Scientific and Social Research, 2026, Volume 8, Issue 4
http://ojs.bbwpublisher.com/index.php/ssr
Online ISSN: 2981-9946
Print ISSN: 2661-4332

Three Core Characteristics of Big Data Analysis and Their Significance

Wenying Jing*

Qingdao University, Qingdao 266071, China

Authors to whom correspondence should be addressed.

Abstract: In people's daily lives, big data is no longer a distant technological concept, but a real part of daily life that permeates clothing, food, housing, and transportation. When one opens their phone in the morning to read the news, the platform is pushing content that interests them; When ordering takeout, the app will prioritize the stores they often eat at; After placing an order online, the system can accurately predict the delivery time; Even when taking a taxi, the software can instantly calculate the best route, and so on. These seemingly simple and convenient experiences are all backed by big data analysis. In the past, people processed information through manual statistics and sampling surveys, which had small samples, low efficiency, and were prone to errors. The Internet, smartphones, and Internet of Things devices generate massive amounts of data every day, which contain users' habits, market trends, and industry patterns. Whoever can read this data will be able to seize opportunities, reduce risks, and improve efficiency. To understand big data, the key is to grasp its core features. Many people, upon hearing the term "big data", find it unfathomable and full of technical jargon that they can neither understand nor apply. In fact, big data is not that complicated. Its core logic is hidden in the three most critical characteristics, namely holomorphism, confounding, and correlation. These three characteristics are not only the fundamental difference between big data and traditional statistics, but also the core reason why big data can be valuable. This article will explain the three core characteristics of big data in the most common and down-to-earth language, and at the same time talk about what these characteristics are used for, what problems they can solve, and what changes they have brought to people's lives and work, so that everyone can truly understand the essence of big data and know why it has changed the way society operates in just over a decade.

Keywords: Big data analysis; Core characteristics; Massive scale

Online publication: May 25, 2026

1. Introduction

As early as 1980, futurist Alvin Toffler predicted the brilliant future of big data in his masterpiece "The Third Wave", in which he wrote: "If IBM's mainframe opened the curtain of the information revolution,

then big data is the brilliant chapter of the Third wave" [1–4]. In May 2011, EMC, known for its advocacy of cloud computing, introduced the concept of big data at its annual meeting "Cloud Computing Meets Big Data." In June of the same year, many foreign institutions, such as IBM and McKinsey, released big data-related research reports to actively follow up. "Data has permeated every industry and business function and is gradually becoming an important factor of production, and the use of massive data will herald the arrival of a new wave of productivity growth and consumer surplus", McKinsey noted in its research report. As a result, the "big data era" has gradually come into the public eye as a formal concept and triggered a series of subsequent social impacts [5].

2. The essential meaning of big data

Many people think that big data means "a lot of data", but this understanding is only half right. Traditional data statistics can also accumulate a lot of data, but they are completely different from big data [6]. Traditional data is "small but precise", aiming for accuracy, standardization, and structuring, such as a company's financial statements, a school student's grades, and a bank's deposit records. These data are all manually organized, with a uniform format and no impurities, and are easy to process, but can reflect very limited problems.

Big data, on the other hand, is not just "big in volume", but more importantly, "comprehensive" and "dynamic." It encompasses every click, every search, every stay, every consumption people make on the web, the temperature, humidity, location, speed recorded by sensors, and various forms of information such as text, images, audio, and video. These data may seem chaotic, but they are a complete record of the real behavior of people and things.

So the essence of big data is not simply the accumulation of data, but the use of comprehensive, real, real-time data to restore the world as it is and discover patterns that traditional statistics cannot find [7].

The traditional sampling survey is like "seeing the leopard through a tube", and it is easy to misjudge the appearance of the leopard by looking at just one spot on it. Big data, on the other hand, is like a panoramic view of the leopard, seeing its entire body, its movements, and its habits clearly, and the conclusions drawn are more reliable [8]. The significance of big data is to break the limitations of traditional data, allowing people to shift from "guessing trends" to "seeing facts", and from "making decisions based on experience" to "speaking with data."

Whether it is enterprises making products, governments managing, or individuals making choices, big data can provide the most authentic reference. It is not about complexity for the sake of complexity, but about making decisions more accurate, more efficient, and less costly — that is the core value of big data.

3. The three core features of big data

3.1. Sameness

One of the core features of big data is holomorphism. This is the most fundamental difference between big data and traditional data, and the most powerful aspect of big data.

What is holomorphism? To put it simply, it means not conducting a sample survey and directly analyzing all the data [9].

Before big data, if someone wanted to know about a group, they could only rely on "sampling." For example, if

someone wants to know what flavors the whole nation likes, it is impossible to ask every single person. They can only randomly ask a few thousand people, and then use the results of those few thousand people to represent the preferences of billions of people. To know if a product is good or not, they have to test it with a few hundred users and use the small range of feedback to judge the overall situation of the product [10].

This sampling method has a fatal problem: sampling bias. The results can be completely different depending on the people, the time, and the place. Just like looking into people's incomes, if one only goes to high-end neighborhoods in first-tier cities, the results will surely be higher; If one only goes to the countryside, the results will be too low to reflect the real situation at all.

Big data's homogeneity directly solves this problem. It does not need to "pick some people", but takes all users, all behaviors, all data for analysis. For example, if an e-commerce platform wants to know what products users like, it does not need sampling [11]. It directly analyzes the browsing, adding, and purchasing records of hundreds of millions of users, and every operation of each person is not missed. Only in this way can the results be the most genuine and comprehensive.

Take the most common example: recommendations on short-video platforms. Traditional video recommendations may rely on editors manually picking or sampling the preferences of a portion of users to recommend content that many people do not like. The current short-video platform uses big data homosameness, which records every user's likes, comments, shares, and duration of stay, whether it is the elderly, children, office workers, or students. All users' data is analyzed, and then the content that each person likes is precisely recommended. So one will find that the more they browse, the more they want to watch, because the platform knows them better than the person know themself, that is the power of homogeneity [12].

Another example is traffic management. In the past, to know which roads were congested, one could only rely on on-site inspections by traffic police and feedback from drivers. The information was lagging and incomplete. The current intelligent transportation system, through cameras on the road, GPS positioning, and ride-hailing data, collects driving data of all vehicles, all time periods, all sections of road, analyzes congestion in real time, automatically adjusts the duration of traffic lights, and plans the optimal route. This is where holality comes into play [13].

The core of holomorphism is to abandon the "small to big" guesswork and restore the truth with "all data." It does not omit any individual, does not overlook any detail, and ensures that the data results are no longer biased, which is the basis for big data to make precise predictions and decisions.

3.2. Confounding

The second core feature of big data is hybridity. Many people do not understand this feature and even consider it a drawback. In fact, it is an advantage of big data. What is confounding? Simply put, big data does not aim for 100% precision and standardization, allowing for impurities, errors, and inconsistent formats in the data.

Traditional data processing has particularly high requirements for data, which must be clean, standardized, and accurate. For example, when entering user information, names, phone numbers, and addresses must be exactly the same, and every wrong character must be corrected; The data format must be uniform, either numbers or text, and there must be no messy content. Because of the small sample size of traditional data, once there is an error, the result will be outrageously wrong. But big data is different. It has a huge volume of data and a wide range of sources, including

information that users input casually, fuzzy data automatically collected by sensors, data of different formats from different platforms, text, images, voice, and video, and is bound to have errors, omissions, and clutter. If, as with traditional data, all data is cleaned up before analysis, not only will it be time-consuming and laborious, but also a great deal of valuable information will be lost. So the logic of big data is: not to be the most precise, but to be the most comprehensive. Allowing the data to be mixed, accepting a small amount of error, and offsetting the error with a large amount of data can actually lead to more accurate results [14].

For example: product reviews on online shopping platforms. The reviews are full of positive, negative, teasing, pictures, videos, and even random typing and emojis. The data is very chaotic, with emotional expressions, repetitive content, and even maliciously inflated reviews. If one follows the traditional data requirements and clears out all these messy reviews, leaving only the standard text, then they would not be able to see the real thoughts of users. Big data accommodates all the mix, analyzing all the reviews, all the text, all the pictures, even if there are a small number of false reviews, in the face of a vast number of real reviews, the error will be offset, and ultimately, it can accurately judge the quality of the product and the pain points of the users. That is the value of mixality.

Another example is speech recognition. People speak with accents, pauses, catchphrases, and the surrounding noise. The data people recognize is bound to have errors, and that is the data mixality. If a speech recognition system pursues 100% accuracy and does not accept any errors, it simply cannot be used. But big data accommodates this mix. By analyzing massive amounts of voice data and constantly optimizing algorithms, it can accurately recognize what people say, even if there are accents and noises. Many people think that big data must be precise, but that is not the case. A small amount of data must be precise, because one wrong is all wrong; A large amount of data allows for mixing, because a large amount can cover errors, and comprehensiveness is more important than precision. Confounding enables big data to collect more dimensions and forms of information without discarding it because it is not standardized, which makes big data richer and closer to reality.

3.3 Relevance

What is relevance? It is about finding the connection between two things, knowing that “if A happens, B will most likely happen too”, without insisting on “why A causes B.”

Traditional logical thinking is particularly fond of finding “cause-and-effect relationships.” For example, when it gets cold, the cold is the cause and the cold is the effect. Study hard and get good grades. Effort is the cause, and good grades are the effect. People always get used to asking “why” and do not stop until they find out the reason.

But in the age of big data, the cause-and-effect relationships of many things are so complex that they are impossible to find or take a long time to find. This is where correlation comes in handy — regardless of why, as long as one knows there is a connection between the two, one can use that connection to solve the problem.

The most classic example is the “beer and diapers” in a supermarket. The supermarket found through big data analysis that men who buy diapers are likely to buy beer along the way. That is the correlation. As for why? It might be that dad is buying diapers for the kids and wants to get a beer to relax by the way; It could be other reasons, and no one can explain the exact cause-and-effect relationship, nor is it necessary to. It is enough for a supermarket to put beer and diapers together and boost sales of both products at the same time.

There are many such examples in people’s lives. For instance, food delivery platforms have found that more people order milk tea on rainy days. This is correlation. There is no need to study why people want to drink milk tea on

rainy days. Just increase the promotion of milk tea on rainy days and prepare more inventory to increase the number of orders.

For instance, short-video platforms have found that users who enjoy watching food-related videos are likely to also like watching kitchenware recommendations. This is what people call relevance. The platform does not have to figure out why users like it. Just promoting kitchenware to users who watch food can increase the conversion rate of sales.

Weather forecasts and disease predictions are all based on relevance. Big data analysis has found that the spread of certain viruses is associated with temperature and humidity. As long as changes in temperature and humidity are monitored, the risk of disease can be predicted in advance without having to fully understand all the complex causes and effects of virus transmission.

The core of correlation is to abandon complex cause-and-effect tracing and quickly identify patterns and solve problems using simple correlations. In a vast amount of data, causal relationships are deeply hidden and hard to find, but correlations are easy to discover and practical enough. Big data is not for research or getting to the bottom of things, but for quick implementation and creating value, and relevance is the key to achieving that goal.

To sum up the three core features of big data: Full sample means looking at all, not sampling, and ensuring the authenticity of the data; Confounding is to tolerate impurities, not be picky, and ensure the completeness of the data; Correlation is about finding connections, not getting entangled in cause and effect, and ensuring the practicality of the data. These three characteristics complement each other and are indispensable, together forming the core logic of big data.

4. Directions of action of the three core features of big data

4.1. Eliminating sampling bias

The drawbacks of traditional sampling surveys, as mentioned earlier, are that if the sample is not selected properly, the results will be distorted. For example, if one wants to survey the consumption concepts of young people, they only sample college students and ignore working people; If only first-tier cities are sampled, third- and fourth-tier cities are ignored, and the results are not representative at all. Big data, on the other hand, bypasses sampling and analyzes all the data, eliminating the bias caused by sampling from the root. Regardless of the group, region, or time period, all data are included in the analysis, and the results are naturally more authentic and reliable.

For example, when the government conducts population censuses and surveys of people's livelihood, it used to rely on manual household visits and sample statistics, which were time-consuming and had large errors. Now, by integrating big data and collecting all the data, such as household registration, medical insurance, social security, travel, and consumption, people can quickly and accurately grasp population structure and people's livelihood needs, and formulate policies that are more in line with reality. For instance, when companies conduct market research, it is easy to be out of touch with the market by sampling user demands before. Now, with big data full-sample analysis, people can precisely grasp the needs, pain points, and habits of all users. Product design and marketing promotion can precisely hit the users and avoid blind decision-making. Eliminating sampling bias and turning data from roughly accurate to completely true is the most fundamental change that big data brings to decision-making [15].

4.2. Unleashing the data source dividend

In the era of traditional data, a lot of data was simply discarded because it was messy, unregulated, and error-prone. Such as casual comments from users, blurry sensor data, unstructured images, and voices — these are all "junk data" in traditional processing methods, useless and taking up space. However, big data enables accepting these imperfect data, collecting them all, and analyzing them. These seemingly disorganized data contain a lot of details and patterns that were wasted before and are now being mined by big data to become the "data source dividend."

For instance, customer service chat records were previously merely evidence for solving problems, disorganized, and no one analyzed them. Now, big data analyzes all chat records and can identify the most frequently encountered problems and most concerning demands of users, allowing enterprises to optimize products and improve services. For instance, surveillance videos were previously only used for security purposes, and the vast amount of video data was not watched by anyone. Now, big data analysis of the flow of people, vehicles, and behaviors in videos can be used for traffic management, security early warning, and commercial location selection, making the otherwise useless video data of great value. The hybrid nature of big data makes it not picky, and it can be utilized regardless of the form or state of the data, revitalizing previously wasted data resources. This is the data source dividend.

4.3. Empower prediction and recommendation

In terms of predictions, big data can predict trends in advance through correlation analysis. For example, e-commerce companies predict which products will be bestsellers by using users' browsing and add-purchase data and stock up in advance; Meteorological departments predict the weather precisely by associating data such as temperature, air pressure, and humidity; Financial institutions predict credit risks and prevent fraud by associating users' consumption and repayment data.

In terms of recommendations, big data precisely matches user needs through correlation. For example, short video and news apps recommend the content someone is interested in, online shopping apps recommend the goods someone wants to buy, music apps recommend the songs someone likes to listen to, and food delivery apps recommend the food someone likes, all of which are driven by relevance.

Predictions help avoid risks and seize opportunities in advance, and recommendations help save time and enhance the experience. Relevance makes big data no longer a passive recorder but an active guide, which is the key to how big data can change lives and industries.

5. Summary

Big data is no longer a distant technological concept, but a fundamental tool integrated into our lives, work, and society. The core logic of it lies in the three characteristics of sameness, hybridity, and correlation. In the future, there will be more and more data, and the application of big data will become more and more widespread. As long as we grasp these three core features, we will be able to understand big data, make good use of big data, and seize opportunities and meet challenges in the data age. The essence of big data is never to show off technology, but to serve people and society with data, which is the most fundamental meaning of it.

Disclosure statement

The author declares no conflict of interest.

References

[1] Yu B, Yang R, Zuo J, et al., 2026, Impact Risk Analysis and Pressure Relief Evaluation of Deep Rockburst Mines Based on Multi-factor Data Fusion and Combination Empowerment. Rock Mechanics and Rock Engineering, 59(4): 4921–4944. https://doi.org/10.1007/s00603-025-05097-0

[2] Sayn H, Muhammed Zeynel ztürk, Zorlu K, 2026, Morphological Characteristics and Geoheritage Value of the Dereii Travertine Terraces (Van, Türkiye). Geoheritage, 2026(18): 2. https://doi.org/10.1007/s12371-026-01318-2

[3] Wang J, Cai Z, Chen X, et al., 2026, Study on Extreme Wave Load Distribution Characteristics and Parameter Sensitivity of Cross-Sea T-Beam. Periodical of Ocean University of China, 56(4): 124–137. https://doi.org/10.16441/j.cnki.hdxb.20240380

[4] Chen Y, Yang Z, Zhai S, et al., 2026, Research on Transport AC Loss Characteristics of Bent Conductor on Round Core Cable. Energies (19961073), 2026(19): 3. https://doi.org/10.3390/en19030841

[5] Jie L, Kecheng L, Shihuang Y, et al., 2026, Safety Profile and Signal Detection of Tadalafil: A Real-world Analysis Based on the Food and Drug Adverse Event Reporting System. Sexual Medicine, 2026(2): 2. https://doi.org/10.1093/sexmed/qfag014

[6] Myers AY, Colello MJ, Wren TAL, et al., 2026, Delays in Pediatric Supracondylar Humerus Fracture Management: A Comparison of Pre- and Postpandemic Trends. Journal of the American Academy of Orthopaedic Surgeons, 34(7): e1057–e1064. https://doi.org/10.5435/JAAOS-D-25-01173

[7] Xu W, Li W, Chen G, et al., 2026, Energy-efficiency Optimization and Internal Flow Characteristics with a Variable Outlet Width Strategy in Large-scale Centrifugal Impeller System. Energy, 2026(350): 140669.

[8] Wan F, Dai Z, Liu W, et al., 2026, Integrating the Belief Rule Base and Fuzzy Comprehensive Evaluation Health Assessment method to Enhance the Operational Efficiency of Rapier Loom Maintenance. Engineering Research Express, 8(6): 065526. https://doi.org/10.1088/2631-8695/ae4852

[9] Wentzcovitch R, Cobden L, Houser C, et al., 2026, A Quiet Quantum Revolution in Earth's Deep Interior. arXiv.

[10] Su Y, 2025, Accurate Marketing Algorithm of Network Video Based on User Big Data Analysis. Mathematical Problems in Engineering: Theory, Methods and Applications, 2022(1): 3317234. https://doi.org/10.1155/2022/3317234

[11] Zhang Z, 2025, Analysis of Upper Paleozoic Trace Element Characteristics and Paleoenvironmental Significance in the Gubei Area of Jiyang Depression. EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-4295. https://doi.org/10.5194/egusphere-egu24-4295

[12] Shahnawaz M, Kumar M, 2025, A Comprehensive Survey on Big Data Analytics: Characteristics, Tools and Techniques. ACM Computing Surveys, 2025(57): 8. https://doi.org/10.1145/3718364

[13] Cao L, Song J, Li X, et al., 2015, Geochemical Characteristics of Soil C, N, P, and Their Stoichiometrical Significance in the Coastal Wetlands of Laizhou Bay, Bohai Sea. Clean Soil Air Water: A Journal of Sustainability & Environmental Safety, 2015(4)3: 260–270.

[14] Khedr L, Halim A, Emara A, et al., 2026, Gender Differences in Biochemical Characteristics and Health-related Quality of Life among Hemodialysis Patients. BMC Nephrology, 27(1): 1–9. https://doi.org/10.1186/s12882-026-04867-4

[15] Sporek P, Konieczny M, 2025, Artificial Intelligence Versus Human Analysis: Interpreting Data in Elderly Fat Reduction Study. Advances in Integrative Medicine, 12(1): 13–18. https://doi.org/10.1016/j.aimed.2024.12.011