Scientific Big Data—The Cornerstone of the National Big Data Strategy (Postprint)

Authors: Guo Huadong

Date: 2023-03-19T00:00:00+00:00

Abstract

As a novel strategic resource for humanity, big data has emerged as a strategic high ground in the knowledge economy era. Characterized by a new paradigm that minimally relies on causal relationships and primarily leverages data correlations for knowledge discovery, it represents a quintessential exemplar of the data-intensive scientific paradigm succeeding empirical, theoretical, and computational paradigms, thereby catalyzing a transformation in research methodology and emerging as a new engine for scientific discovery. As a critical branch of big data, scientific big data exhibits intrinsic features including non-repeatability, high uncertainty, high dimensionality, and high complexity in computational analysis, alongside extrinsic characteristics concerning data content, data volume, data acquisition, and data analysis, which collectively pose novel challenges to the processing technologies and methodologies for scientific big data. Building upon the aforementioned analysis, this paper proposes recommendations for the scientific cognition of scientific big data, the construction of scientific big data infrastructure, the establishment of scientific data research centers, and the development of academic platforms for scientific big data.

Full Text

Scientific Big Data—A Footstone of National Strategy for Big Data

Abstract: Big data occupies the strategic high ground in the era of knowledge economies and also constitutes a new national and global strategic resource. It is a new pattern for scientific discovery with less dependence on causality and heavy dependence on data correlation. It has become a data-intensive scientific paradigm, following previous paradigms of empirical, theoretical and computational science. The paradigm has shifted the methodology of scientific research from theories and models based on causal analysis to comprehensive mechanistic

scientific discovery including correlation analysis. As a branch of big data, scientific big data includes internal characteristics such as non-repeatability, high uncertainty, high dimensionality, and computational complexity. External characteristics include data type, data volume, data acquisition, and data analysis. All these characteristics bring new challenges for the techniques and methods of processing scientific big data. On the basis of the above analysis, we raise four recommendations: scientific cognition of scientific big data, construction of scientific big data infrastructure, establishment of a scientific data research center, and the structuring of a scientific big data academic platform.

Keywords: big data, scientific big data, big earth data, data-intensive science

GUO Huadong Professor of Institute of Remote Sensing and Digital Earth (RADI), Chinese Academy of Sciences (CAS), Academician of CAS, Foreign Member of the Russian Academy of Sciences, Foreign Member of the Finnish Society of Sciences and Letters, and Fellow of the World Academy of Sciences for the advancement of science in developing countries (TWAS). He presently serves as President of the International Society for Digital Earth (ISDE), Member of UN 10-Member Group to Support the Technology Facilitation Mechanism, Chairman of the International Committee on Remote Sensing of Environment (ICORSE), Director of the International Centre on Space Technologies for Natural and Cultural Heritage (HIST) under the Auspices of UNESCO, Chair of Science Committee of Digital Belt and Road Program (DBAR), Editor-in-Chief of the International Journal of Digital Earth and Big Earth Data. He served as President of ICSU Committee on Data for Science and Technology (CODATA). He specializes in remote sensing science and its applications, and has a series of achievements in remote sensing information mechanisms, radar for Earth observation, and Digital Earth science. He has published more than 600 papers and sixteen books, and is the principal awardee of sixteen domestic and international prizes. E-mail: hdguo@radi.ac.cn

General Overview

As a novel strategic resource for humanity, big data has become a strategic high ground in the era of the knowledge economy. Its new pattern of knowledge discovery, which relies less on causality and more on data correlation, represents a typical data-intensive scientific paradigm following empirical, theoretical, and computational paradigms. This paradigm shift has transformed research methodology and become a new engine for scientific discovery. As an important branch of big data, scientific big data possesses internal characteristics such as non-repeatability, high uncertainty, high dimensionality, and high computational complexity in analysis, as well as external characteristics in data content, volume, acquisition, and analysis. These characteristics pose new challenges for processing techniques and methods. Based on this analysis, this article proposes four recommendations: scientific cognition of scientific big data, construction of scientific big data infrastructure, establishment of scientific data research centers, and building of scientific big data academic platforms.

The Booming Development of Big Data

On July 17, 2013, General Secretary Xi Jinping pointed out: "The vast ocean of data is like oil resources in industrial society, containing enormous productive forces and business opportunities. Whoever masters big data technology masters the resources and initiative for development." Big data has become a manifestation of information sovereignty and another arena for major power competition, following border, maritime, and air defense [1].

The volume of data doubles every three years. The current convergence of a new round of information technology revolution with human social activities has led to an explosion of semi-structured and unstructured data, whose generation is no longer constrained by time or space, exceeding the capacity of traditional data management systems and processing models [2]. Humanity is embarking on a new journey in the big data era. According to IDC's 2017 Big Data White Paper, global data scale will reach 163 ZB by 2025, ten times that of 2016, demonstrating robust growth [3]. China holds a significant position internationally in data volume, accounting for 13% globally by 2012, and is projected to produce 20% of worldwide data by 2020 [4].

Search trend data clearly shows global attention to big data in recent years. International interest remained low before 2012, grew rapidly from 2012-2015, and has maintained near-peak attention since 2016. The second industrial revolution caused text-based data to double approximately every ten years; entering the information age from the industrial era, data volume growth accelerated further. Big data is changing human life and deep understanding of the world.

Figure 1 [Figure 1: see original paper] Global data volume growth from 2016-2025 [3]

Big data's influence has reached all research fields in natural sciences, social sciences, humanities, and engineering, with big data research centers established across different domains [6]. China has deployed a series of big data science and technology projects, established laboratories for different research directions, and the Chinese Academy of Sciences launched the "Scientific Big Data Engineering" initiative. Research based on the big data value chain advocates innovation mechanisms driven by big data.

Understanding Scientific Big Data

Scientific big data exhibits characteristics of the data-intensive paradigm, including non-repeatability, high uncertainty, high dimensionality, and high computational complexity in analysis [7]. Using correlations among massive data can replace causality and theoretical models, enabling new knowledge and discoveries through data correlation [8]. For instance, in 1609, Johannes Kepler, assistant to Tycho Brahe, discovered planetary motion laws from Brahe's systematic observations, publishing the seminal work *New Astronomy*. Similarly, the Large Hadron Collider helped physicists test hypotheses about particle physics

and confirmed the existence of the Higgs boson. Big data has also made scientific discoveries in genomics possible, and spatiotemporal big data is playing a major role in global environmental change research [9].

Large scientific facilities are producing massive amounts of data, making the relationship between scientific big data and large facilities increasingly close. Earth observation satellites, large telescopes, the Large Hadron Collider, high-throughput scientific instruments, sensor networks, and other major facilities have successfully generated enormous datasets. In recent years, China's major facilities such as the 500-meter Aperture Spherical Telescope (FAST) and space science satellite series have provided powerful foundations for understanding nature through scientific big data. To meet massive and rapidly growing application demands, there is an urgent need to establish open systems that can share data, algorithms, and models to enable scientific analysis and integrated applications of existing data. A typical example is the European Space Agency's Sentinel-5P satellite, launched in October 2017, which acquires nearly 20 million observations of air pollutants and gases daily—more than ten times the data volume of previous missions. At current processing speeds, a single computer would require 1,200 years to process 3 million scenes of global satellite imagery, whereas cloud computing facilities can complete the same task in 45 days, demonstrating the critical importance of major infrastructure [10].

Realizing the great value of scientific big data still faces a series of technical challenges, which are reflected in five main aspects: (1) Data storage and management: The inherent characteristics of scientific big data urgently require databases capable of efficiently storing and managing massive, unstructured or semi-structured data. (2) Data analysis methods: The separation between data generation and analysis processes increases data noise, gradually replacing problem-driven research with data-driven approaches. (3) Models and algorithms: As semi-structured and unstructured data proportions increase, feature learning methods for such data are gradually surpassing and replacing traditional data models and algorithms. (4) Computing architecture: The continuous emergence of new storage and computing devices is transitioning general-purpose processors and single-architecture standalone systems to specialized processors, multi-core systems, and distributed large-scale heterogeneous clusters. (5) Computing and services: Cloud computing models and distributed high-performance data centers mediated by the internet are becoming new paradigms for big data processing [2].

The Chinese Academy of Sciences is conducting practical research on scientific big data. For example, the Strategic Priority Research Program (Category A) "Earth Big Data Science Engineering" is currently underway. Earth big data is a typical form of scientific big data—Earth science big data with spatial attributes. This program aims to break through technical bottlenecks in ultra-large-scale, cross-domain distributed resources, effectively promote Earth big data technology innovation, multi-spatiotemporal data management and association fusion, and problem-oriented data mining and analysis. Its goal is to

enable anyone, anywhere, with a terminal and internet access, to enjoy diverse services provided by Earth big data, achieving major scientific discoveries and one-stop comprehensive macro decision-making support services [11].

Another example is international scientific programs based on scientific big data. The "Digital Belt and Road" (DBAR) international program, initiated in 2016, aims to achieve big data aggregation, services, analysis, and presentation support to form a "Belt and Road" scientific big data platform. This ten-year scientific program will provide scientific decision-making support for sustainable development, food security, ecological environmental protection, climate change monitoring, disaster risk response, and cultural-natural heritage protection and development along the Belt and Road [12].

Reflections on Scientific Big Data

With the accumulation of data and improvement of computing capabilities, obtaining knowledge directly from big data has become possible. In September 2013, the author and his team proposed the concept of "scientific big data," which was published in *Chinese Science Bulletin* in January 2014 under the title "Scientific Big Data and Digital Earth." We believe that scientific big data differs essentially from internet big data and commercial big data in attributes and characteristics, possessing unique scientific connotations and features [9].

Overall, scientific big data has the following external features: In terms of data content, it generally represents natural objective objects and change processes; in terms of data volume, there are significant differences across disciplines; in terms of growth rate, it varies considerably by discipline; in terms of acquisition methods, it generally comes from observation and experimental records and subsequent processing; in terms of analysis methods, knowledge discovery from scientific big data generally requires scientific principle models.

Through summarizing these external features, the internal features become relatively clear, mainly summarized as: (1) Non-repeatability of data content: As the philosopher Heraclitus said, "No man ever steps in the same river twice," observations of natural and physical objective processes have a certain non-repeatability. (2) High uncertainty of data: Due to direct or indirect observation methods, sampling techniques, and recording technologies, systematic observation errors and data recording errors are often introduced. (3) High dimensionality of data: Due to the temporal and spatial attributes of observation objects and sampling methods themselves, as well as multi-channel characteristics of observation sensors, scientific big data often has spatiotemporal continuity and spectral multidimensionality, leading to the curse of dimensionality. (4) High computational complexity of data analysis: The high uncertainty and dimensionality of data, combined with the complexity of principle models accompanying scientific data analysis, result in computational complexity in scientific data processing and analysis. In summary, scientific big data has characteristics different from general big data, and its internal mechanisms and application to

knowledge discovery require in-depth research [7].

In June 2014, under our initiative and chairmanship, the "International Scientific Program Big Data Workshop: Challenges and Opportunities" was held in Beijing. Hosted by the Committee on Data for Science and Technology (CODATA) and co-hosted by seven international organizations, the conference issued a declaration emphasizing that scientific research should strengthen understanding of big data and enhance international big data science cooperation through developing research, policies, and frameworks related to big data to promote social development. Although this was only a starting point at the time, the declaration represented a substantive step toward recognizing big data's potential. Key points included: responding to the importance of big data for international scientific programs; developing big data's potential to serve society; enhancing understanding of big data through international cooperation; promoting big data accessibility through global research infrastructure; exploring and addressing challenges in big data management; encouraging capacity building in big data science; and promoting policy development to maximize big data utilization.

Since then, we have organized or co-organized a series of conferences on scientific big data, including the "Xiangshan Science Conference on Frontiers of Scientific Big Data," "CAS Academic Division Forum on Frontiers of Spatial Earth Big Data Science and Technology," "Roundtable Conference on Frontiers of Natural Science and Humanities Big Data," and "Xiangshan Science Conference on Earth Big Data." Relevant departments and organizations have also held various meetings related to scientific big data for in-depth discussions.

Particularly important is that under the organization of the Chinese Academy of Sciences, we proposed developing "scientific big data," which received government attention after submission. In 2015, the State Council's "Notice on Issuing the Action Outline for Promoting Big Data Development" included scientific big data as part of the outline, proposing to "develop scientific big data: actively promote the gradual opening and sharing of scientific data obtained and generated by public welfare scientific research activities supported by national public finance, build major national infrastructure for scientific big data, achieve authoritative aggregation, long-term preservation, integrated management, and comprehensive sharing of important national scientific and technological data, and develop scientific big data application service centers oriented toward economic and social development needs to support solutions to major issues in economic and social development and national security" [13].

Recommendations for Developing Scientific Big Data

With big data flourishing globally and China implementing its national big data strategy, scientific big data has become an important component of the national big data strategy. Against the backdrop of General Secretary Xi Jinping's higher requirements for implementing the national big data strategy, the State Coun-

cil issued the "Scientific Data Management Measures" in March 2018. We have welcomed an important historical opportunity for developing scientific big data. To better promote the development of scientific big data, we offer four recommendations.

(1) **Scientifically understand scientific big data in the big data world.** Scientific big data in the big data world has unique characteristics. It provides innovative research methodology, serves as a new engine driving scientific discovery, represents a frontier field for occupying future scientific high ground, offers a completely new way of thinking for human understanding of the world, and cultivates a new breed of scientists. Currently, China ranks first globally in computer users, internet users, and mobile internet users. China's data volume may reach 20% of the global total in the coming years, and China's published big data papers currently rank second internationally. The Chinese government attaches great importance to big data, and China's big data has relatively high international discourse power, laying a solid foundation for scientific big data research to reach international frontiers.

(2) **Construct major national infrastructure for scientific big data.** Major facilities produce big data, big data breeds big science, and big science drives big discoveries. National unified planning and construction of major infrastructure for scientific big data is of great significance. This includes ensuring acquisition and updating of scientific big data, authoritative aggregation and efficient processing, and achieving long-term preservation and integrated management of important scientific and technological data. Meanwhile, massive scientific data generated during research activities need to be shared with scientists via networks for analysis and processing. However, under current network information security environments and conditions, sharing and transmission processes for massive data result in low efficiency of scientific data transfer, affecting the quality of scientific discovery. Core technologies for collection, storage, maintenance, management, analysis, and sharing of scientific big data require support from major infrastructure.

(3) **Establish national scientific big data research centers.** China currently has dozens of major scientific facilities, hundreds of national key laboratories, and numerous departmental key laboratories, and is building national laboratories. These should be the primary targets for scientific big data "initiatives." Establishing scientific big data centers to serve research institutions in different fields could involve setting up domain-specific centers such as life big data centers, Earth big data centers, and astronomy big data centers, developing corresponding disciplines like bioinformatics, geoinformatics, and astroinformatics. Regional scientific big data centers could also be established. Considering the national positioning of the Chinese Academy of Sciences, we recommend establishing the National Scientific Big Data Research Center based on CAS. Meanwhile, a key fac-

tor for successful development of scientific big data is data sharing, which should include attention to scientific data publishing as a new mechanism for data integration and open sharing.

(4) **Launch international forums and alliances for scientific big data.** Improving methodology, theoretical foundations, and technical research for scientific big data in practical applications, and conducting bilateral or multilateral international exchanges and cooperation are important pathways to enhance scientific big data research levels. International scientific forums are important platforms to guarantee these implementations, facilitating frontier theory discussions and strengthening collaboration with international scientific and technological organizations and international scientific programs to gather expertise from more fields and disciplines, maintaining an excellent international scientific cooperation environment. Meanwhile, establishing an international scientific big data alliance should be considered. For example, building a big data alliance oriented toward the Belt and Road Initiative, using scientific big data as a starting point to make big data an engine for Belt and Road construction, a peace messenger jointly built by countries, and a light illuminating the present and future.

References

1. Guo H D, Wang L Z, Chen F, et al. Scientific big data and digital earth. Science Bulletin, 2014, 59(35): 5066-5073.
2. Guo Huadong, Chen Runsheng, Xu Zhiwei, et al. Big data in natural science and humanities: A review of the 6th China-Germany Frontiers of Science Roundtable Symposium. Bulletin of Chinese Academy of Sciences, 2016, 31(6): 707-716.
3. Reinsel D, Gantz J, Rydning J. Data age 2025: The evolution of data to life-critical don't focus on big data. Framingham: IDC Analyze the Future, 2017.
4. Gantz J, Reinsel D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. Framingham: IDC Analyze the Future, 2014.
5. GRDI2020. Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures. [2018-08-16]. http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf.
6. Li Xuelong, Gong Haigang. Survey of big data systems. Scientia Sinica (Informationis), 2015, 45(1): 1-44.
7. Guo Huadong. Big data, big science, big discovery: A review of the International Conference on Big Data and Scientific Discovery. Bulletin

of Chinese Academy of Sciences, 2014, 29(4): 500-506.

8. Hey T, Tansley S, Tolle K. The fourth paradigm: Data-intensive scientific discovery. Washington DC: Microsoft Research, 2009.
9. Guo Huadong, Wang Lizhe, Chen Fang, et al. Scientific big data and digital Earth. Chinese Science Bulletin, 2014, 59(12): 1047-1054.
10. Guo H D. Steps to the digital Silk Road. Nature, 2018, 554: 25-27.
11. Guo H D. Big Earth data: A new frontier in Earth and information sciences. Big Earth Data, 2017, 1(1-2): 4-20.
12. Guo H D, Liu J, Qiu Y B, et al. The Digital Belt and Road program in support of regional sustainability. International Journal of Digital Earth, 2018, 11(7): 657-669.
13. State Council of the People's Republic of China. Notice on Issuing the Action Outline for Promoting Big Data Development. [2015-09-05]. http://www.gov.cn/zhengce/content/2015-09/05/content_{10137}.htm.

Acknowledgments: Comrade Liang Dong contributed significantly to this work, for which we express our gratitude.

Note: Figure translations are in progress. See original paper for figures.