Mathematical Geosciences (2023) 55:287–293
https://doi.org/10.1007/s11004-023-10054-0

EDITORIAL

Special Issue: Data-Driven Discovery in Geosciences: Opportunities and Challenges

Guoxiong Chen[1] · **Qiuming Cheng**[1,2] · **Steve Puetz**[3]

Published online: 27 March 2023

Abstract
With the rapid expansion in big data and artificial intelligence (AI), Earth sciences are undergoing unprecedented advances in data processing and interpretation techniques, as well as in facilitating data-driven discoveries of complex Earth systems. This special collection explores scientific research related to data-driven discoveries in geosciences and provides a timely presentation of progress in developments and/or applications of AI and big data approaches to multiple aspects of geosciences. These include geohazards monitoring, mineral resource exploration, and environmental assessments. We hope this collection will inspire researchers and will transform the work undertaken in the field of data-driven Earth science. While many challenges remain, including the formidable tasks of transforming the deluge of geoscience data into useable information and furthering knowledge via cutting-edge AI techniques, we envision that data-driven discovery will revolutionize conventional methods of observation, analysis, modeling, and prediction in geosciences, and will further advance scientific understanding of our complex Earth system.

Keywords Data-driven discovery · Big data · Artificial intelligence · Geosciences

1 Introduction

The global community is presently facing considerable challenges in obtaining natural resources (e.g., minerals, energy, water, foods, etc.) while risking unwanted environmental effects of extreme events, such as global warming, loss of biodiversity, and

✉ Guoxiong Chen
gxchen@cug.edu.cn

1 State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Wuhan 430074, China

2 State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Beijing 10083, China

3 Progressive Science Institute, Honolulu, HI 96814, USA

natural/anthropogenic hazards (Sorkhabi 2022). When confronted with these challenges, the sustainable development of our world will require a deeper understanding of how the Earth operates in order to better tackle or predict extreme events (which must be better understood in terms of both the present and in deep-time Earth) (Cheng 2022). Progress in science and technology is a primary driving force in societal development, as they help provide vital solutions to challenges. In the past decade, big data and artificial intelligence (AI) have significantly altered the lifestyles and overall prosperity of society, while also influencing a fourth scientific paradigm—data-intensive or data-driven science. This fourth paradigm follows the traditional paradigms of experimental, theoretical, and computational sciences (Bell et al. 2009; Hey et al. 2009). Driven by ever-expanding arrays of data and armed with digital technologies (e.g., AI, big data analysis, and supercomputing), geoscientists are advancing the traditional approaches of thinking in both industry and academia (Bergen et al. 2019; Sun et al. 2022).

Historically, geosciences have progressed using inductive, knowledge-driven (or theory-guided) models by first generating a hypothesis and then collecting evidence to prove or disprove these hypotheses (Agterberg 2020). In geosciences, knowledge-driven models rely on logical reasoning based on prior knowledge gained by geologists, such as plate tectonics, evolutionary theory, and mineral deposit models. However, constructing prior geoscience knowledge is subject to the paucity of (preserved or exposed) rocks and limited observations, which hinder inferences and knowledge discovery. Nonetheless, data-driven science, based on abduction with big data, offers an opportunity for discovering new knowledge through AI techniques (e.g., machine learning and knowledge graphs) without a specific hypothesis (or theory) in mind. The advantages of data-driven discovery include transforming human learning by itself into an integration of both human learning and machine learning, as well as providing answers to known questions and formulating unknown answers to unknown questions (Cheng and Zhao 2020).

In general, data-driven science consists of several basic activities, including data capture, data curation, and data analysis (Hey et al. 2009). Data capture is the basis of data-intense science. Over recent decades, the rapid development of remote and in situ sensing techniques and the subsequent swift deployment of these technologies have led to the explosive growth of geoscience data for both industry and academia. Simultaneously, researchers have accumulated vast amounts of engineering and scientific data; however, these legacy data, for example, might represent gold deposits yet to be mined. Data curation, including data cleaning, data aligning, and converting meta-data, aims to build a data life cycle and form the data basis for conducting data-driven discovery via AI techniques. Considerable efforts have already been dedicated to quantifying geosciences. Examples include Macrostrat, EarthChem, along with many other data portals and databases. Data analysis and mining are key features of the fourth paradigm of scientific research. These types of approaches are imperative in tackling data deluge through cutting-edge AI techniques, including machine learning, deep learning, and knowledge graphs.

Although the geoscience community has been slow in adopting AI and big data techniques (relative to other disciplines), data-driven discovery is gaining popularity amongst industry, government, and academia. This is reflected in the ever-increasing

number of programs initiated, conferences and workshops convened, and in the number of related published research papers. For example, the United States Geological Survey (USGS) proposed four innovation areas in their future work, including big data, critical mineral resources, ecological resources, and natural hazards. Furthermore, the USGS suggested several potential directions for developing data-driven geosciences (Bristol et al. 2012). In 2019, the Deep-Time Digital Earth (DDE) big science program was initiated by the International Union of Geological Sciences (IUGS) with the intention of reconstructing the co-evolution of life, geography, matter, and climate over the 4.6 billion years of Earth history using data-driven abductive discovery, as well as by identifying the spatiotemporal distribution of global mineral and energy resources via AI techniques (Cheng and Zhao 2020; Wang et al. 2021).

Over the past decade, the data-driven discovery paradigm has received considerable attention for solving a variety of geoscience questions and challenges (Sun et al. 2022). These solutions range from addressing fundamental questions of Earth science to technical bottlenecks in engineering, such as the exploration of mineral and oil/gas resources. As a practical application, data-driven techniques, such as machine learning and deep learning, have played crucial roles in improving data processing methods and approaches to interpreting data in the field of remote sensing (Zhu et al. 2017), applied geophysics (Wang and Chen 2021; Yu and Ma 2021), and mineral prospectivity modeling (Chen et al. 2022c; Cheng and Agterberg 1999). In addition to transforming industries, data-driven science is also beginning to play an important role in advancing scientific discoveries of complex Earth systems, such as earthquake forecasting (Mousavi and Beroza 2022), global climate change (Reichstein et al. 2019), planetary interior structure (Wilding et al. 2022), the evolution of mass, life, and climate in the early Earth (e.g., Chen et al. 2022b; Chiaradia 2014; Fan et al. 2020; Hazen 2014; Keller and Schoene 2012; Puetz et al. 2018), and the search for extraterrestrial life (Ma et al. 2023). As yet another example, a high-resolution history of Earth's atmospheric oxygenation was reconstructed using machine learning and big data from mafic igneous rocks for the past 4 billion years (Chen et al. 2022a).

The special collection on "Data-driven Discovery in Geosciences" gathers six research papers that showcase new developments and novel applications of data-driven AI techniques in multiple aspects of geosciences. In Sect. 2, we summarize the highlights of these papers in addressing the specific challenges in different domains when using data-driven AI techniques. In Sect. 3, we outline future challenges for facilitating data-driven Earth science and then speculate about possible directions for these advances.

2 Summary of Articles in This Special Issue

The paper entitled "Geographically Optimal Similarity" by Song (2022) develops a mathematical model of geographically optimal similarity (GOS) for accurate and reliable spatial prediction of geological variables (e.g., trace elements) based on the Third Law of Geography—namely, the geographical similarity principle, which describes the comprehensive degree of approximation of a geographical structure instead of alternative explicit relationships between variables. GOS employs a small number of

samples and then derives better spatial predictions compared to the traditional methods. An R package named "geosimilarity" was developed for GOS-based predictions and uncertainty assessments. This work demonstrates the potential for applying the GOS model to spatial predictions, such as geochemical mapping in environmental assessments and mineral exploration.

The paper entitled "Revealing Geochemical Patterns Associated with Mineralization Using t-Distributed Stochastic Neighbor Embedding and Random Forest" by Shi et al. (2022) focuses on mineral prospectivity modeling using both unsupervised and supervised learning algorithms. A hybrid model combining t-distributed stochastic neighbor embedding (t-SNE) and the random forest (RF) method addresses data redundancy and the curse of dimensionality in geochemical mapping for mineral exploration. The application to the exploration of gold deposits in the northwestern Hubei Province of China demonstrates that the hybrid model combining t-SNE and RF can identify geochemical anomalies associated with gold mineralization efficiently. The high agreement with known gold deposits suggests that the areas targeted by t-SNE + RF can guide future mineral exploration in this area of study.

The paper entitled "Robust Optimal Well Control using an Adaptive Multigrid Reinforcement Learning Framework" by Dixit and Elsheikh (2022) focuses on optimal control problems using cutting-edge data-driven deep learning techniques. An adaptive multigrid reinforcement learning (RL) framework was introduced to address the computational challenge of robust control policies for uncertain, partially observable well control attributes. RL-based control policies are initially learned using computationally efficient low-fidelity simulations with coarse grid discretization of the underlying partial differential equations. The proposed RL framework was demonstrated by using the state-of-the-art Proximal Policy Optimization algorithm. Its application to two cases of well control problems suggests significant gains in computational efficiency. The improved efficiency is estimated to be between 60 and 70% when compared to single fine-grid methods.

The paper entitled "Ensemble and Self-supervised Learning for Improved Classification of Seismic Signals from the Åknes Rockslope" by Lee et al. (2022) focuses on geohazard monitoring using data-driven deep learning techniques. The fast and reliable identification of seismic events and their classification provide crucial information for monitoring rock slopes and early warning systems for potential rock slides. In this paper, a classifier for seismic geophone data was built to distinguish between different types of microseismic events using deep convolutional neural networks. With ensemble learning, the classification accuracy has been improved in comparison to the aggregation for a form 1 single spectrogram. This work also demonstrates the value of applying self-supervised learning. This is particularly relevant for datasets with insufficient labeling.

The paper entitled "Random Noise Attenuation by Self-supervised Learning from Single Seismic DatRandom Noise Attenuation" by Wang et al. (2022) focuses on reflection seismic data denoising using deep learning algorithms in the field of oil/gas exploration. A dropout-based self-supervised (DSS) deep learning method was introduced for single seismic data random noise attenuation to address the challenges arising from limited clean labels (i.e., noise-free) when using supervised algorithms in practice. Compared to the traditional *f–x* deconvolution and deep image prior methods,

the DSS method achieves better denoising results for preserving details of synthetic seismic data and field data. Moreover, numerical experiments indicate that the DSS method is stable for seismic denoising and reduces the over-fitting phenomenon.

The paper entitled "Construction and Application of a Knowledge Graph for Iron Deposits Using Text Mining Analytics and Deep Learning Algorithm" by Qiu et al. (2023) explores one of the frontiers for applying AI techniques in geoscience, that is, building a knowledge graph for facilitating knowledge discovery. A deep learning model was introduced to automate the extraction of geological entity relations from ore deposits, while creating a prototype question–answer system (Q&A) for ore-forming circumstances. This approach establishes annotation specifications for iron ore deposit entity relationships and a human-annotated corpus of geological entities of iron ore deposits. The constructed geological knowledge graphs were applied to analyze the mineralization characteristics of the Daye iron deposits in China.

3 Outlook

This special collection showcases a variety of data-driven research and/or applications in geosciences including seismic data processing, mineral prospectivity modeling, environmental pollution assessments, and geohazard monitoring, by using many data-driven AI techniques, such as unsupervised, supervised, self-supervised, and reinforcement learning. While recent advances in big data and AI approaches offer wonderful new opportunities for accelerating scientific discoveries and predictions via abductive, data-driven models and techniques, we face unique challenges specific to the geoscience domain, in addition to common difficulties pertaining to data capture, storage, searching, sharing, and visualization. The first challenge arises from transforming complex geoscience data into usable information because of the heterogeneity of the multivariate data as well as complex patterns obscured in data. The second challenge stems from converting information into knowledge due to gaps between predictions and the current understanding. Geoscience big data can be a gold mine. Whether this gold mine can be discovered by geoscientists depends on how effectively we overcome these challenges. Simply put, tackling the above challenges calls for domain-specific mathematical (statistical) models, advanced machine learning algorithms capable of learning with limited, weak, or biased labels, as well as a combination of data-driven and knowledge-driven models (Karniadakis et al. 2021). Given that many AI techniques are deeply rooted in mathematical and computational models (De Iaco et al. 2022; Dramsch 2020), it is an important mission for mathematical geoscientists to seize the strategic opportunity of the ongoing data revolution and bridge the gap between geoscientific data and AI models to further promote the new paradigm of data-driven Earth science research. Overall, while the research and development of data-driven discovery in geosciences are still in their infancy, we envision this new science paradigm to play ever-greater roles in the future. Such roles include but are not limited to protecting our society from various geohazards (e.g., major earthquakes, explosive volcanos, and landslides), providing resources for future generations, tackling environmental degradation and climate change, and searching for habitable planets.

Acknowledgements GXC and QMC are grateful for the support from the National Natural Science Foundation of China (Nos. 41972305 and 42050103), UNESCO Chair Program, Deep-time Digital Earth (DDE) Big Science Program, and MOST Special Fund from the State Key Laboratory of Geological Processes and Mineral Resources (MSFGPMR2022-3).

References

Agterberg F (2020) Induction, deduction, and abduction. In: Daya-Sagar BS, Cheng Q, McKinley J, Agterberg F (eds) Encyclopedia of mathematical geosciences. Springer, Cham, pp 1–12. https://doi.org/10.1007/978-3-030-26050-7_159-1

Bell G, Hey T, Szalay A (2009) Beyond the data deluge. Science 323:1297–1298

Bergen KJ, Johnson PA, Maarten V, Beroza GC (2019) Machine learning for data-driven discovery in solid Earth geoscience. Science 363:eaau0323

Bristol RS, Euliss Jr NH, Booth NL, Burkardt N, Diffendorfer JE, Gesch DB, McCallum BE, Miller DM, Morman SA, Poore BS (2012) Science strategy for core science systems in the US Geological Survey, 2013–2023. US Geological Survey

Chen G, Cheng Q, Lyons TW, Shen J, Agterberg F, Huang N, Zhao M (2022a) Reconstructing Earth's atmospheric oxygenation history using machine learning. Nat Commun 13:5862

Chen G, Cheng Q, Peters SE, Spencer CJ, Zhao M (2022b) Feedback between surface and deep processes: insight from time series analysis of sedimentary record. Earth Planet Sci Lett 579:117352

Chen G, Huang N, Wu G, Luo L, Wang D, Cheng Q (2022c) Mineral prospectivity mapping based on wavelet neural network and Monte Carlo simulations in the Nanling W-Sn metallogenic province. Ore Geol Rev 143:104765

Cheng Q (2022) Quantitative simulation and prediction of extreme geological events. Sci China Earth Sci 65:1012–1029

Cheng Q, Agterberg F (1999) Fuzzy weights of evidence method and its application in mineral potential mapping. Nat Resour Res 8:27–35

Cheng Q, Zhao M (2020) A new international initiative for facilitating data-driven Earth science transformation. Geol Soc Lond Spec Publ 499:225–240

Chiaradia M (2014) Copper enrichment in arc magmas controlled by overriding plate thickness. Nat Geosci 7:43–46

De Iaco S, Hristopulos DT, Lin G (2022) Geostatistics and machine learning. Math Geosci 54:459–465

Dixit A, Elsheikh AH (2022) Robust optimal well control using an adaptive multigrid reinforcement learning framework. Math Geosci. https://doi.org/10.1007/s11004-022-10033-x

Dramsch JS (2020) 70 years of machine learning in geoscience in review. Adv Geophys 61:1–55

Fan J, Shen S, Erwin DH, Sadler PM, MacLeod N, Cheng Q, Hou X, Yang J, Wang X, Wang Y (2020) A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity. Science 367:272–277

Hazen RM (2014) Data-driven abductive discovery in mineralogy. Am Miner 99:2165–2170

Hey AJ, Tansley S, Tolle KM (2009) The fourth paradigm: data-intensive scientific discovery, vol 1. Microsoft Research, Redmond

Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L (2021) Physics-informed machine learning. Nat Rev Phys 3:422–440

Keller CB, Schoene B (2012) Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gyr ago. Nature 485:490–493

Lee D, Aune E, Langet N, Eidsvik J (2022) Ensemble and self-supervised learning for improved classification of seismic signals from the Åknes rockslope. Math Geosci. https://doi.org/10.1007/s11004-022-10037-7

Ma PX, Ng C, Rizk L, Croft S, Siemion AP, Brzycki B, Czech D, Drew J, Gajjar V, Hoang J (2023) A deep-learning search for technosignatures from 820 nearby stars. Nat Astron. https://doi.org/10.1038/s41550-022-01872-z

Mousavi SM, Beroza GC (2022) Deep-learning seismology. Science 377:eabm4470

Puetz SJ, Ganade CE, Zimmermann U, Borchardt G (2018) Statistical analyses of global U–Pb database 2017. Geosci Front 9:121–145

Qiu Q, Ma K, Lv H, Tao L, Xie Z (2023) Construction and application of a knowledge graph for iron deposits using text mining analytics and a deep learning algorithm. Math Geosci. https://doi.org/10.1007/s11004-023-10050-4

Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566:195–204

Shi Z, Zuo R, Xiong Y, Sun S, Zhou B (2022) Revealing geochemical patterns associated with mineralization using t-distributed stochastic neighbor embedding and random forest. Math Geosci. https://doi.org/10.1007/s11004-022-10024-y

Song Y (2022) Geographically optimal similarity. Math Geosci. https://doi.org/10.1007/s11004-022-10036-8

Sorkhabi R (2022) Geoscience: what remains to be discovered? Epis J Int Geosci 45:173–180

Sun Z, Sandoval L, Crystal-Ornelas R, Mousavi SM, Wang J, Lin C, Cristea N, Tong D, Carande WH, Ma X (2022) A review of earth artificial intelligence. Comput Geosci 159:105034

Wang D, Chen G (2021) Seismic stratum segmentation using an encoder–decoder convolutional neural network. Math Geosci 53:1355–1374

Wang C, Hazen RM, Cheng Q, Stephenson MH, Zhou C, Fox P, Shen S, Oberhänsli R, Hou Z, Ma X, Feng Z, Fan J, Ma C, Hu X, Luo B, Wang J (2021) The deep-time digital earth program: data-driven discovery in geosciences. Natl Sci Rev 8:nwab027

Wang X, Sui Y, Wang W, Ma J (2022) Random noise attenuation by self-supervised learning from single seismic data. Math Geosci. https://doi.org/10.1007/s11004-022-10032-y

Wilding JD, Zhu W, Ross ZE, Jackson JM (2022) The magmatic web beneath Hawai 'i. Science 379:462–468

Yu S, Ma J (2021) Deep learning for geophysics: current and future trends. Rev Geophy 59:e2021RG000742

Zhu XX, Tuia D, Mou L, Xia G-S, Zhang L, Xu F, Fraundorfer F (2017) Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci Remote Sens Mag 5:8–36