algorithms can predict future development if the future is similar to the past.[204] 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.[204] 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".[205] Additionally, it has been suggested to combine big data approaches with computer simulations, such as agent-based models[66][*page needed*] 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.[206][207] 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.[208] 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.[209][210] In the massive approaches it is the formulation of a relevant hypothesis to explain the data that is the limiting factor.[211] The search logic is reversed and the limits of induction ("Glory of Science and Philosophy scandal", C. D. Broad, 1926) are to be considered.[*citation needed*]

Privacy advocates are concerned about the threat to privacy represented by increasing storage and integration of personally identifiable information; expert panels have released various policy recommendations to conform practice to expectations of privacy.[212] 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.[213]

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.[214]

Critiques of the "V" model[edit]

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:[215]

- 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[edit]

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[edit]

Ulf-Dietrich Reips and Uwe Matzat wrote in 2014 that big data had become a "fad" in scientific research.[185] 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.[216] This approach may lead to results that have a bias in one way or another.[217] 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.[218] In the provocative article "Critical Questions for Big Data",[219] 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".[219] 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.[220] Big structures are full of spurious correlations[221] either because of non-causal coincidences (law of truly large numbers), solely nature of big randomness[222] (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.[223] 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."[224]

Big data analysis is often shallow compared to analysis of smaller data sets.[225] 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.[225]

Big data is a buzzword and a "vague term",[226][227] but at the same time an "obsession"[227] 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"[228] 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[229] with varying degrees of success.

Critiques of big data policing and surveillance[edit]

Big data has been used in policing and surveillance by institutions like law enforcement and corporations.[230] 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*,[231] big data policing can reproduce existing societal inequalities in three ways:

- Placing people 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.