The Geological Society of America
Special Paper 558

GOLD OPEN ACCESS

Data science for geoscience: Recent progress and future trends from the perspective of a data life cycle

Xiaogang Ma*
Department of Computer Science, University of Idaho, 875 Perimeter Drive, MS 1010, Moscow, Idaho 83844-1010, USA

ABSTRACT

Data science is receiving increased attention in a variety of geoscience disciplines and applications. Many successful data-driven geoscience discoveries have been reported recently, and the number of geoinformatics and data science sessions at many geoscience conferences has begun to increase. Across academia, industry, and government, there is strong interest in knowing more about current progress as well as the potential of data science for geoscience. To address that need, this paper provides a review from the perspective of a data life cycle. The key steps in the data life cycle include concept, collection, preprocessing, analysis, archive, distribution, discovery, and repurpose. Those subjects are intuitive and easy to follow even for geoscientists with very limited experience with cyberinfrastructure, statistics, and machine learning. The review includes two key parts. The first addresses the fundamental concepts and theoretical foundation of data science, and the second summarizes highlights and sharable experience from existing publications centered on each step in the data life cycle. At the end, a vision about the future trends of data science applications in geoscience is provided that includes discussion of open science, smart data, and the science of team science. We hope this review will be useful to data science practitioners in the geoscience community and will lead to more discussions on the best practices and future trends of data science for the geosciences.

1. INTRODUCTION

Data-driven discovery has received a lot of attention in geoscience research in the past decade, as reflected in the increasing number of projects funded, facilities constructed, data sets shared, and scientific findings published. Cyberinfrastructure, data portals, databases, workflow platforms, statistical models, machine learning algorithms, data management, and data sharing are becoming the new normal in many geoscientists' daily work. Various success stories of data-driven geoscience discovery in recent years have demonstrated the enormous potential of the data revolution. It is obvious that to scale up the innovation and accelerate new findings in geoscience, data science will play an important role in the coming decades. Nevertheless, as the theoretical foundation of data science is still under development, discussion and review of data science in the geosciences is limited. In contrast, data science methods and tools are currently in high demand among geoscientists. To address that need, this paper reviews progress in both data science and data-driven geoscience and discusses the future trends.

*max@uidaho.edu

Ma, X., 2022, Data science for geoscience: Recent progress and future trends from the perspective of a data life cycle, *in* Ma, X., Mookerjee, M., Hsu, L., and Hills, D., eds., Recent Advancement in Geoinformatics and Data Science: Geological Society of America Special Paper 558, p. 57–69, https://doi.org/10.1130/2022.2558(05). © The Author. Gold Open Access: This chapter is published under the terms of the CC-BY license and is available open access on www.gsapubs.org.

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

Data science is the study of extracting value from data (Wing, 2019). A primary driving force of data science in geoscience is the fast growing volume, velocity, and variety of data, i.e., big data. Hey et al. (2009) stated that data exploration is the key feature of "the fourth paradigm" in science for tackling the data deluge, as compared with the previous three scientific paradigms in which empirical, theoretical, and computational approaches were the key features. There are several factors in their vision of this new paradigm. Big data are captured by instruments or generated by simulators. Advanced infrastructures are deployed to store and transmit data, along with data analysis software and knowledge systems. Scientists, with the support and assistance of those resources, will focus more on scientific discovery in the midstream to downstream of the data flow. Another point raised by Hey et al. (2009) is that data-intensive science in the fourth paradigm is not only computational science but should also incorporate theories and methods from many other disciplines. Many later publications (Drineas and Huo, 2016; Kelleher and Tierney, 2018; NASEM, 2018a) resonate with Hey et al.'s (2009) vision of the theoretical foundation of data science. It is now commonly understood that data science will set its root in the basic research of computer science, mathematics, statistics, information science, and other disciplines. Successful data-driven scientific discovery also requires an open cyberinfrastructure and innovative pathways to enable the synergy of data science methods and domain-specific research questions.

Researchers of geoinformatics and geomathematics have also reviewed and discussed the evolution of information technologies in their work. Merriam (2004) listed six stages for the history of quantitative geology: origins (1650–1833), formative (1833–1895), exploration (1895–1941), development (1941–1958), automated (1958–1982), and integration (1982). Ma (2018) added that since the early 2010s, geoinformatics has been in the intelligent stage. Recently, there have been several review articles summarizing the latest trends of different aspects of data science in geoscience. Chan et al. (2016) and Shipley and Tikoff (2019) analyzed the changes that open data and cyberinfrastructure can bring to the workflow of geoscience, such as sedimentary geology and structural geology. Gil et al. (2019) analyzed the characteristics of research challenges in geoscience and then proposed a roadmap for developing and deploying knowledge-rich intelligent systems to address those challenges. In Karpatne et al. (2019), Bergen et al. (2019), and Reichstein et al. (2019), the challenges and opportunities of machine learning and deep learning for geoscience were thoroughly reviewed. Each of those three articles also has its own highlights. Karpatne et al. (2019) pointed out the synergistic advancement that such applications can bring to both machine learning and geoscience. Bergen et al. (2019) analyzed the larger function space and data-processing capability of machine learning in comparison to the conventional approaches in geoscience. Reichstein et al. (2019) asserted that data-driven machine learning should be coupled with the spatial and temporal context to obtain better understanding of Earth system processes and thus to improve prediction.

The quick progress of big data and data science has inspired plans and schemes for data-driven geoscience research at a larger scale. In 2018, the Carnegie Institution for Science started the Deep-time Data Driven Discovery (4D) Initiative (4D Initiative, 2018). In 2019, the International Union of Geological Sciences initiated the Deep-Time Digital Earth (DDE) big science program (Cheng et al., 2020). In the vision (NASEM, 2020) for the next decade of Earth science priorities within the U.S. National Science Foundation (NSF), key recommendations were made regarding open data and community practices for cyberinfrastructure needs and advances. We are now at a dramatic tipping point in science—a time when the open data resources, cyberinfrastructure facilities, and new data science methods for analysis and visualization will change the way geoscientists conduct their research. Keys to discovery lie in the continued development, integration, and exploitation of facilities, data, and expertise to build and explore pathways for a deeper understanding of the evolving Earth (Hazen et al., 2019). The review and analysis presented in this paper aim to answer questions such as, "What changes can data science bring to geoscience?" "What are the fundamental data science skills that a geoscientist should learn?" "What will be the patterns of data science applications in the next five or ten years?" and "As a student of geoscience, how can I quickly learn the data science methods and use them in my work?"

The perspective of this paper is from the point of view of a data life cycle. The data life cycle includes key steps such as concept, data collection, preprocessing, archive, distribution, discovery, analysis, and repurposing. The theme of each step is intuitive and easy to follow. Through this structure, this article summarizes sharable experience from existing studies with regards to data science workflows in geoscience. In the writing, the author has tried to present a comprehensive list and review of existing publications; however, the analysis presented may not cover all of the highlights of the cited publications. The remainder of the paper is organized as follows. Section 2 summarizes key concepts in data science. Section 3 reviews a number of recent publications on each step of a data life cycle. Section 4 analyzes the trends of data science in geoscience, and Section 5 offers a conclusion.

2. THE SCIENCE OF DATA SCIENCE

To better understand the workflows in data science, it is necessary to know a few fundamental concepts. The author has taught database and data science classes for senior undergraduate and graduate students in recent years. The experience has shown that even students majoring in computer science may confuse the meanings of data, metadata, information, and knowledge. Data are the recorded representation of facts. In the current digital era, the records are normally presented in a digital form, such as plain text, spreadsheet, relational database, and graph database. In addition to a hard disk, data can also be recorded on other types of media, such as paper and tape. Archived records from the old days, such as literature printed on hardcopies, can be digitized. Metadata are data about data. Metadata are important in data sharing and reuse

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

because they give an overview of the background of the data. An end user can get a quick summary and understanding of a piece of data just by reading the metadata. Structured metadata can improve the performance of search engines and enable them to accurately index records and find the best match for a request. Information is the meaning or message extracted from data. The information extraction process often depends on the purpose of data analysis, the methods and tools used, and the interpretation of data analysis results. It is not strange to see the same piece of data used in studies of different topics to generate varied information. Knowledge is the expertise and familiarity with a topic. In traditional understanding, a human can attain knowledge by learning, practice, and experience. In data science, there are now knowledge bases that can save knowledge in quantitative and qualitative formats, which can in turn be used in the data analysis process. The three concepts of data, information, and knowledge are also used in combination with other concepts, such as wisdom and action, to form a pyramid or flowchart and depict the ability of using knowledge and insight gained from data to think and act in real-world practices (Fig. 1A).

Many researchers and communities have depicted the data life cycle and the data science process. Figure 1 presents a number of diagrams from the existing publications (Chapman et al., 2000; Schutt and O'Neil, 2013; Berman et al., 2018; Wing, 2019; DDI Alliance, 2021). Most are easy to read and understand, and we will omit the detailed description for each of them. Nevertheless, some shared topics in those diagrams are worthy of highlighting. For instance, the data life cycles presented in Figures 1B and 1E both include the steps of data sharing, publication, and reuse. The step of data processing in Figures 1B and 1C actually means data cleansing, wrangling, and munging, which is similar to the step of data preprocessing in Figure 1F. In Figures 1D and 1F, the steps of visualization and interpretation address the needs of meaningful data science, i.e., to appropriately interpret the results of data analysis. This includes not only the precision and efficiency of algorithms but also the domain-specific meaning in the outputs of those algorithms. Also, the issues of data privacy and ethics have received more attention and discussion in recent publications to highlight data science as an ecosystem (Figs. 1D–1E).

Interdisciplinary collaboration led to the emergence and evolution of data science. Donoho (2017) offered a thorough review of data science's evolution over the past decades. In particular, he summarized the perspectives of several statisticians on

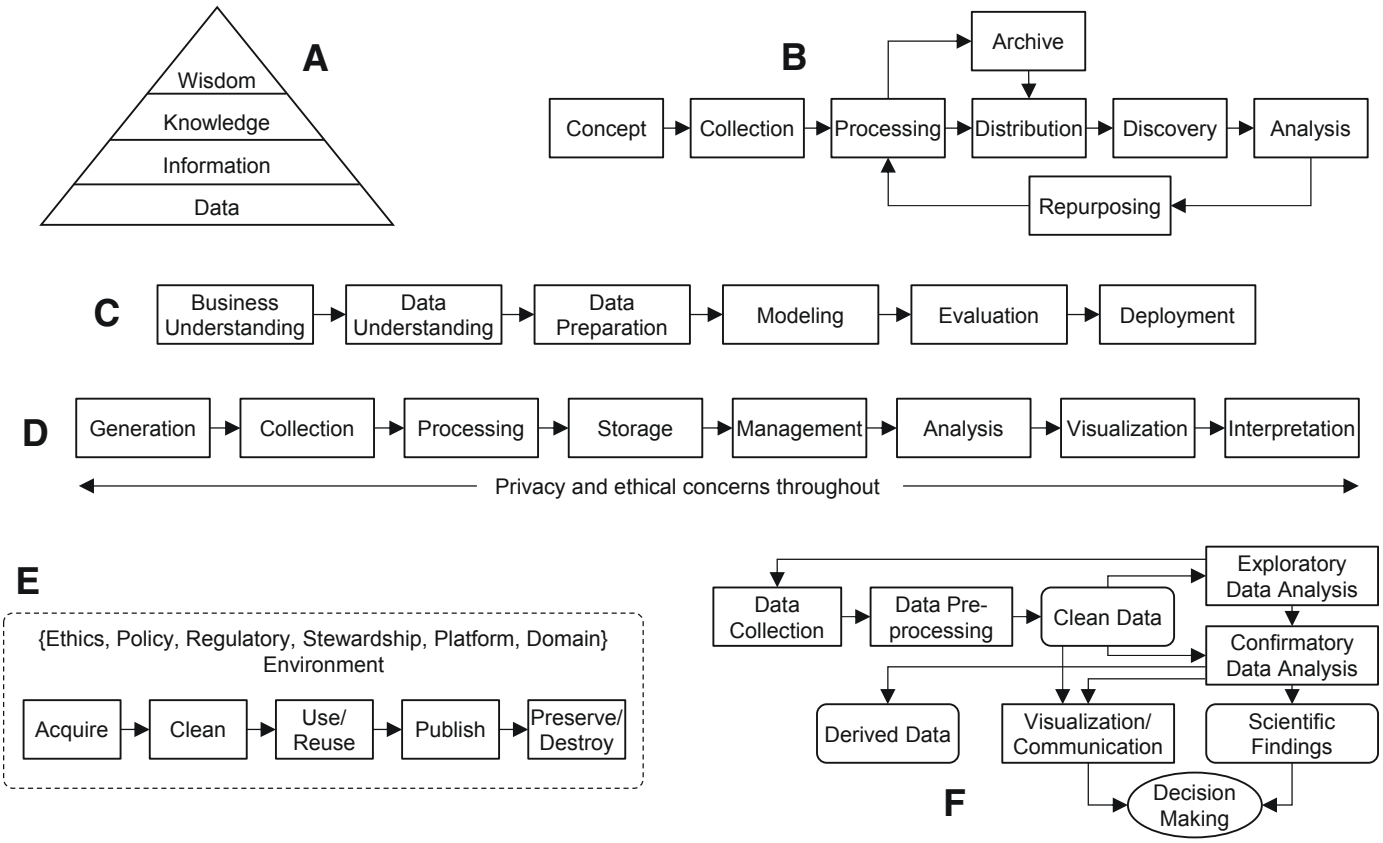

Figure 1. Different depictions of the data life cycle and the data science process are shown. (A) The DIKW model; (B) the Data Documentation Initiative (DDI) data life cycle (DDI Alliance, 2021); (C) the cross-industry standard process for data mining (CRISP-DM) (Chapman et al., 2000); (D) the data life cycle in data science (Wing, 2019); (E) the data life cycle and surrounding data ecosystem (Berman et al., 2018); and (F) the data science process (Schutt and O'Neil, 2013).

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

the need to expand the boundaries of classical statistics to cover topics of data preparation, presentation, and prediction. In the review it was mentioned that the term "Data Science" had been used two decades ago by Cleveland (2001) for the envisioned new field. Recent discussions have made clear that the field of data science should be interdisciplinary, including computer science, statistics, mathematics, information science, and progress in subject matter applications (Drineas and Huo, 2016; Kelleher and Tierney, 2018). Those discussions were reflected in the list of data science courses and the curricula of those courses. A recent National Academies of Sciences, Engineering, and Medicine report (NASEM, 2018a) stated that a critical task of data science education is to establish data acumen, which includes these key concepts: mathematical foundations, computational foundations, statistical foundations, data management and curation, data description and visualization, data modeling and assessment, workflow and reproducibility, communication and teamwork, domain-specific considerations, and ethical problem solving. Those topics of data acumen are reflected in the data life cycle and data science process (Fig. 1) to address the real-world needs of data science applications. Several universities already offer data science courses. For example, the University of California at Berkeley offers Data 8: Foundation of Data Science to entry-level undergraduates in any major (Adhikari and DeNero, 2017). Its curriculum covers most of the subjects in the above data acumen list.

Many geoscience and geoinformatics researchers have analyzed the science of data science from the perspective of their experiences with real-world practices. Mattmann (2013) discussed four advancements that are necessary to tackle the challenges of big data: algorithm integration, software development and stewardship, automated data format identification and reading, and the training of data scientists. Fox and Hendler (2014) addressed that the field of data science includes not only the disciplinary foundations but also strategies for real-world challenges. They provided details about four cross-cutting data science challenges: understanding scale in systems, sparse systems with incomplete and heterogeneous data, abductive reasoning, and next-generation semantic data infrastructure. Here, abduction reasoning is similar to the "Exploratory Data Analysis" proposed by Tukey (1977, p. v), "It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights." Ho (1994) summarized that abduction creates, deduction explicates, and induction verifies. This means that abduction is a good way of finding clues to scientific questions through the activities of data exploration. Hazen (2014), based on his experience of data-driven studies in mineralogy, further summarized that deduction and induction are to discover what we know we do not know while abduction is to discover what we do not know we do not know. Recognition of the data science myths pointed out by Kitchin (2014) and Kelleher and Tierney (2018) is important for avoiding unrealistic expectations. The myths are: (1) data science is an autonomous process without human oversight; (2) every data science project needs big data and machine learning; (3) data science software is easy to use, and data science is an easy job; and (4) data science pays for itself quickly. Awareness of those myths will help geoscientists understand the limitations of data science and be better prepared to problem solve in the real world.

3. A REFLECTION ON THE KEY STEPS OF A DATA LIFE CYCLE

Focusing on the theme of data science for geoscience, the following sub-sections review a list of recent publications for each key step in the data life cycle and summarize the shareable experiences from them.

3.1. Business Understanding and Concept

The steps labeled "concept" in Figure 1B and "business understanding" in Figure 1C are intended to determine the objectives of a data science project and estimate the data needs (Chapman et al., 2000; DDI Alliance, 2021). They are about turning business goals into data science plans. If the planned activities include database construction, this step will also include the work of developing data structures, such as a conceptual model, logical model, physical model, as well as controlled vocabularies for data standardization. Cyberinfrastructure researchers recognize that consideration and action regarding data semantics in the early stage will help improve data interoperability when data are generated, collected, integrated, and shared in a later stage (Reitsma et al., 2009; Narock and Shepherd, 2017).

The Semantic Web extends the World Wide Web by adding structures and meaning to terms in documents on the web (Berners-Lee et al., 2001). The key technical approach to enable the Semantic Web is the use of ontologies, which are formal specifications of a shared conceptualization of a domain (Gruber, 1995). Researchers have suggested a semantic spectrum that consists of a sequence of items such as catalog, glossary, taxonomy thesaurus, conceptual schema, and formal logical models, for constructing and implementing ontology in practice (Welty, 2002; McGuinness, 2003; Obrst, 2003; Uschold and Gruninger, 2004). The items in this spectrum provide a roadmap for increasing the semantic precision and interoperability of data in a variety of applications.

Data interoperability has received tremendous attention in recent years. The widely accepted FAIR (Findable, Accessible, Interoperable, and Reusable) data principles (Wilkinson et al., 2016; Stall et al., 2019) are closely related to the discussion of data interoperability in the past decades (Fig. 2). Several researchers presented the layered structure of data interoperability, including systems, syntax, schematics, semantics, and pragmatics (Bishr, 1998; Sheth, 1999; Ludäscher et al., 2003; Brodaric, 2007, 2018). A few other researchers explained those layers in layman's terms, including discoverable, accessible, decodable, understandable, and usable (Wood et al., 2010; Ma et al., 2011). The layered structures of data interoperability and

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

the FAIR principles can also be compared with the technical architecture of the Semantic Web (Berners-Lee, 2000). Many best practices of data interoperability can be seen in the domain of geoscience. The U.S. National Geologic Map Database of the U.S. Geological Survey (USGS) has adopted the North American Geologic Map Data Model (NADM) (NADM Steering Committee, 2004) as a common schema for coordinating state-level geologic map databases. Such efforts to determine standards are continuously active at USGS, such as the recently released Geologic Map Schema (GeMS) (USGS NCGMP, 2020). Similarly, NASA has implemented the Global Change Master Directory (GCMD) Keywords as a hierarchical set of controlled vocabularies to ensure the interoperability of its data and services (GCMD, 2020). In Europe, the INSPIRE Directive aims to create a European Union spatial data infrastructure (Bartha and Kocsis, 2011; Ma and Fox, 2014). Its data and metadata specifications cover 34 data themes in Earth and environmental sciences, with full implementation required by 2021 across all of the participating European nations. Scientific communities such as the World Wide Web Consortium and the Open Geospatial Consortium have also summarized best practices for publishing and serving data on the web (Loscio et al., 2017; Tandy et al., 2017).

3.2. Data Understanding, Generation, and Collection

Along with the quick development of hardware and software in the cyberinfrastructure, data are now generated at an ever increasing speed. Sensor networks (Martinez et al., 2004; Hart and Martinez, 2006) greatly facilitate the generation, transmission, and integration of Earth and environmental data. NASA organizes ~100 missions and thousands of platforms, instruments, and sensors around the Earth and nearby space and is one of the biggest geoscience data producers worldwide. It was reported (Shannon, 2019) that in 2016, NASA was already generating 12.1 TB of data every day. The same article also reported that NASA is deploying new sensors that alone will be able to generate 24 TB of data daily. Similar advances in instruments and facilities for data generation, transmission, and management were also seen in field-based geological surveys (Mookerjee et al., 2015). Wing (2019) made a distinction between data generation and collection and pointed out that not all data generated are collected (Fig. 1D). That may be because we only want to collect a certain part of the data or because the velocity of data streams is too high to be processed with existing tools.

Crowd-sourcing platforms, such as social media and community portals, are generating massive data. Many of our daily activities, such as posting on Twitter or Facebook, watching and commenting on a video on YouTube, and searching on Google, all generate digital records in a way in which many of us are even not aware. A great deal of social media data are used for scientific studies. For example, Twitter data were used for wildfire disaster management (Wang et al., 2016). Google search data were used to predict seasonal influenza trends (Carneiro and Mylonakis, 2009). The community collaboration on OpenStreetMap greatly helped the rescue work after the 2010 Haitian earthquake (Ahmouda et al., 2018). Images on Flickr were used for ecosystem assessment in remote areas (Rossi et al., 2020). Besides the public social media, another type of crowd-sourcing platform focuses on a certain subject and is normally maintained by a community of enthusiasts. For example, Mindat.org is such a community platform focused on mineral species. It has a small team of database administrators and data reviewers and is open to thousands of data contributors and users across the world. Researchers have used Mindat data in many recent studies on mineral evolution and mineral ecology (Hazen et al., 2011; Morrison et al., 2020).

The massive collection of geoscience literature is another good source of data. For example, GeoDeepDive (Zhang et al., 2013; Peters et al., 2014) is a machine learning package for discovering data and knowledge from published documents. By

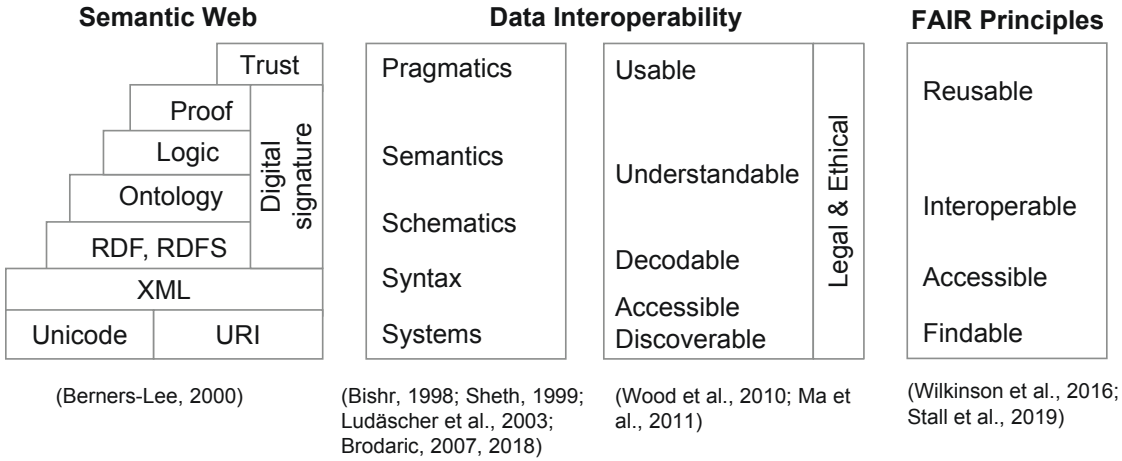

Figure 2. Comparison shows the layered structure of data interoperability with the Semantic Web architecture and the FAIR data principles (from Ma et al., 2020; CC BY 4.0 license). For sources of sub-diagrams, see description in text.

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

January 2021 it had preprocessed more than 13 million documents. Peters et al. (2014) successfully used the fossil records extracted from GeoDeepDive to enhance the Paleobiology Database. GeoDeepDive also allows other researchers to use the data resource to explore their own scientific topics. Recently, there have also been many studies on using text mining technologies to extract knowledge graphs from the geoscience literature (Wang et al., 2018; Qiu et al., 2019; Fan et al., 2020b).

3.3. Data Preprocessing and Preparation

Data preprocessing is an increasingly important step in data science. It is also referred to by several alternative names such as data cleansing, data wrangling, and data munging. The general purpose of data preprocessing is to ensure the quality of data before any data analysis is conducted. In real-world practice, it may involve tasks such as removing noisy and unreliable records, reducing data dimensionality, transforming data formats, selecting records of interest, enriching the existing data with additional attributes, and combing data from different sources to build a new piece of data (Wang et al., 2018). Many researchers (Press, 2016; Mons, 2018), including geoscientists (Fox, 2019), spend 80% of their time cleansing and preparing data before analyzing the data (i.e., the 80/20 rule). Good data preprocessing can significantly increase the efficiency of data analysis and lead to remarkable scientific discoveries. For example, the above-mentioned Mindat data portal was used as a source for the Mineral Evolution Database (Golden et al., 2019). Nevertheless, a limitation of the original Mindat is that it does not include an age attribution for a mineral species' first occurrence on Earth. Golden et al. (2019) searched over 1600 publications and several existing databases to extract such age data and then used them to enrich the Mineral Evolution Database. The updated database underpinned many new research discoveries, including mineral evolution and ecology (Morrison et al., 2019, 2020) and the co-evolution between the geosphere and the biosphere (Spielman and Moore, 2020). The database also led to new designs of mineral species databases and discussions on better methods for data curation and sharing (Prabhu et al., 2021).

Applying data standards to transform existing data or mediate between databases is also a widely used approach in data preprocessing and preparation. The above-mentioned metadata and data specifications in the INSPIRE Directive is a good use case for that approach. Another example is the global "OneGeology" project for improving the accessibility of geological maps on the Internet (Jackson, 2010). OneGeology has developed a tool kit to set up online geologic map services. More than 110 countries have participated in the project, and about half of them are serving map data to a web map portal. The original maps are heterogeneous because they are recorded in different formats and use different data models, terminology, and language. Through the OneGeology map service tool kit, the online services of those maps are made consistent, and they can be browsed in a centralized map window. The "OneGeology-Europe" project (Laxton, 2017) has utilized multilingual vocabularies to develop innovative capabilities for the online geologic maps of participating European nations. New functions of OneGeology-Europe include the multilingual user interface, federated queries across distributed geologic map services, consistency with other regional and international data standards, and more. As reflected in those examples, well-organized data preprocessing preparation can significantly change the 80/20 rule in data science activities or even reverse it.

3.4. Data Archive, Distribution, and Discovery

Nowadays, it is a new normal that funding agencies require researchers to include a data management plan in their grant proposals (Dietrich et al., 2012; NSF, 2015). Increasingly, data are treated as a formal research output and receive the same attention as paper publications. The FAIR data principles (Wilkinson et al., 2016) are now well received in almost all scientific disciplines, including geoscience (Stall et al., 2019; Lannom et al., 2020). The FAIR data principles represent many preceding efforts on data management and stewardship and represent a systematic approach to sharing and reusing scientific data in an open scientific environment. Those efforts include data infrastructure construction (Cutcher-Gershenfeld et al., 2016), persistent and resolvable identifiers for data publication (Klump et al., 2016), metadata standardization (Starr and Gastl, 2011), provenance documentation (Lebo et al., 2013), data citation (Parsons et al., 2010), and more. There are many general-purpose data portals where researchers can upload and share their data. Moreover, there are specific data portals that only focus on one or a few subjects, such as petrology, geochemistry, and geophysics. Data-producing agencies such as NASA, USGS, the National Oceanic and Atmospheric Administration (NOAA), and the U.S. Department of Agriculture (USDA) all have their own data archives and data portals that allow users to search and access data of interest. For instance, USGS enables federated query to a long list of mineral resource spatial databases through a central portal (USGS MRDATA, 2021). As workflow platforms such as Jupyter Notebook and R Markdown are increasingly used, many data portals have also developed packages to enable data access from workflow platforms, such as the paleobioDB R package for the Paleobiology Database (Varela et al., 2015) and the neotoma R package for the Neotoma Paleoecology Database (Goring et al., 2015).

The FAIR data principles prioritize findability. It is true that from the perspective of a user, data discovery is a key step if the user's work needs to access data on external databases or data portals. A top-down approach can be used to search records in data portals with specific themes, such as EarthChem (earthchem.org), PANGAEA (pangaea.de), Neotoma (neotomadb.org), PaleoBioDB (paleobiodb.org), and many data portals organized by the federal agencies. Moreover, there are also registries for metadata from multiple data portals, such as DataONE (dataone.org), as well as registries of data portals, such as RE3DATA

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

(re3data.org). On those data portals, a user can quickly narrow down the scope of a search by selecting disciplines, subjects, geospatial range, time span, and other attributes. Another approach of data discovery is the free-style search, such as those enabled by the Schema.org (Noy et al., 2019). By providing metadata through the Schema.org specifications, the records in a data portal will be made indexable to search engines. For example, Google has indexed millions of data sets on thousands of data portals and made them searchable through the Google Dataset Search engine (Noy et al., 2019). A user can search data sets with any combination of keywords. Once a data set is identified on the search engine, the user can access the data set through the web address provided in the metadata. Recently, there have also been discussions about dataguides, which are a type of computer-aided analysis that can inform researchers about what data to collect and where to find them (Shipley and Tikoff, 2019).

3.5. Data Analysis and Result Interpretation

Many people would simply think of data science just as data analysis. Indeed, data analysis is a key step in the data life cycle, but it is just a part of the process. In past decades, many studies focused on the theories and applications of statistical models and data mining in geoscience (Merriam, 2004; Sagar et al., 2018). In recent years, the fast-growing methods and technologies of big data (Yang et al., 2017, 2019), cloud computing (Li et al., 2015; He et al., 2019), machine learning (Lary et al., 2016; Bergen et al., 2019; Karpatne et al., 2019), and deep learning (Reichstein et al., 2019) have been widely used in geoscience with achievement of significant outcomes. Many innovative, data-driven discoveries were seen in paleobiology (Peters et al., 2017), paleontology (Fan et al., 2020a), mineralogy (Hystad et al., 2015, 2019), water resources (Wen et al., 2018; Sun and Scanlon, 2019), forest cover change (Hansen et al., 2013), and public health (Goovaerts, 2008, 2021). Data analysis often includes two steps: exploratory and confirmatory data analysis (Fig. 1F). This conventional statistical method can still be very useful for data science applications today. Exploratory data analysis is used to get a better understanding of the data and draw plausible research questions or hypotheses (Tukey, 1977; Camizuli and Carranza, 2018). Confirmatory data analysis, in contrast, is where the complicated models and/or algorithms are applied to prove or disprove the hypotheses.

Data visualization has been increasingly discussed as an efficient way to improve the understandability of a data science process and the interpretability of the data science results (Fox and Hendler, 2011; Ma et al., 2015; Wing, 2019). Data visualization not only means to make the information visible, but also that the visualization should make the information easy to perceive by a reader. Many may think of visualization just as a way to present data science results, but in actual practice, many data visualization techniques can also be used in data preprocessing and analysis. For example, box plot is a widely used visualization in exploratory data analysis. Ma et al. (2017) used a three-dimensional cube matrix to explore the co-relationship between elements and mineral species and generated new research questions for detailed analyses. Morrison et al. (2017) applied network analysis to visualize the patterns of co-existence of minerals. In Dutkiewicz et al. (2015), machine learning was used to generate new hypotheses based on the analysis of big seafloor sediment data. GPlates software (Müller et al., 2018) was used as a data visualization platform in that study, which generated impressive results. These examples show that data visualization is an efficient approach for facilitating collaboration among geoscientists, data scientists, mathematicians, and data managers and for making the data science process and results understandable to a broader audience.

3.6. Repurposing

Repurposing means that a piece of data can be reused in other projects either by external users or the data producers themselves. Data interoperability and reusability will be the focus in this step. The FAIR data principles as well as the open data and open science campaigns suggest that metadata should include the provenance information of the original research activities that generated the data (Di et al., 2013; Gil et al., 2016; Wilkinson et al., 2016; Zeng et al., 2019; Lehmann et al., 2020). According to best practice, besides sharing data, researchers should also document their software packages, workflow setup, and the context information that interconnects the entities, agents, and activities involved in a research program. Open data and open science are helping to change the culture of research and create a virtuous data ecosystem in geoscience (Sinha et al., 2010; Welle Donker and van Loenen, 2017; Caron, 2020). Many new scientific discoveries are based on research activities that use "other people's data." For example, the work of Muscente et al. (2018) on the ecological impacts of mass extinctions used fossil community data from the Paleobiology Database. The work of Keller and Schoene (2012) on disruption in secular lithospheric evolution and Keller et al. (2015) on volcanic–plutonic parity and continental crust both used data from EarthChem. The work of Hazen et al. (2019) on mineral evolution used data from Mindat and other open data resources. To promote a healthy open data ecosystem, legal and ethical issues are also discussed (Berman et al., 2018; Kelleher and Tierney, 2018; Wing, 2019).

4. FROM BIG DATA TO DATA SCIENCE ECOSYSTEM: A VISION FOR THE NEXT DECADE

Along with the evolution of data science theory and methodology, the upgrading of computational facilities and capabilities, the thriving of big data and open data in geoscience, and the training of geoscientists with data science skill sets, it is certain that data science will be applied more frequently in geoscience, which will lead to more scientific discoveries. What will be the trends in methodology and technology, and what should a geoscientist be aware of to be better prepared for the data revolution? This section offers a few thoughts.

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

4.1. Open Data and Open Science Will Be the New Normal

The concept of open science is being widely accepted in academia (Donoho, 2017; NASEM, 2018b; Aspesi and Brand, 2020). Open science is an umbrella concept for a long list of "open" activities, including open access to publications, open source software programs, open data, open samples, and open workflows, just to name a few. Many open science activities will take place on the internet and the web (Berendt et al., 2020). For example, data will become more open, accessible, and interactive through various protocols and interfaces, such as those maintained by the World Wide Web Consortium and the Open Geospatial Consortium. Facilitated by the FAIR principles and other associated efforts, the shared data will be better curated, which will save researchers time on data preprocessing and preparation. The USGS mineral resources spatial data portal is an example of that trend (USGS MRDATA, 2021). The Giovanni infrastructure of NASA (Acker and Leptoukh, 2007) has also been working toward cooperation among NASA's distributed data archives to enable federated data exploration and comparison (Lynnes, 2020). For reflection, a key idea in the vision of the Semantic Web (Berners-Lee et al., 2001) is the persistence and traceability of resources on the web. Similar to the digital object identifier (DOI) for publications, many other entities and agents in open science, such as data, software packages, samples, researchers, organizations, and research grants, will also have their persistent and resolvable identifiers on the web. By connecting those identifiers, we can easily weave a graph for all of the objects, steps, and workflows involved in generating a scientific finding.

Workflow platforms such as Jupyter Notebook, R Markdown, and others will be widely used in geoscience from research projects to classroom education. Those workflow platforms are not only good tools for collaborative and reproducible research activities, they also provide well-organized environments for students to learn and use programming languages. Many geoscience data portals now have Python or R packages to enable users to search and access data directly from a workflow, and there have been various successful applications in geoscience (Varela et al., 2015; Peters and McClennen, 2016; Choi et al., 2021; Rosenberg et al., 2020). We anticipate that workflow platforms will become more popular in geoscience in the future. Similar to the needs of computer scientists and data scientists for trustworthy artificial intelligence (Floridi, 2019; Wing, 2020), geoscientists also express the request for provenance in their workflows (Gil et al., 2019). Recently, packages have been developed in workflow platforms to capture provenance. For example, the MetaClip (Bedia et al., 2019) framework is able to capture the provenance description of a climate product and then append the provenance information inside the resulting image. Once that image is loaded to the MetaClip Web portal, the provenance information inside it will be read and visualized. To tackle large data sets, researchers have begun to deploy workflow platforms in the cloud environment (Hamman et al., 2018; Sun et al., 2020). This will be a trend in big geoscience data processing in the near future.

4.2. Big Data, Smart Data, Data Science, and the Changes They Bring to Geoscience

Big data does not mean we can dump and share data while simply relying on machine learning to identify patterns in the chaos. Many researchers have discussed the idea of smart data (Iafrate, 2014; Sheth, 2014; Maskey et al., 2020). That is, the application of metadata and semantics to add more machine-readable structures in data generation and collections and the deployment of intelligent algorithms to improve the precision of data discovery and analysis. Smart data will bring refreshing changes to the data life cycle and help researchers quickly identify the data to be used and extract value from the data. Many geoscience data portals, such as EarthChem, Neotoma, and the Paleobiology Database have already applied controlled vocabularies to improve the precision of data search and query. The Google Dataset Search engine, enabled by Schema.org, offers a playground for developing more innovative functions in data search. The geoscience community has already begun to work on approaches to expose Schema.org-compatible metadata on their data portals (Shepherd et al., 2019; Valentine et al., 2020) and make the metadata indexable to the Google Dataset Search engine. When more data portals enable such functions, an end user will be able to search a variety of data on the Google Dataset Search engine. Metadata portals for specific geoscience disciplines or subjects such as deep time (Stephenson et al., 2020) can also be built with those indexable metadata from various data portals. Those improved functionalities will greatly benefit end users (Chapman et al., 2020). With more provenance information about workflows documented and shared, smart search engines can be developed that use such information to provide recommendations not only on data, but also on software packages that can be used to analyze the data, potential research topics for the data, and researchers with whom to collaborate. For example, Mookerjee et al. (this volume) discussed that by using machine learning, data management systems will be able to make connections to other data sets that can potentially build collaborations or suggest other geographical areas to study.

The smart data will save researchers time on data discovery and allow them to put more efforts toward proposing research questions and conducting data analysis. This will be possible whether working with a small amount of data and identified research questions or a large amount of data that requires exploratory data analysis and hypothesis generation (Kitchin and Lauriault, 2015). Ma (2018) compared the data science process with conventional science approaches and pointed out that a unique feature of data science in the big data era is that while a lot of data are collected, we may not yet have formed a specific research question. Bergen et al. (2019) discussed that machine learning provides the means to discover high-dimensional and complex relationships in data and enables exploration of more scientific hypotheses. If the conventional approaches are small data and small knowledge (i.e., domain experts and personal computers), then the data science process can enable big data and big

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

knowledge (i.e., domain experts, smart data, machine learning, and cloud environment). In big data–enabled, multidisciplinary geoscience research projects, interpretability of the workflow will help people from different disciplinary backgrounds better understand the results and findings (Reichstein et al., 2019). This overlaps with the work on explainable and meaningful artificial intelligence in computer science (Hagras, 2018; Holzinger, 2018; Chari et al., 2020). In the geoscience community, there has been some initial work on this topic in workflow platforms, such as the "Meaning Spatial Statistics" initiative (Stasch et al., 2014), and we anticipate that more projects will be launched in the near future.

4.3. Science of Team Science to Facilitate Data-Driven Geoscience Discovery

In the data ecosystem underpinned by open science, there will be small data science projects that only require a small team, personal computers, and open source software packages. There will also be large-scale data science projects that cross disciplinary boundaries and require the collaboration of researchers from different institutions, high-performance computing facilities, efficient infrastructure for data storage and transmission, and large software programs for data management and processing. To succeed in such data science projects, the science of team science is recommended by many communities (NASEM, 2015). Key elements of the science of team science include (1) clear communication to reach consensus on the objective among team members, (2) regular brainstorming activities to identify and specify research questions, (3) complementary expertise from team members on problem solving, (4) regular team meetings to review progress and seek alternative approaches, and (5) positive and supportive working relationships within the team. The recent collaboration on data-driven mineral evolution study (Hazen et al., 2019) shows successful real-world practices of team science. In that work, a list of activities was organized to create an environment where people from different knowledge backgrounds could quickly step out of their comfort zones, get familiar with each other, and work together on focused scientific topics.

Geoscience communities also need some cultural change to fully embrace open data and open science. The NASEM (2020) "Earth in Time" report envisioned a list of science priority questions for the NSF Earth science programs in the next decade. The report also made two recommendations on cyberinfrastructure. One is about a strategy to support FAIR data practices in community data efforts and the other is about the initiation of a community-based standing committee to provide advice on cyberinfrastructure needs and advances. Community of practice has received increasing attention in many academic associations and has been discussed as a catalyzer for open science (Cutcher-Gershenfeld et al., 2017). Many researchers have been actively promoting open science in geoscience (Caron, 2020). For instance, the Earth Science Information Partners, through collaboration with EarthCube, the American Geophysical Union, European Geosciences Union, Geological Society of America, American Meteorological Society, and other organizations, has successfully organized many successful Data Help Desk activities recently and archived a long list of reusable resources (ESIP, 2020). Hundreds of researchers across the world have joined those activities as volunteers to answer questions and share research outcomes. We anticipate that more such activities will be organized in the future to promote data science applications and cultural change in the geosciences.

5. CONCLUDING REMARKS

This paper presents a review of recent data science activities in geoscience from the perspective of a data life cycle. It first provides a description of the basic concepts and theoretical foundation of data science. Then, by following the process of the data life cycle, it reviews a number of the latest publications on each step in the data life cycle and summarizes the shareable experience from them. Finally, a vision of the trends in data science applications in geoscience is discussed, including open science, smart data, and the science of team science. The author hopes the review from the aspect of a data life cycle will lower the barrier of data science for geoscientists, especially newcomers to data science applications. Individual geoscientists can gain awareness of resources available in the cyberinfrastructure, explore representative examples of data science, and initiate ideas for their own work. Research teams can learn methods for collaboration and team science. Geoscientists have been successfully embracing the strategy of community of practice to share data science resources and promote best practices. The author hopes the open science campaign will further facilitate data science applications in geoscience and lead to more data-driven scientific discoveries.

ACKNOWLEDGMENTS

The work presented in this paper was supported by the National Science Foundation under grants 1835717, 2019609, and 2126315. Additional support was provided by the International Union of Geological Sciences Deep-Time Digital Earth (DDE) Big Science program, the Deep Carbon Observatory, the Alfred P. Sloan Foundation, and the Carnegie Institution for Science for communicating research progress at several workshops and meetings.

REFERENCES CITED

4D Initiative, 2018, White Paper of the 4D Initiative: Deep-Time Data Driven Discovery: https://4d.carnegiescience.edu/sites/default/files/4D_materials/4D_WhitePaper.pdf (March 04, 2020).

Acker, J.G., and Leptoukh, G., 2007, Online analysis enhances use of NASA earth science data: Eos (Transactions, American Geophysical Union), v. 88, no. 2, p. 14–17, https://doi.org/10.1029/2007EO020003.

Adhikari, A., and DeNero, J., 2017, Computational and inferential thinking: The foundations of data science: https://www.inferentialthinking.com (accessed January 2021).

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

Ahmouda, A., Hochmair, H.H., and Cvetojevic, S., 2018, Analyzing the effect of earthquakes on OpenStreetMap contribution patterns and tweeting activities: Geo-Spatial Information Science, v. 21, no. 3, p. 195–212, https://doi.org/10.1080/10095020.2018.1498666.

Aspesi, C., and Brand, A., 2020, In pursuit of open science, open access is not enough: Science, v. 368, no. 6491, p. 574–577, https://doi.org/10.1126/science.aba3763.

Bartha, G., and Kocsis, S., 2011, Standardization of geographic data: The European INSIPIRE Directive: European Journal of Geography, v. 2, no. 2, p. 79–89.

Bedia, J., San-Martín, D., Iturbide, M., Herrera, S., Manzanas, R., and Gutiérrez, J.M., 2019, The METACLIP semantic provenance framework for climate products: Environmental Modelling & Software, v. 119, p. 445–457, https://doi.org/10.1016/j.envsoft.2019.07.005.

Berendt, B., Gandon, F., Halford, S., Hall, W., Hendler, J., Kinder-Kurlanda, K., Ntoutsi, E., and Staab, S., eds., 2020, Web futures: Inclusive, intelligent, sustainable: The 2020 manifesto for web science: http://webscience.org/the-2020-manifesto-for-web-science/ (accessed January 2021).

Bergen, K.J., Johnson, P.A., Maarten, V., and Beroza, G.C., 2019, Machine learning for data-driven discovery in solid Earth geoscience: Science, v. 363, no. 6433, https://doi.org/10.1126/science.aau0323.

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., Stodden, V., and Szalay, A.S., 2018, Realizing the potential of data science: Communications of the Association for Computing Machinery, v. 61, no. 4, p. 67–72, https://doi.org/10.1145/3188721.

Berners-Lee, T., 2000, Semantic Web on XML. Presentation at XML 2000 Conference: Washington, D.C., World Wide Web Consortium, http://www.w3.org/2000/Talks/1206-xml2k-tbl (accessed 24 January 2021).

Berners-Lee, T., Hendler, J., and Lassila, O., 2001, The Semantic Web: Scientific American, v. 284, no. 5, p. 34–43, https://doi.org/10.1038/scientificamerican0501-34.

Bishr, Y., 1998, Overcoming the semantic and other barriers to GIS interoperability: International Journal of Geographical Information Science, v. 12, no. 4, p. 299–314, https://doi.org/10.1080/136588198241806.

Brodaric, B., 2007, Geo-pragmatics for the Geospatial Semantic Web: Transactions in GIS, v. 11, no. 3, p. 453–477, https://doi.org/10.1111/j.1467-9671.2007.01055.x.

Brodaric, B., 2018, Interoperability of representations, *in* Richardson, D., Castree, N., Goodchild, M.F., Kobayashi, A., Liu, W., and Marston, R.A., eds., The International Encyclopedia of Geography: Hoboken, New Jersey, John Wiley & Sons, 18 p., https://doi.org/10.1002/9781118786352.wbieg0894.pub2.

Camizuli, E., and Carranza, E.J., 2018, Exploratory Data Analysis (EDA), *in* Varela, S.L.L., ed., The Encyclopedia of Archaeological Sciences: Hoboken, New Jersey, Wiley Online Library, 7 p., https://doi.org/10.1002/9781119188230.saseas0271.

Carneiro, H.A., and Mylonakis, E., 2009, Google trends: A web-based tool for real-time surveillance of disease outbreaks: Clinical Infectious Diseases, v. 49, no. 10, p. 1557–1564, https://doi.org/10.1086/630200.

Caron, B.C., 2020, Open Scientist Handbook, 305 p., https://doi.org/10.21428/8bbb7f85.35a0e14b.

Chan, M.A., Peters, S.E., and Tikoff, B., 2016, The future of field geology, open data sharing and cybertechnology in Earth science: The Sedimentary Record, v. 14, p. 4–10, https://doi.org/10.2110/sedred.2016.1.4.

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., and Groth, P., 2020, Dataset search: A survey: The VLDB Journal, v. 29, no. 1, p. 251–272, https://doi.org/10.1007/s00778-019-00564-x.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R., 2000, CRISP-DM 1.0: Step-by-Step Data Mining Guide: CRISP-DM Consortium, 78 p.

Chari, S., Seneviratne, O., Gruen, D.M., Foreman, M.A., Das, A.K., and McGuinness, D.L., 2020, November. Explanation ontology: A model of explanations for user-centered AI, *in* Pan, J.Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., and Polleres, A., eds., The Semantic Web—ISWC 2020: Cham, Switzerland, Springer, p. 228–243.

Cheng, Q., Oberhänsli, R., and Zhao, M., 2020, A new international initiative for facilitating data-driven Earth science transformation, *in* Hill, P.R., Lebel, D., Hitzman, M., Smelror, M., and Thorleifson, H., eds., The Changing Role of Geological Surveys: Geological Society, London, Special Publication 499, p. 225–240, https://doi.org/10.1144/SP499-2019-158.

Choi, Y.D., Goodall, J.L., Sadler, J.M., Castronova, A.M., Bennett, A., Li, Z., Nijssen, B., Wang, S., Clark, M.P., Ames, D.P., and Horsburgh, J.S., 2021, Toward open and reproducible environmental modeling by integrating online data repositories, computational environments, and model Application Programming Interfaces: Environmental Modelling & Software, v. 135, https://doi.org/10.1016/j.envsoft.2020.104888.

Cleveland, W.S., 2001, Data science: An action plan for expanding the technical areas of the field of statistics: International Statistical Review, v. 69, no. 1, p. 21–26, https://doi.org/10.1111/j.1751-5823.2001.tb00477.x.

Cutcher-Gershenfeld, J., Baker, K.S., Berente, N., Carter, D.R., DeChurch, L.A., Flint, C.G., Gershenfeld, G., Haberman, M., King, J.L., Kirkpatrick, C., and Knight, E., 2016, Build it, but will they come? A geoscience cyberinfrastructure baseline analysis: Data Science Journal, v. 15, p. 8, https://doi.org/10.5334/dsj-2016-008.

Cutcher-Gershenfeld, J., Baker, K.S., Berente, N., Flint, C., Gershenfeld, G., Grant, B., Haberman, M., King, J.L., Kirkpatrick, C., Lawrence, B., and Lewis, S., 2017, Five ways consortia can catalyse open science: Nature, v. 543, no. 7647, p. 615–617, https://doi.org/10.1038/543615a.

DDI Alliance, 2021, Why use DDI?: https://ddialliance.org/training/why-use-ddi (accessed January 2021).

Di, L., Yue, P., Ramapriyan, H.K., and King, R.L., 2013, Geoscience data provenance: An overview: IEEE Transactions on Geoscience and Remote Sensing, v. 51, no. 11, p. 5065–5072, https://doi.org/10.1109/TGRS.2013.2242478.

Dietrich, D., Adamus, T., Miner, A., and Steinhart, G., 2012, De-mystifying the data management requirements of research funders: Issues in Science & Technology Librarianship, no. 70, Summer, https://doi.org/10.5062/F44M92G2.

Donoho, D., 2017, 50 years of data science: Journal of Computational and Graphical Statistics, v. 26, no. 4, p. 745–766, https://doi.org/10.1080/10618600.2017.1384734.

Drineas, P., and Huo, X., 2016, NSF Workshop Report: Theoretical Foundations of Data Science (TFoDS): http://www.cs.rpi.edu/TFoDS/TFoDS_v5.pdf (accessed January 2021).

Dutkiewicz, A., Müller, R.D., O'Callaghan, S., and Jónasson, H., 2015, Census of seafloor sediments in the world's ocean: Geology, v. 43, no. 9, p. 795–798, https://doi.org/10.1130/G36883.1.

ESIP (Earth Science Information Partners), 2020, Data Help Desk: Connecting researchers and data experts to enhance research and make data and software more open and FAIR: https://www.esipfed.org/data-help-desk (accessed January 2021).

Fan, J.X., Shen, S.Z., Erwin, D.H., Sadler, P.M., MacLeod, N., Cheng, Q.M., Hou, X.D., Yang, J., Wang, X.D., Wang, Y., and Zhang, H., 2020a, A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity: Science, v. 367, no. 6475, p. 272–277, https://doi.org/10.1126/science.aax4953.

Fan, R., Wang, L., Yan, J., Song, W., Zhu, Y., and Chen, X., 2020b, Deep learning-based named entity recognition and knowledge graph construction for geological hazards: ISPRS International Journal of Geo-Information, v. 9, no. 1, p. 15, https://doi.org/10.3390/ijgi9010015.

Floridi, L., 2019, Establishing the rules for building trustworthy AI: Nature Machine Intelligence, v. 1, no. 6, p. 261–262, https://doi.org/10.1038/s42256-019-0055-y.

Fox, P., 2019, Disruption in biogeosciences: Conceptual, methodological, digital, and technological: Acta Geologica Sinica, v. 93, no. S3, p. 17–18, https://doi.org/10.1111/1755-6724.14231.

Fox, P., and Hendler, J., 2011, Changing the equation on scientific data visualization: Science, v. 331, no. 6018, p. 705–708, https://doi.org/10.1126/science.1197654.

Fox, P., and Hendler, J., 2014, Science of data science: Big Data, v. 2, no. 2, p. 68–70, https://doi.org/10.1089/big.2014.0011.

GCMD (Global Change Master Directory), 2020, GCMD Keywords, Version 9.1. Earth Science Data and Information System, Earth Science Projects Division, Goddard Space Flight Center (GSFC), National Aeronautics and Space Administration (NASA): https://wiki.earthdata.nasa.gov/display/gcmdkey (accessed January 2021).

Gil, Y., David, C.H., Demir, I., Essawy, B.T., Fulweiler, R.W., Goodall, J.L., Karlstrom, L., Lee, H., Mills, H.J., Oh, J.H., and Pierce, S.A., 2016, Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance: Earth and Space Science, v. 3, no. 10, p. 388–415, https://doi.org/10.1002/2015EA000136.

Gil, Y., Pierce, S.A., Babaie, H., Banerjee, A., Borne, K., Bust, G., Cheatham, M., Ebert-Uphoff, I., Gomes, C., Hill, M., and Horel, J., 2019, Intelligent systems for geosciences: An essential research agenda: Communications of the Association for Computing Machinery, v. 62, no. 1, p. 76–84, https://doi.org/10.1145/3192335.

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

Golden, J.J., Downs, R.T., Hazen, R.M., Pires, A.J., and Ralph, J., 2019, Mineral Evolution Database: Data-driven age assignment, how does a mineral get an age?: Geological Society of America Abstracts with Programs, v. 51, no. 5, https://doi.org/10.1130/abs/2019AM-334056.

Goovaerts, P., 2008, Geostatistical analysis of health data: State-of-the-art and perspectives, *in* Soares, A., Pereira, M.J., and Dimitrakopoulos, R., eds., geoENV VI—Geostatistics for Environmental Applications: Dordrecht, Netherlands, Springer, p. 3–22.

Goovaerts, P., 2021, From natural resources evaluation to spatial epidemiology: 25 years in the making: Mathematical Geosciences, v. 53, p. 239–266, https://doi.org/10.1007/s11004-020-09886-x.

Goring, S., Dawson, A., Simpson, G., Ram, K., Graham, R., Grimm, E., and Williams, J., 2015, Neotoma: A programmatic interface to the Neotoma Paleoecological Database: Open Quaternary, v. 1, no. 1, p. 2, https://doi.org/10.5334/oq.ab.

Gruber, T.R., 1995, Toward principles for the design of ontologies used for knowledge sharing?: International Journal of Human-Computer Studies, v. 43, no. 5–6, p. 907–928, https://doi.org/10.1006/ijhc.1995.1081.

Hagras, H., 2018, Toward human-understandable, explainable AI: Computer, v. 51, no. 9, p. 28–36, https://doi.org/10.1109/MC.2018.3620965.

Hamman, J., Rocklin, M., and Abernathy, R., 2018. Pangeo: A big-data ecosystem for scalable earth system science: 2014 EGU General Assembly Conference Abstracts, v. 20, no. EGU2018-12146.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., and Kommareddy, A., 2013, High-resolution global maps of 21st-century forest cover change: Science, v. 342, p. 850–853.

Hart, J.K., and Martinez, K., 2006, Environmental sensor networks: A revolution in the earth system science?: Earth-Science Reviews, v. 78, no. 3–4, p. 177–191, https://doi.org/10.1016/j.earscirev.2006.05.001.

Hazen, R.M., 2014, Data-driven abductive discovery in mineralogy: The American Mineralogist, v. 99, no. 11–12, p. 2165–2170, https://doi.org/10.2138/am-2014-4895.

Hazen, R.M., Bekker, A., Bish, D.L., Bleeker, W., Downs, R.T., Farquhar, J., Ferry, J.M., Grew, E.S., Knoll, A.H., Papineau, D., and Ralph, J.P., 2011, Needs and opportunities in mineral evolution research: The American Mineralogist, v. 96, no. 7, p. 953–963, https://doi.org/10.2138/am.2011.3725.

Hazen, R.M., Downs, R.T., Eleish, A., Fox, P., Gagne, O., Golden, J.J., Grew, E.S., Hummer, D.R., Hystad, G., Krivovichev, S.V., Li, C., Liu, C., Ma, X., Morrison, S.M., Pan, F., Pires, A.J., Prabhu, A., Ralph, J., Rumyon, S.E., and Zhong, H., 2019, Data-driven discovery in mineralogy: Recent advances in data resources, analysis, and visualization: Engineering, v. 5, no. 3, p. 397–405, https://doi.org/10.1016/j.eng.2019.03.006.

He, Z., Liu, G., Ma, X., and Chen, Q., 2019, GeoBeam: A distributed computing framework for spatial data: Computers & Geosciences, v. 131, p. 15–22, https://doi.org/10.1016/j.cageo.2019.06.003.

Hey, T., Tansley, S., and Tolle, K., eds., 2009, The Fourth Paradigm: Data-Intensive Scientific Discovery: Redmond, Washington, Microsoft Corporation, 252 p.

Ho, Y.C., 1994, Abduction? Deduction? Induction? Is there a logic of exploratory data analysis?, *in* Proceedings of the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, 28 p.

Holzinger, A., 2018, From machine learning to explainable AI, *in* Proceedings of the 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), Kosice, Slovakia, p. 55–66.

Hystad, G., Downs, R.T., and Hazen, R.M., 2015, Mineral species frequency distribution conforms to a large number of rare events model: Prediction of Earth's missing minerals: Mathematical Geosciences, v. 47, no. 6, p. 647–661, https://doi.org/10.1007/s11004-015-9600-3.

Hystad, G., Eleish, A., Hazen, R.M., Morrison, S.M., and Downs, R.T., 2019, Bayesian estimation of Earth's undiscovered mineralogical diversity using noninformative priors: Mathematical Geosciences, v. 51, no. 4, p. 401–417, https://doi.org/10.1007/s11004-019-09795-8.

Iafrate, F., 2014, A journey from big data to smart data, *in* Benghozi, P.-J., Krob, D., Lonjon, A., and Panetto, H., eds., Digital Enterprise Design & Management: Cham, Switzerland, Springer, p. 25–33, https://doi.org/10.1007/978-3-319-04313-5_3.

Jackson, I., 2010, OneGeology: Improving access to geoscience globally: Earthwise, v. 26, p. 14–15.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., and Kumar, V., 2019, Machine learning for the geosciences: Challenges and opportunities: IEEE Transactions on Knowledge and Data Engineering, v. 31, no. 8, p. 1544–1554, https://doi.org/10.1109/TKDE.2018.2861006.

Kelleher, J.D., and Tierney, B., 2018, Data Science: Cambridge, Massachusetts, MIT Press, 280 p., https://doi.org/10.7551/mitpress/11140.001.0001.

Keller, C.B., and Schoene, B., 2012, Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gyr ago: Nature, v. 485, no. 7399, p. 490–493, https://doi.org/10.1038/nature11024.

Keller, C.B., Schoene, B., Barboni, M., Samperton, K.M., and Husson, J.M., 2015, Volcanic–plutonic parity and the differentiation of the continental crust: Nature, v. 523, no. 7560, p. 301–307, https://doi.org/10.1038/nature14584.

Kitchin, R., 2014, The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences: London, Sage, 222 p., https://doi.org/10.4135/9781473909472.

Kitchin, R., and Lauriault, T.P., 2015, Small data in the era of big data: GeoJournal, v. 80, no. 4, p. 463–475, https://doi.org/10.1007/s10708-014-9601-7.

Klump, J., Huber, R., and Diepenbroek, M., 2016, DOI for geoscience data—How early practices shape present perceptions: Earth Science Informatics, v. 9, no. 1, p. 123–136, https://doi.org/10.1007/s12145-015-0231-5.

Lannom, L., Koureas, D., and Hardisty, A.R., 2020, FAIR data and services in biodiversity science and geoscience: Data Intelligence, v. 2, no. 1–2, p. 122–130, https://doi.org/10.1162/dint_a_00034.

Lary, D.J., Alavi, A.H., Gandomi, A.H., and Walker, A.L., 2016, Machine learning in geosciences and remote sensing: Geoscience Frontiers, v. 7, no. 1, p. 3–10, https://doi.org/10.1016/j.gsf.2015.07.003.

Laxton, J.L., 2017, Geological map fusion: OneGeology-Europe and INSPIRE, *in* Riddick, A.T., Kessler, H., and Giles, J.R.A., eds., Integrated Environmental Modelling to Solve Real World Problems: Methods, Vision and Challenges: Geological Society, London, Special Publication 408, p. 147–160, https://doi.org/10.1144/SP408.16.

Lebo, T., Sahoo, S., and McGuinness, D., 2013, PROV-O: The PROV Ontology. W3C recommendation: https://www.w3.org/TR/2013/REC-prov-o-20130430 (accessed January 2021).

Lehmann, A., Nativi, S., Mazzetti, P., Maso, J., Serral, I., Spengler, D., Niamir, A., McCallum, I., Lacroix, P., Patias, P., and Rodila, D., 2020, GEOEssential—Mainstreaming workflows from data sources to environment policy indicators with essential variables: International Journal of Digital Earth, v. 13, no. 2, p. 322–338, https://doi.org/10.1080/17538947.2019.1585977.

Li, Z., Yang, C., Jin, B., Yu, M., Liu, K., Sun, M., and Zhan, M., 2015, Enabling big geoscience data analytics with a cloud-based, MapReduce-enabled and service-oriented workflow framework: PLoS One, v. 10, no. 3, https://doi.org/10.1371/journal.pone.0116781.

Loscio, B.F., Burle, C., and Calegari, N., eds., 2017, Data on the web best practices, W3C recommendation: https://www.w3.org/TR/dwbp.

Ludäscher, B., Lin, K., Brodaric, B., and Baru, C., 2003, GEON: Toward a cyberinfrastructure for the geosciences—A prototype for geological map interoperability via domain ontologies, *in* Soller, D.R., ed., Digital Mapping Techniques '03—Workshop Proceedings, 1–4 June, Millersville, Pennsylvania: U.S. Geological Survey Open-File Report 03-471 p. 223–229, https://pubs.usgs.gov/of/2003/of03-471/report.pdf.

Lynnes, C., 2020, Federated Giovanni for multi-sensor data exploration: https://earthdata.nasa.gov/esds/competitive-programs/access/federated-giovanni (accessed January 2021).

Ma, X., 2018, Data science for geoscience: Leveraging mathematical geosciences with semantics and open data, *in* Sagar, B.S.D., Cheng, Q., and Agterberg, F.D., eds., Handbook of Mathematical Geosciences: Fifty Years of IAMG: Cham, Switzerland, Springer, p. 687–702, https://doi.org/10.1007/978-3-319-78999-6_34.

Ma, X., and Fox, P., 2014, A jigsaw puzzle layer cake of spatial data: Eos (Transactions, American Geophyscial Union), v. 95, no. 19, p. 161–162, https://doi.org/10.1002/2014EO190006.

Ma, X., Asch, K., Laxton, J.L., Richard, S.M., Asato, C.G., Carranza, E.J.M., van der Meer, F.D., Wu, C., Duclaux, G., and Wakita, K., 2011, Data exchange facilitated: Nature Geoscience, v. 4, no. 12, p. 814, https://doi.org/10.1038/ngeo1335.

Ma, X., Chen, Y., Wang, H., Zheng, J., Fu, L., West, P., Erickson, J.S., and Fox, P., 2015, Data visualization in the Semantic Web, *in* Narock, T., and Fox, P., eds., The Semantic Web in Earth and Space Science: Current Status and Future Directions: Berlin, IOS Press, p. 149–167.

Ma, X., Hummer, D., Golden, J.J., Fox, P.A., Hazen, R.M., Morrison, S.M., Downs, R.T., Madhikarmi, B.L., Wang, C., and Meyer, M.B., 2017, Using visual exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research: ISPRS International Journal of Geo-Information, v. 6, no. 11, p. 368, https://doi.org/10.3390/ijgi6110368.

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

Ma, X., Ma, C., and Wang, C., 2020, A new structure for representing and tracking version information in a deep time knowledge graph: Computers & Geosciences, v. 145, https://doi.org/10.1016/j.cageo.2020.104620.

Martinez, K., Hart, J.K., and Ong, R., 2004, Environmental sensor networks: Computer, v. 37, no. 8, p. 50–56, https://doi.org/10.1109/MC.2004.91.

Maskey, M., Alemohammad, H., Murphy, K.J., and Ramachandran, R., 2020, Advancing AI for Earth science: A data systems perspective: Eos (Transactions, American Geophysical Union), v. 101, https://doi.org/10.1029/2020EO151245.

Mattmann, C.A., 2013, A vision for data science: Nature, v. 493, no. 7433, p. 473–475, https://doi.org/10.1038/493473a.

McGuinness, D.L., 2003, Ontologies come of age, *in* Fensel, D., Hendler, J., Lieberman, H., and Wahlster, W., eds., Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential: Cambridge, Massachusetts, MIT Press, p. 171–196.

Merriam, D., 2004, The quantification of geology: From abacus to pentium: A chronicle of people, places, and phenomena: Earth-Science Reviews, v. 67, no. 1–2, p. 55–89, https://doi.org/10.1016/j.earscirev.2004.02.002.

Mons, B., 2018, Data Stewardship for Open Science: Implementing FAIR Principles: New York, Chapman and Hall, 244 p., https://doi.org/10.1201/9781315380711.

Mookerjee, M., Vieira, D., Chan, M.A., Gil, Y., Pavlis, T.L., Spear, F.S., and Tikoff, B., 2015, Field data management: Integrating cyberscience and geoscience: Eos (Transactions, American Geophysical Union), v. 96, no. 20, p. 18–21, https://doi.org/10.1029/2015EO036703.

Mookerjee, M., Chan, M.A., Gil, Y., Gill, G., Goodwin, C., Pavlis, T.L., Shipley, T.F., Swain, T., Tikoff, B., and Vieira, D., 2022, this volume, Cyberinfrastructure for collecting and integrating geology field data: Community priorities and research agenda, *in* Ma, X., Mookerjee, M., Hsu, L., and Hills, D., eds., Recent Advancement in Geoinformatics and Data Science: Geological Society of America Special Paper 558, https://doi.org/10.1130/2022.2558(01).

Morrison, S.M., Liu, C., Eleish, A., Prabhu, A., Li, C., Ralph, J., Downs, R.T., Golden, J.J., Fox, P., Hummer, D.R., and Meyer, M.B., 2017, Network analysis of mineralogical systems: The American Mineralogist, v. 102, no. 8, p. 1588–1596, https://doi.org/10.2138/am-2017-6104CCBYNCND.

Morrison, S.M., Prabhu, A., Eleish, A., Pan, F., Zhong, H., Huang, F., Fox, P., Ma, X., Ralph, J., Golden, J.J., and Downs, R.T., 2019, Application of advanced analytics and visualization in mineral systems: Acta Geologica Sinica, v. 93, no. S3, p. 55, https://doi.org/10.1111/1755-6724.14243.

Morrison, S.M., Buongiorno, J., Downs, R.T., Eleish, A., Fox, P., Giovannelli, D., Golden, J.J., Hummer, D.R., Hystad, G., Kellogg, L.H., and Kreylos, O., 2020, Exploring carbon mineral systems: Recent advances in C mineral evolution, mineral ecology, and network analysis: Frontiers of Earth Science, v. 8, p. 208, https://doi.org/10.3389/feart.2020.00208.

Müller, R.D., Cannon, J., Qin, X., Watson, R.J., Gurnis, M., Williams, S., Pfaffelmoser, T., Seton, M., Russell, S.H., and Zahirovic, S., 2018, GPlates: Building a virtual Earth through deep time: Geochemistry, Geophysics, Geosystems, v. 19, no. 7, p. 2243–2261, https://doi.org/10.1029/2018GC007584.

Muscente, A.D., Prabhu, A., Zhong, H., Eleish, A., Meyer, M.B., Fox, P., Hazen, R.M., and Knoll, A.H., 2018, Quantifying ecological impacts of mass extinctions with network analysis of fossil communities *in* Proceedings of the National Academy of Sciences of the United States of America, v. 115, no. 20, p. 5217–5222, https://doi.org/10.1073/pnas.1719976115.

NADM Steering Committee, 2004, NADM Conceptual Model 1.0—A Conceptual Model for Geologic Map Information: U.S. Geological Survey Open-File Report 2004-1334, 58 p., https://doi.org/10.3133/ofr20041334.

Narock, T., and Shepherd, A., 2017, Semantics all the way down: The Semantic Web and open science in big earth data: Big Earth Data, v. 1, no. 1–2, p. 159–172, https://doi.org/10.1080/20964471.2017.1397408.

NASEM (National Academies of Sciences, Engineering, and Medicine), 2015, Enhancing the Effectiveness of Team Science: Washington, D.C., The National Academies Press, 268 p., https://doi.org/10.17226/19007.

NASEM (National Academies of Sciences, Engineering, and Medicine), 2018a, Data Science for Undergraduates: Opportunities and Options: Washington, D.C., The National Academies Press, 107 p., https://doi.org/10.17226/25104.

NASEM (National Academies of Sciences, Engineering, and Medicine), 2018b, Open Science by Design: Realizing a Vision for 21st Century Research: Washington, D.C., The National Academies Press, 216 p., https://doi.org/10.17226/25116.

NASEM (National Academies of Sciences, Engineering, and Medicine), 2020, A Vision for NSF Earth Sciences 2020–2030: Earth in Time: Washington, D.C., The National Academies Press, 172 p., https://doi.org/10.17226/25761.

Noy, N., Burgess, M., and Brickley, D., 2019, Google Dataset Search: Building a search engine for datasets in an open web ecosystem, *in* Proceedings of the 2019 World Wide Web Conference, San Francisco, California: New York, Association for Computing Machinery, p. 1365–1375.

NSF (National Science Foundation), 2015, NSF Public Access Plan: Today's Data, Tomorrow's Discoveries—Increasing Access to the Results of Research Funded by the National Science Foundation, 31 p., https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf (accessed May 2019).

Obrst, L., 2003, Ontologies for semantically interoperable systems, *in* Proceedings, Twelfth International Conference on Information and Knowledge Management, 3–8 November, New Orleans, Louisiana: New York, Association for Computing Machinery, p. 366–369.

Parsons, M.A., Duerr, R., and Minster, J.B., 2010, Data citation and peer review: Eos (Transactions, American Geophysical Union), v. 91, no. 34, p. 297–298, https://doi.org/10.1029/2010EO340001.

Peters, S.E., and McClennen, M., 2016, The Paleobiology Database application programming interface: Paleobiology, v. 42, no. 1, p. 1–7, https://doi.org/10.1017/pab.2015.39.

Peters, S.E., Zhang, C., Livny, M., and Ré, C., 2014, A machine reading system for assembling synthetic paleontological databases: PLoS One, v. 9, no. 12, https://doi.org/10.1371/journal.pone.0113523.

Peters, S.E., Husson, J.M., and Wilcots, J., 2017, The rise and fall of stromatolites in shallow marine environments: Geology, v. 45, no. 6, p. 487–490, https://doi.org/10.1130/G38931.1.

Prabhu, A., Morrison, S.M., Eleish, A., Zhong, H., Huang, F., Golden, J.J., Perry, S.N., Hummer, D.R., Ralph, J., Runyon, S.E., and Fontaine, K., 2021, Global earth mineral inventory: A data legacy: Geoscience Data Journal, v. 8, no. 1, p. 74–89, https://doi.org/10.1002/gdj3.106.

Press, G., 2016 (23 March), Cleaning big data: Most time-consuming, least enjoyable data science task, survey says: Forbes, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/ (accessed January 2021).

Qiu, Q., Xie, Z., Wu, L., and Li, W., 2019, Geoscience keyphrase extraction algorithm using enhanced word embedding: Expert Systems with Applications, v. 125, p. 157–169, https://doi.org/10.1016/j.eswa.2019.02.001.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., and Carvalhais, N., 2019, Deep learning and process understanding for data-driven Earth system science: Nature, v. 566, no. 7743, p. 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Reitsma, F., Laxton, J., Ballard, S., Kuhn, W., and Abdelmoty, A., 2009, Semantics, ontologies and eScience for the geosciences: Computers & Geosciences, v. 35, no. 4, p. 706–709, https://doi.org/10.1016/j.cageo.2008.03.014.

Rosenberg, D.E., Filion, Y., Teasley, R., Sandoval-Solis, S., Hecht, J.S., Van Zyl, J.E., McMahon, G.F., Horsburgh, J.S., Kasprzyk, J.R., and Tarboton, D.G., 2020, The next frontier: Making research more reproducible: Journal of Water Resources Planning and Management, v. 146, no. 6, https://doi.org/10.1061/(ASCE)WR.1943-5452.0001215.

Rossi, S.D., Barros, A., Walden-Schreiner, C., and Pickering, C., 2020, Using social media images to assess ecosystem services in a remote protected area in the Argentinean Andes: Ambio, v. 49, p. 1146–1160, https://doi.org/10.1007/s13280-019-01268-w.

Sagar, D.B.S., Cheng, Q., and Agterberg, F., eds., 2018, Handbook of Mathematical Geosciences: Fifty Years of IAMG: Cham, Switzerland, Springer, 914 p., https://doi.org/10.1007/978-3-319-78999-6.

Schutt, R., and O'Neil, C., 2013, Doing Data Science: Straight Talk from the Frontline: New York, O'Reilly, 406 p.

Shannon, M., 2019, How does NASA use big data?: Big Data Made Simple, https://bigdata-madesimple.com/how-does-nasa-use-big-data (accessed January 2021).

Shepherd, A., Minnett, R., Jarboe, N., Koppers, A., Tauxe, L., Constable, C., and Jonestrask, L., 2019, Thorough Annotation of Magnetics Information Consortium (MagIC) contributions with Schema.org structured metadata: Abstract IN22B-01 presented at 2019 Fall Meeting, American Geophysical Union, San Francisco, California, 9–13 December.

Sheth, A., 2014, Transforming big data into smart data: Deriving value via harnessing volume, variety, and velocity using semantic techniques and technologies, *in* Proceedings of the 2014 IEEE 30th International Conference on Data Engineering (ICDE), Chicago, p. 2–2.

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

Sheth, A.P., 1999, Changing focus on interoperability in information systems: From system, syntax, structure to semantics, *in* Goodchild, M., Egenhofer, M., Fegeas, R., and Kottman, C., eds., Interoperating Geographic Information Systems: Dordrecht, Netherlands, Kluwer Academic Publishers, p. 5–29, https://doi.org/10.1007/978-1-4615-5189-8_2.

Shipley, T.F., and Tikoff, B., 2019, Collaboration, cyberinfrastructure, and cognitive science: The role of databases and dataguides in 21st century structural geology: Journal of Structural Geology, v. 125, p. 48–54, https://doi.org/10.1016/j.jsg.2018.05.007.

Sinha, A.K., Malik, Z., Rezgui, A., Barnes, C.G., Lin, K., Heiken, G., Thomas, W.A., Gundersen, L.C., Raskin, R., Jackson, I., and Fox, P., 2010, Geoinformatics: Transforming data to knowledge for geosciences: GSA Today, v. 20, no. 12, p. 4–10, https://doi.org/10.1130/GSATG85A.1.

Spielman, S.J., and Moore, E.K., 2020, dragon: A new tool for exploring redox evolution preserved in the mineral record: Frontiers of Earth Science, v. 8, https://doi.org/10.3389/feart.2020.585087.

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., and Wyborn, L., 2019, Make scientific data FAIR: Nature, v. 570, p. 27–29, https://doi.org/10.1038/d41586-019-01720-7.

Starr, J., and Gastl, A., 2011, isCitedBy: A metadata scheme for DataCite: D-Lib Magazine: The Magazine of the Digital Library Forum, v. 17, no. 1/2, https://doi.org/10.1045/january2011-starr.

Stasch, C., Scheider, S., Pebesma, E., and Kuhn, W., 2014, Meaningful spatial prediction and aggregation: Environmental Modelling & Software, v. 51, p. 149–165, https://doi.org/10.1016/j.envsoft.2013.09.006.

Stephenson, M.H., Cheng, Q., Wang, C., Fan, J., and Oberhänsli, R., 2020, Progress towards the establishment of the IUGS Deep-Time Digital Earth (DDE) programme: Episodes Journal of International Geoscience, v. 43, no. 4, p. 1057–1062, https://doi.org/10.18814/epiiugs/2020/020057.

Sun, A.Y., and Scanlon, B.R., 2019, How can big data and machine learning benefit environment and water management: A survey of methods, applications, and future directions: Environmental Research Letters, v. 14, no. 7, 073001, https://doi.org/10.1088/1748-9326/ab1b7d.

Sun, Z., Di, L., Burgess, A., Tullis, J.A., and Magill, A.B., 2020, Geoweaver: Advanced cyberinfrastructure for managing hybrid geoscientific AI workflows: ISPRS International Journal of Geo-Information, v. 9, no. 2, p. 119, https://doi.org/10.3390/ijgi9020119.

Tandy, J., van den Brink, L., and Barnaghi, P., eds., 2017, Spatial data on the web best practices, W3C Working Group note, https://www.w3.org/TR/sdw-bp.

Tukey, J.W., 1977, Exploratory Data Analysis: Reading, Pennsylvania, Addison-Wesley, 688 p.

Uschold, M., and Gruninger, M., 2004, Ontologies and semantics for seamless connectivity: SIGMOD Record, v. 33, no. 4, p. 58–64, https://doi.org/10.1145/1041410.1041420.

USGS MRDATA, 2021, Mineral resources online spatial data, https://mrdata.usgs.gov (accessed January 2021).

USGS NCGMP (U.S. Geological Survey National Cooperative Geologic Mapping Program), 2020, GeMS (Geologic Map Schema)—A Standard Format for the Digital Publication of Geologic Maps: Reston, Virginia, U.S. Geological Survey, 74 p., https://doi.org/10.3133/tm11B10.

Valentine, D., Zaslavsky, I., Richard, S., Meier, O., Hudman, G., Peucker-Ehrenbrink, B., and Stocks, K., 2020, EarthCube Data Discovery Studio: A gateway into geoscience data discovery and exploration with Jupyter notebooks: Concurrency and Computation, v. 33, no. 19, https://doi.org/10.1002/cpe.6086.

Varela, S., González-Hernández, J., Sgarbi, L.F., Marshall, C., Uhen, M.D., Peters, S., and McClennen, M., 2015, paleobioDB: An R package for downloading, visualizing and processing data from the Paleobiology Database: Ecography, v. 38, no. 4, p. 419–425, https://doi.org/10.1111/ecog.01154.

Wang, C., Ma, X., Chen, J., and Chen, J., 2018, Information extraction and knowledge graph construction from geoscience literature: Computers & Geosciences, v. 112, p. 112–120, https://doi.org/10.1016/j.cageo.2017.12.007.

Wang, Z., Ye, X., and Tsou, M.H., 2016, Spatial, temporal, and content analysis of Twitter for wildfire hazards: Natural Hazards, v. 83, no. 1, p. 523–540, https://doi.org/10.1007/s11069-016-2329-6.

Welle Donker, F., and van Loenen, B., 2017, How to assess the success of the open data ecosystem?: International Journal of Digital Earth, v. 10, no. 3, p. 284–306, https://doi.org/10.1080/17538947.2016.1224938.

Welty, C., 2002, Ontology-driven conceptual modeling, *in* Pidduck, A.B., Mylopoulos, J., Woo, C.C., and Ozsu, M.T., eds., Advanced Information Systems Engineering, Lecture Notes in Computer Science, Volume 2348: Berlin, Springer, p. 3.

Wen, T., Niu, X., Gonzales, M., Zheng, G., Li, Z., and Brantley, S.L., 2018, Big groundwater data sets reveal possible rare contamination amid otherwise improved water quality for some analytes in a region of Marcellus shale development: Environmental Science & Technology, v. 52, no. 12, p. 7149–7159, https://doi.org/10.1021/acs.est.8b01123.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., and Bouwman, J., 2016, The FAIR Guiding Principles for scientific data management and stewardship: Scientific Data, v. 3, no. 1, 160018, https://doi.org/10.1038/sdata.2016.18.

Wing, J.M., 2019, The data life cycle: Harvard Data Science Review, v. 1, no. 1, https://doi.org/10.1162/99608f92.e26845b4.

Wing, J.M., 2020, Ten research challenge areas in data science: Harvard Data Science Review, v. 2, no. 3, https://doi.org/10.1162/99608f92.c6577b1f.

Wood, J., Andersson, T., Bachem, A., Best, C., Genova, F., Lopez, D.R., Los, W., Marinucci, M., Romary, L., Van de Sompel, H., and Vigen, J., 2010, Riding the wave: How Europe can gain from the rising tide of scientific data, *in* Final Report of the High Level Expert Group on Scientific Data–A Submission to the European Commission: European Union, 36 p.

Yang, C., Huang, Q., Li, Z., Liu, K., and Hu, F., 2017, Big data and cloud computing: Innovation opportunities and challenges: International Journal of Digital Earth, v. 10, no. 1, p. 13–53, https://doi.org/10.1080/17538947.2016.1239771.

Yang, C., Yu, M., Li, Y., Hu, F., Jiang, Y., Liu, Q., Sha, D., Xu, M., and Gu, J., 2019, Big Earth data analytics: A survey: Big Earth Data, v. 3, no. 2, p. 83–107, https://doi.org/10.1080/20964471.2019.1611175.

Zeng, Y., Su, Z., Barmpadimos, I., Perrels, A., Poli, P., Boersma, K.F., Frey, A., Ma, X., de Bruin, K., Goosen, H., and John, V.O., 2019, Towards a traceable climate service: Assessment of quality and usability of essential climate variables: Remote Sensing, v. 11, no. 10, p. 1186, https://doi.org/10.3390/rs11101186.

Zhang, C., Govindaraju, V., Borchardt, J., Foltz, T., Ré, C., and Peters, S., 2013, GeoDeepDive: Statistical inference using familiar data-processing languages, *in* Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, p. 993–996.

Manuscript Accepted by the Society 17 March 2022
Manuscript Published Online 23 November 2022

Printed in the USA

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user

Downloaded from http://pubs.geoscienceworld.org/gsa/books/book/2377/chapter-pdf/5743375/spe558-05.pdf
by INGEMMET user