

Chapter 34

Data Science for Geoscience: Leveraging Mathematical Geosciences with Semantics and Open Data



Xiaogang Ma

Abstract Mathematical geosciences are now in an intelligent stage. The freshly new data environment enabled by the Semantic Web and Open Data poses both new challenges and opportunities for the conduction of geomathematical research. As an interdisciplinary domain, mathematical geosciences share many topics in common with data science. Facing the new data environment, will data science inject new blood into mathematical geosciences, and can data science benefit from the achievements and experiences of mathematical geosciences? This chapter presents a perspective on these questions and introduces a few recent case studies on data management and data analysis in the geosciences.

34.1 Introduction

The global science community is facing a fresh data environment that never existed before. New generations of sensors, instruments and platforms extend the range of exploration and speed up the frequency of data collection. The quick updates in data storage facilities make it possible to archive and retrieve massive datasets in digital formats. The wide coverage of Internet and World Wide Web services allow researchers to share datasets and communicate with colleagues efficiently both in the office and from the field. As transparency, openness and reproducibility of research results and methods receive increasing attention, the science community is now promoting an open science culture (Nosek et al. 2015) and encouraging actions on open access, open data, open code and open samples (Easterbook 2014; Hey and Payne 2015; McNutt et al. 2016). In the domain of geoscience, significant progress has been achieved on open data, including those emanating from federal agencies such as data services of NASA, USGS, NOAA and community-built data portals such as OneGeology, EarthChem, RRUFF, PANGAEA, PaleoBioDB, and more.

X. Ma (✉)

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

e-mail: max@uidaho.edu

© The Author(s) 2018

B. S. Daya Sagar et al. (eds.), *Handbook of Mathematical Geosciences*,

https://doi.org/10.1007/978-3-319-78999-6_34

687

A clear trend in open data actions is that the World Wide Web is used as the space for data storage, publication, discovery and access. Data resources on the Web provide convenience for geoscience researchers, and lay out the platform for cross-disciplinary collaboration and new scientific discoveries.

In addition to focused research topics within each discipline, geoscience researchers in the 21st century are now able to tackle more grand research questions (Fig. 34.1) that need broad perspectives, multidisciplinary collaboration and sustained data support. Studies on these questions will lead to the extension of our fundamental knowledge and understanding about the Earth system, which in turn will contribute to the application of geoscience in tackling social and economic issues that are relevant to human welfare. For example, the Future Earth, a ten-year initiative (2015–2025) coordinated by several international organizations, proposed eight key challenges to the global sustainability (Future Earth 2014): water-energy-food nexus, decarbonization, natural assets, cities, rural futures, human health, consumption and production, and social resilience. To grasp these tremendous opportunities and make innovative discoveries, geoscience researchers need the necessary data resources and skills. Although geoscience data are increasingly made available online, due to the heterogeneities inside them, many data are not ready for use by end users. The heterogeneities of geoscience data are reflected in the vast number of subjects, varied data structures and formats, and diverse terminologies (Berg-Cross et al. 2012; Ramachandran et al. 2006; Reitsma and Albrecht 2005). Methods and skills of both data management and data analysis are needed for conducting science within the inspiring and complex data environment of today.

Data management and data analysis are the two key concepts in data science (cf. Schutt and O’Neil 2013), which involves knowledge of library and information science, computer science, mathematics, statistics, and domain-specific disciplines. While the theoretical foundations of data science are still under development (Drineas and Huo 2016), there have already been many applications and

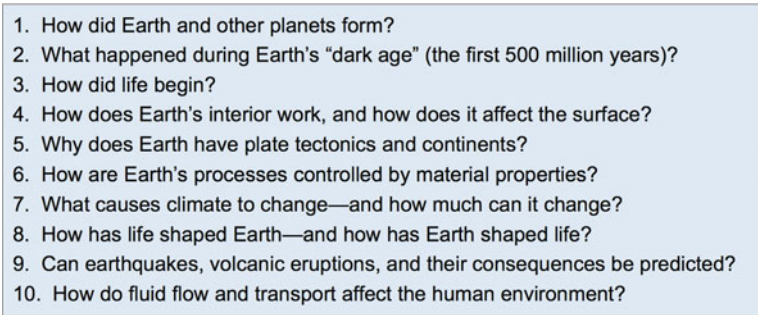
- 
1. How did Earth and other planets form?
 2. What happened during Earth’s “dark age” (the first 500 million years)?
 3. How did life begin?
 4. How does Earth’s interior work, and how does it affect the surface?
 5. Why does Earth have plate tectonics and continents?
 6. How are Earth’s processes controlled by material properties?
 7. What causes climate to change—and how much can it change?
 8. How has life shaped Earth—and how has Earth shaped life?
 9. Can earthquakes, volcanic eruptions, and their consequences be predicted?
 10. How do fluid flow and transport affect the human environment?

Fig. 34.1 The 10 grand research questions for the 21st century Earth sciences (National Research Council 2008)

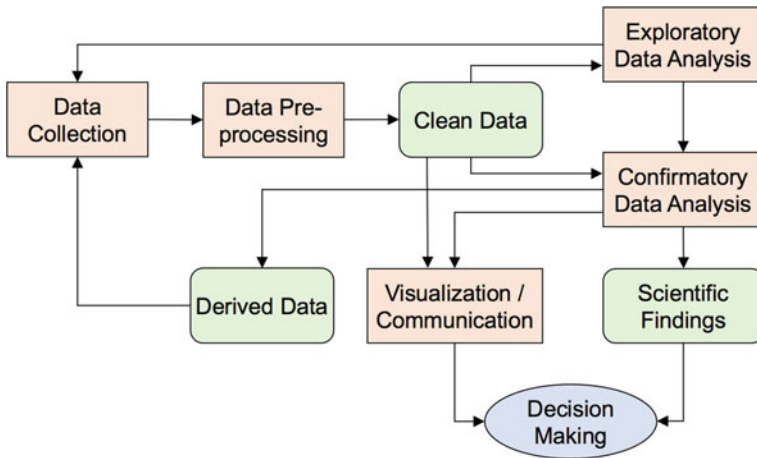


Fig. 34.2 Primary steps in a data science process. From Schutt and O’Neil (2013) with changes

discussions of data science in recent years (Schutt and O’Neil 2013), and a general process of data science is emerging (Fig. 34.2). The steps and processes in Fig. 34.2 would be familiar to researchers in all disciplines mentioned above, as they are comparable to the widely-adopted hypothesis-driven research method in modern science. Nevertheless, there could remain many questions to be asked as we are now in the “inspiring and complex data environment”: Do we have methods and techniques to improve the efficiency in each step? How to create a space and design an approach where researchers from the different disciplines can collaborate and leverage their individual capabilities to achieve a focused objective? What is the feature of data science in a domain-specific context, including geoscience?

Researchers of mathematical geosciences or geomathematics can have a lot to say about their experience and understanding of data science, because mathematical geosciences is a domain with a long history of incorporating knowledge from computer science, mathematics and statistics with geoscience (Agterberg 2014; Bonham-Carter 1994; Loudon 2000; Merriam 2004). Will the latest research progress of data science inject some new blood into the mathematical geosciences; and vice versa, can the methods and experiences in mathematical geosciences contribute to the theoretical developments of data science? The purpose of this chapter is to present a perspective on questions based on a review of the evolution of mathematical geosciences and a summary of the latest discussions of data science within the geoscience community. To support the presented perspective, a few recent case studies will be introduced in the second half of the chapter, with a focus on how data science can help leverage the existing capabilities in geoscience research and achieve new goals.

34.2 The Intelligent Stage of Mathematical Geosciences

34.2.1 *Evolution of Mathematical Geosciences*

Retrospection on the evolution of mathematical geosciences will help us understand the characteristics of this discipline as well as the opportunities it faces today. In an informative review, Merriam (2004) summarized the six stages in the development of quantitative geology: Origins (1650–1833), Formative (1833–1895), Exploration (1895–1941), Development (1941–1958), Automated (1958–1982), and Integration (1982–). The three earlier stages, over a period of almost 300 years, made use of various developments in both geoscience and mathematics, and more importantly the co-evolution between them. The latter three stages were characterized by the application of computers, first in geostatistics, simulation and modeling, and the organization of large datasets and later in all aspects of the geoscience workflow, including data capture, manipulation, analysis and documentation. Merriam (2004) also briefly mentioned the Internet and the potential challenges and opportunities in the connected virtual world, and he stated, “There is seemingly no limit to the information and communication revolution.”

Indeed, coming to today, which is just about 12 years after Merriam’s review paper, geomathematical researchers as well as the broad geoscience community already face the fresh data environment. We now have new instruments for measurement and observation, powerful facilities in data storage and transmission, improved interoperability of online datasets, and effective algorithms for data processing and analysis. New methods and technologies such as big data, open data, machine learning, data mining, data science, semantic web, natural language processing have been increasingly used in geoscience studies. The functionality of computers is being leveraged to a new level, where they are not only capable to represent “what is” known but can also show us “why” and help generate ideas on “how to” explore new findings. Ma (2015) proposed that the mathematical geosciences is now in an Intelligent stage (2014–). Besides these accelerated developments and applications of geomathematical methods within the geoscience disciplines, there are growing needs for using these methods in cross-disciplinary programs to address socio-economic issues that are of public concern (Freedman 2010).

In this intelligent stage, what we can do to leverage mathematical geosciences in various multidisciplinary studies? In this chapter, the author wants to address the need of refreshing our knowledge about the latest progress in open data and data science. For geoscience researchers, especially those who are not familiar with data science, knowing open data will be a key to understanding the general data science process and some featured works using datasets retrieved from the Web.

34.2.2 Characteristics of Open Data and Semantic Web

Most geoscience studies are driven by data. The term “open data” reflects people’s desire of access to freely available datasets. Some open data are made accessible with specified licenses and copyrights, and others are without any limits or restrictions. The popularity of the Internet and the Web creates a wide space for the implementation of open data. For end users of open data, an issue of extreme concern is the data interoperability (Fig. 34.3). Researchers have discussed the levels of data interoperability from different aspects. The levels in the center of Fig. 34.3 (Brodaric 2007) are from a technical point of view. Systems level is fundamental, which means there should be the necessary protocols (e.g. TCP/IP for the Internet and HTTP for the Web) supporting data discovery and transmission. Syntax and Schematics levels are relevant to the data structures and models, for which an end user should be able to parse and analyze. Semantics level indicates that the meaning of data reflected in data model, terminology and encoding are made readable to machines and thus understandable to users. Pragmatics level means the data are suitable for the user’s purpose and can contribute value in applications. The right part of Fig. 34.3 (Ma et al. 2011) explains these technical levels with layman’s language, and it also adds that all the technologies and implementations at those levels should be legal and ethical from a point of view of social science.

The Semantic Web (Berners-Lee 2000) provides technological support to each level of data interoperability (Fig. 34.3). For geoscience researchers, the Semantic Web creates a space where datasets can be more efficiently annotated, published, discovered and accessed. The Semantic Web is an extension to the current World Wide Web (Berners-Lee et al. 2001). The Web is now in the transition from a Web of Documents to a Web of Data because of the embedded structures and meanings that did not exist before. Nevertheless, to add structure and meaning to the

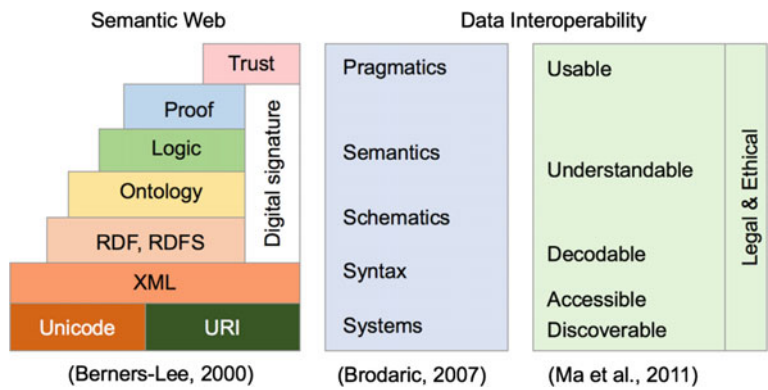


Fig. 34.3 Levels of data interoperability and a comparison with the architecture of the Semantic Web. From Berners-Lee (2000), Brodaric (2007) and Ma et al. (2011)

information on the Web, definitions and representations of concepts and the interrelationships among concepts are needed (Berners-Lee, 2006). In the Semantic Web such definitions and representations are called ontologies. Each ontology is the formal specification of the shared conceptualization of a domain of study (Gruber 1995). In practice, ontologies can be of different forms, such as glossary, controlled vocabulary, conceptual schemas and detailed logic constraints, depending on the level of detail on conceptual specification. Semantic Web technologies provide the essential elements for modeling and encoding ontologies in machine-readable formats.

In the context of cross-disciplinary program with datasets from various resources and subjects and researchers from different knowledge domains, there could be a large number of ontologies addressing the various needs on knowledge engineering and concept representation. Those ontologies can be implemented to build innovative functions to support the discoverability, accessibility, understandability and usability of open data. For example, there can be projects on categorizing datasets and publications based on their subjects and keywords, recommending datasets or publications to a user based on his research interests, suggesting matches between datasets and scientific questions, and more. The data science domain recently also has proposed the topic “smart data” (Sheth 2014), which aims at using Semantic Web technologies to improve the efficiency in the transformation from massive datasets into actionable information.

34.2.3 Methodology of Deploying Data Science in Geoscience

Although data science has already attracted significant attention in both academia and the industry, the theoretical foundations and technological systems of data science are still under development. In the summary report of a recent NSF-funded workshop (Drineas and Huo 2016), the emergence of data science as a discipline was compared to the rise of computer science in the 1950s along with the wide availability of computers, especially personal computers (PCs). The data deluge of today and its great potential for academia and industry are, in the report authors’ language, a “forcing function” that will catalyze the emergence of data science departments in universities and nurture the development of data science as a discipline. At the current time, since we do not have established theoretical foundations for data science, we can understand the core of data science as a cross-disciplinary topic, or a blend of massive datasets with methodologies in existing disciplines, such as computer science, library and information science, statistics and mathematics. The application of data science will further extend the coverage of disciplines to other domains, such as geoscience.

In most scientific researches, including those in geoscience, a general research process includes the following steps: (1) Choose a general direction and do

background research; (2) Generate a hypothesis; (3) Conduct experiments and collect data; (4) Analyze data and revise hypothesis; (5) Communicate results. We can compare those steps with the data science process in Fig. 34.2. Both processes follow a direction of data collection, data analysis and result communication, but there are also a few items worthy of further discussion. First, data science often faces a situation in which massive datasets are already in existence while we do not yet have a hypothesis. Second, the data science process addresses a step called data pre-processing, which detects the inconsistent, incomplete and incorrect parts in the datasets and takes actions to ensure the data quality before doing analysis. Data pre-processing is an essential step for large datasets collected from multiple sources. Third, the step of exploratory data analysis (EDA) offers clues for hypotheses in scientific research. EDA is a widely-used approach in statistics, and it covers many methods, such as scatterplot, box plot, residual plot, smoother, bag plot, and more (Brillinger 2011). The term “exploratory” explains the purpose of the method: it is flexible and can help look for things that we believe are not there or to be there (Tukey 1977). EDA helps address the shortage of research hypotheses for massive data that already exist. The functionality of EDA is comparable to the approach of data-driven abductive discovery (Hazen 2014). Abduction means the formation of a plausible explanation for an observation. Charles S. Pierce (1839–1914) viewed abduction as the first stage of scientific reasoning, i.e. to create a hypothesis. Then deduction will be carried out to determine the specific evidence needed to prove the hypothesis. After that, induction will be used to extrapolate a general rule or principle from the findings. Hazen (2014) summarized that abduction is to discover what we do not know we do not know, while deduction and induction are to discover what we know we do not know. This is comparable to Tukey’s point of view on EDA (Tukey 1977).

One of the most significant challenges to deploy data science in geoscience is to create a space (physical and/or virtual) and establish an approach so that researchers from different disciplines can talk to each other. Science of today is highly compartmented into disciplines and there are considerable gaps between these, as reflected by differences in scientific subjects, research methods, terminologies used and even styles of working. The challenge of cross-disciplinary collaboration is like encouraging people to step out from their “comfort zones”. Researchers in geoinformatics (Fox and McGuinness 2008; Ma et al. 2014b) have proposed a method called use case-driven iterative approach, and have successfully implemented it to facilitate the collaboration between data scientists and domain scientists in several projects. Each use case is a description of the process of a focused task. It can be used to identify scientific questions to ask, resources to be used to answer these questions and methods to be implemented to determine the answer. Through the documentation and analysis of a use case, data scientists and domain scientists (e.g. geologists) can understand the needs and aims of each other. As each use case is a focused small task, the collaborative team can spend a relatively short time to achieve the goal, and then can review, update and move on to the next use case. The process is iterative until the overall objective of a research program is realized.

34.3 Case Studies of Data Science in Geoscience

When applying data science to leverage current geoscience studies, the focus or highlight can consist of one or a few steps, depending on the target aimed at. For example, the target can be improving data discoverability and accessibility by updating building blocks and frameworks in the cyberinfrastructure. It can also be focused on finding patterns within massive datasets such as those from literature legacy or crowd-sourcing databases. In this section, a few recent efforts and case studies will be introduced.

34.3.1 *Coordinating Standards to Improve Data Interoperability*

In the domain of geoscience, a few recent achievements on data standards and their implementation were led by CGI-IUGS (<http://www.cgi-iugs.org>), the Commission for the Management and Application of Geoscience Information within the International Union for Geological Sciences. GeoSciML was proposed as a markup language for the exchange of general geoscience information on the Web (Sen and Duffy 2005). GeoSciML was built on top of the Geography Markup Language (GML) and the eXploration and Mining Markup Language (XMML). The first geoscience subjects covered in GeoSciML included boreholes and structural geology. Raw datasets such as those in geologic maps can be transformed into GeoSciML formats once the mapping between the original data structure and the GeoSciML schema is set up. This makes it easier for data exchange and sharing among organizations and nations. GeoSciML was successfully implemented in the OneGeology project (Jackson and Wyborn 2008). On the front end of the OneGeology data portal (<http://portal.onegeology.org>), users can access geologic map services in a standard data structure. At the back end of the portal, there are multiple data providers, distributed data servers and different data structures. GeoSciML acts as a mediator between those heterogeneous structures and improves the data interoperability. Another significant contribution from CGI-IUGS is the multi-lingual geoscience vocabularies. Initial projects on geologic time and rock type vocabularies were applied in the OneGeology-Europe project to harmonize geologic maps from around 20 European countries (Laxton et al. 2010). Standards derived from those vocabularies also became a part of INSPIRE, the Infrastructure for Spatial Information in Europe (<http://inspire.jrc.ec.europa.eu>).

Such efforts on data standards are an essential part of informatics, especially applied informatics that has a domain specific background. Comparing with the geoscience community at large, the number of people working on geoinformatics is low. The value and gains that data standard work can provide are often not fully understood within the geoscience community (Jackson and Wyborn 2008). The situation has been changing in recent years since the value of data science was

recognized by increasingly more geoscience researchers. For instance, besides GeoSciML, CGI-IUGS also has developed EarthResourceML for the exchange of information on mineral occurrences, mines and mining activity. CGI-IUGS's Terminology Working Group has published additional standardized vocabularies. The geoscience community has also collaborated with standard organizations to improve the visibility of data standard outputs. In 2017, GeoSciML was published as a standard of the Open Geospatial Consortium (OGC) (OGC 2017), making it one of the first domain-specific standards in OGC. Geoinformatics researchers also take the lead in coordinating data standards among different scientific disciplines. In 2016, CODATA, the International Council for Science's Committee on Data for Science and Technology, set up a task group on coordinating data standards amongst scientific unions (<http://www.codata.org/task-groups/coordinating-data-standards>). The aim of the group is to take stock of the progress on disciplinary data standards in different scientific unions, recognize the best practices and coordinate the development of future work. Data standards provide the basic-level technical support when we collect and analyze datasets in cross-disciplinary projects. They significantly reduce the workload on data pre-processing and data cleansing in a data science process (Fig. 34.2).

34.3.2 Openness, Provenance and Reproducibility of Research

Provenance and reproducibility are both regarded as important research topics in data science (Drineas and Hou 2016), and they are also essential parts of open science. The literal meaning of provenance is the origin of something. In data science, documenting provenance involves the annotation and interconnection of a network of research activities, people, organizations and resources involved in the production of scientific findings (Ma et al. 2014a). In 2013, the Semantic Web community released an ontology called PROV-O (Lebo et al. 2013). The three top classes Entity, Activity and Agent in PROV-O are easy to understand. The ontology also covers a list of subclasses and relationships that can be applied in domain specific applications. A recent successful implementation of PROV-O is the Global Change Information System (GCIS) (Tilmes et al. 2013), which is part of the U.S. Global Change Research Program (USGCRP, <http://www.globalchange.gov>). USGCRP is a multi-agency research program to "assist the Nation and the world to understand, assess, predict, and respond to human-induced and natural processes of global change." Every four or five years, USGCRP releases a National Climate Assessment Report with the latest scientific findings on different aspects global change. The most recent one was released in 2014. The initial aim of GCIS is to present the 2014 report and to incorporate integrated access to interlinked resources underpinning that report. The long-term goal of GCIS is to be a web-based source of authoritative, accessible, usable and timely information about global change.

Semantic Web technologies, including PROV-O, were applied in the design and development of GCIS. The project included four major parts: categorization, annotation, identification and linking (Ma et al. 2014a), which are coherent within the architecture of the Semantic Web (Berners-Lee 2000). With the well-documented provenance information on the GCIS website (<https://data.globalchange.gov>), users will be able to conduct innovative research on provenance tracing data mining. For example, they can seek answers for the question: What is NASA's contribution to the sea-level rise scenarios in the 2014 National Climate Assessment Report?

Reproducibility in data science and open science includes at least two levels of meaning. The first is replicability of a research output by using the datasets and methods in the research. The second is the derived value, which means the open datasets and methods from that research can be reused in new research and make substantial contributions (Beaulieu et al. 2017). To improve the reproducibility of scientific research, several technical frameworks can be applied and/or adapted, such as workflow platforms and provenance documentation. In a recent study about reproducible marine ecosystem assessment (Ma et al. 2017), the PROV-O ontology was extended and implemented in the Jupyter Notebook (<http://jupyter.org>) to capture and interconnect information from various resources in a scientific research project. Jupyter Notebook is an open-source web application that can be used to create workflow documents with codes, formulas, tables, diagrams, interactive visualizations and descriptive text. The developed ontology further enhanced the function of the platform in capturing and presenting scientific provenance information. The work was used in the Ecosystem Assessment Program of the U.S. NOAA Northeast Fisheries Science Center to support assessment reports of Large Marine Ecosystems. In the implementation, a user works within the Jupyter Notebook to write codes and text for data input, analysis, output and documentation. Once the notebook is completed, the provenance information is automatically captured using the structure defined in the ontology. The collected provenance information is machine-readable and can be archived for later use, such as verifying steps and outputs in the workflow or retrieving raw datasets used in any given step.

34.3.3 Leveraging Geoscience Data Legacy for New Discovery

Geoscience is a domain with abundant literature resources, and much useful information can be extracted from the data legacy. A recent study, originally called PaleoDeepDive (Peters et al. 2014) and now GeoDeepDive (<https://geodeepdive.org>), has demonstrated the significant value of geoscience publication archives through the application of machine learning and data mining technologies. The domain of focus in GeoDeepDive is paleontology and its aim is to detect and extract fossil occurrence information from the massive scientific literature. The work leverages methods in natural language processing, entity recognition and extraction and knowledge graph

construction to improve the efficiency of document processing and the quality of output datasets. In several complicated data extrication and reasoning tasks, the outputs of GeoDeepDive were comparable to the results collected by human experts of geologic history (Peters et al. 2014). Most recently, several publishers and research organizations have set up partnerships with GeoDeepDive and provided a huge number of publications for processing. By middle April 2017, the team has already processed more than 3.2 million documents. The extracted fossil records and their interrelationships can provide useful updates to existing databases, such as the Paleobiology Database (PBDB, <https://paleobiodb.org/>). PBDB, in turn, has set up interfaces and libraries such as those for Web-based data query and retrieval (Peters and McClennen 2015) and the R environment (Varela et al. 2015). These projects build up channels through which any geoscience researcher can easily access datasets of interest and integrate them with other datasets in their own projects.

A project ongoing in the author's group is about using an ontology to help integrate datasets from PBDB with geologic map services provided by USGS and, thus, to build an enriched data portal where users can discover and access more information. Previous works already have shown the functionality of ontology and data visualization in geoscience data services (Ma et al. 2012). In the ongoing project the focus is an ontology for the regional geologic time scale of North America, in addition to the established ontology for the global geologic time scale (Cox and Richard 2015). The geologic time scale of North America has unique classification and terminology for the time intervals at the Epoch and Age levels; for the levels of Eon, Era and Period it shares the architecture with the global standard. As the terminology in the regional standard has been used in geoscience research of the North American region, specific terms in the regional standard can now also be used as keywords in data search, such as in queries sent to PBDB. In the ontology for the regional geologic time scale of North America, detailed information on all time intervals and their relationships were captured and represented in a machine-readable format. A Web-based visualization was then developed for the ontology, and interactive functions were developed to deploy the visualization as a control panel for data search. When a user clicks a time term in the panel, a query will be sent to PBDB, and the retrieved fossil records from PBDB will be plotted in a map window. Our project also set up connections to the USGS data services, so the user can load geologic map layers onto the map window and browse the background geologic information of a location where a fossil was discovered. The multi-source information has the potential to stimulate discussion among users and help them propose new research questions.

34.3.4 Cross-Disciplinary Collaboration for Innovative Discoveries

In early 2015, a research project focused on the co-evolution of geo- and biospheres was kicked off at the Carnegie Institution of Washington (<http://dtdi.carnegiescience.edu>).

The researchers in that project are from several universities and institutions and are with diverse knowledge backgrounds, making the research a real cross-disciplinary collaboration. The project proposed to deploy a data-driven abductive approach to discover patterns in the evolution of Earth's environment. A major task in the early stage of the project is to set up a Deep-Time Data Infrastructure (DTDI), which includes the enrichment of attributes (e.g. age information) in existing geo- and bio-databases, connections among geo-databases of petrology, mineralogy and geochemistry, the linkage between geo- and bio-databases, and open access and dissemination protocols for the built data infrastructure. Many open access data resources were considered for DTDI, including ruff.info (mineral species and properties), mindat.org (mineral species and localities), earthref.org (geochemistry and geomagnetism), geokem.com (igneous rock chemistry), metpetdb.rpi.edu (metamorphic petrology), earthchem.org (geochemistry, geochronology, petrology), vamps.mbl.edu (subsurface microbial ecosystem), pdb.org (protein structures), paleobiodb.org (paleobiology), and more. The user case-driven iterative method mentioned in Sect. 34.2.3 has been implemented to organize meetings and promote collaborations among researchers in the group. While the project is still ongoing, several interesting findings have already been achieved. One of them is the pattern of Large Number of Rare Events (LNRE) among the mineral species frequency distribution (Hystad et al. 2015). The work used the records of mineral species, localities and observations (species-locality pairs) from mindat.org and discovered the LNRE pattern. By extrapolating the domain of observation to be about four times the current size, the result in the LNRE model showed that there are about 1,500 new mineral species to be discovered. From that work, further studies on the population probabilities of all mineral species lead to the characterization of Earth-like planets, such as the Mars (Hystad et al. 2017).

As an affiliated scientist in the project mentioned above, the author led a project of using data visualization to study the co-relationships between mineral-forming elements and mineral species. The first study focused on a list of 30 key elements chosen by the research team (Ma et al. 2016). First, we built a $30 \times 30 \times 30$ matrix and visualized it in a three-dimensional coordinate system, which made the matrix a fundamental framework to fill in records. Along each axis in this matrix we plotted the same arranged list of 30 elements as indices. Each cell in the matrix was first filled with the raw number of minerals in which elements X, Y, and Z coexist. A color spectrum was then applied to render each cell according to the value of the number in it. The process was intuitive, and the output in the three-dimensional matrix already showed interesting patterns in the co-relationships between elements and minerals. The visualized matrix was developed to be interactive in a web browser. Researchers can rotate the matrix and zoom into see details of a part, highlight a certain cell and see attributes in it, and slice one or more planes out from the matrix to see two-dimensional patterns. In another study, we extended the scale to all the 72 mineral-forming elements and constructed a $72 \times 72 \times 72$ matrix. We then applied a chi-squared test to generate values to be filled and visualized in that matrix (Hummer et al. 2016). The mineralogical research question in that

study was “Does the presence of element Z affect the correlation between elements X and Y in mineral species, and is the effect positive or negative?” Besides the completed case studies, many other interesting projects can be further developed with the three-dimensional matrix. For example, we can add data on electronegativity, ionic radius, atomic number, period, crustal abundance, etc. as associated parameters to each axis and test for different clustering of elements based those parameters.

34.4 Concluding Remarks

Mathematical geosciences are now in an intelligent stage. As a research domain, mathematical geosciences share many topics in common with the data science of today. A topic of great interest in deploying data science for geoscience is how to generate research questions or hypotheses when massive datasets are already in existence. In this chapter, the role of exploratory data analysis was analyzed for that purpose, and it was compared with the data-driven abductive approach. Semantic Web and Open Data create a freshly new data environment for conducting geomathematical studies. The Web is built as an open space where Anyone can say Anything on Any topic. The Semantic Web aims to facilitate data Interoperability on the Web, to improve Interactivity between humans and machines, and to inspire Intercreativity for exploring new things. For informatics, a major objective is to present the Right information to the Right person in the Right way. We can use the acronym AIR3 to represent those nine words with initial capital letters. AIR3 presents a broad vision of deploying data science for geoscience in the context of the Semantic Web and Open Data. To put this into practice, we need to create a physical and/or virtual space and implement an approach where researchers from different disciplines can step out from their ‘comfort zones’, talk to each other, and collaborate on focused research topics.

Acknowledgements This work was partly supported by W. M. Keck Foundation, the National Science Foundation (NSF) through the NSF Idaho EPSCoR Program (award number IIA-1301792) and by the University of Idaho ORED 2017 Seed Grant Program.

References

- Agterberg F (2014) Geomathematics: theoretical foundations, applications and future developments. Springer, Cham, Switzerland, 553 pp
- Beaulieu SE, Fox PA, Di Stefano M, Maffei A, West P, Hare JA, Fogarty M (2017) Toward cyberinfrastructure to facilitate collaboration and reproducibility for marine integrated ecosystem assessments. *Earth Sci Inf* 10(1):85–97

- Berg-Cross G, Cruz I, Dean M, Finin T, Gahegan M, Hitzler P, Hua H, Janowicz K, Li N, Murphy P, Nordgren B, Obrst L, Schildhauer M, Sheth A, Sinha K, Thessen A, Wiegand N, Zaslavsky I (2012) Semantics and ontologies for EarthCube. In: Janowicz K, Kessler C, Kauppinen T, Kolas D, Scheider S (eds) *Proceedings of the workshop on GIScience in the big data age 2012*, Columbus, OH, 5 pp
- Berners-Lee T (2000) Semantic web on XML. Presentation at XML 2000 conference, Washington DC. <http://www.w3.org/2000/Talks/1206-xml2k-tbl>. Accessed 15 July 2013
- Berners-Lee T (2006) Linked Data. <http://www.w3.org/DesignIssues/LinkedData.html>. Accessed on 02 July 2013
- Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. *Sci Am* 284(5):34–43
- Bonham-Carter GF (1994) *Geographic information systems for geoscientists: modeling with GIS*. Pergamon, Elsevier Science Ltd., Kidlington, UK, 398 pp
- Brillinger DR (2011) Exploratory data analysis. In: Badie B, Berg-Schlosser D, Morlino L (eds) *International encyclopedia of political science*. SAGE Publications, Thousand Oaks, CA, pp 530–537
- Brodaric B (2007) Geo-pragmatics for the geospatial semantic web. *Trans GIS* 11(3):453–477
- Cox SJ, Richard SM (2015) A geologic timescale ontology and service. *Earth Sci Inf* 8(1):5–19
- Drineas P, Huo X (2016) NSF workshop report: theoretical foundations of data science (TFoDS), Arlington, VA, 20 pp. http://www.cs.rpi.edu/TFoDS/TFoDS_v5.pdf
- Easterbrook SM (2014) Open code for open science? *Nat Geosci* 7(11):779–781
- Fox P, McGuinness DL (2008) TWC semantic web technology. http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology. Accessed on 14 Apr 2017
- Freeden W (2010) Geomathematics: its role, its aim, and its potential. In: Freeden W, Nashed MZ, Sonar T (eds) *Handbook of geomathematics*. Springer, Berlin, Heidelberg, pp 3–42
- Future Earth (2014). Future earth 2025 vision. Future earth, Paris, France, 6 pp. http://www.futureearth.org/sites/default/files/future-earth_10-year-vision_web.pdf
- Gruber TR (1995) Toward principles for the design of ontologies used for knowledge sharing. *Int J Hum-Comput Stud* 43(5–6):907–928
- Hazen RM (2014) Data-driven abductive discovery in mineralogy. *Am Miner* 99(11–12):2165–2170
- Hey T, Payne MC (2015) Open science decoded. *Nat Phys* 11(5):367–369
- Hummer DR, Hazen RM, Ma X, Golden JJ, Downs RT, Liu C, Morrison SM, Meyer M (2016) Quantifying and visualizing earth's mineral chemistry through geologic time. GSA Annual Meeting, Denver, Colorado, USA
- Hystad G, Downs RT, Hazen RM (2015) Mineral species frequency distribution conforms to a large number of rare events model: prediction of earth's "missing" minerals. *Math Geosci* 47:647–661
- Hystad G, Downs RT, Hazen RM, Golden JJ (2017) Relative abundances of mineral species: a statistical measure to characterize earth-like planets based on earth's mineralogy. *Math Geosci* 49(2):179–194
- Jackson I, Wyborn L (2008) One planet: OneGeology? The Google Earth revolution and the geological data deficit. *Environ Geol* 53(6):1377–1380
- Laxton J, Serrano JJ, Tellez-Arenas A (2010) Geological applications using geospatial standards—an example from OneGeology-Europe and GeoSciML. *Int J Digit Earth* 3(S1):31–49
- Lebo T, Sahoo S, McGuinness D (2013) PROV-O: the PROV ontology. <http://www.w3.org/TR/prov-o/>
- Loudon TV (2000) *Geoscience after IT: a view of the present and future impact of information technology on geoscience*. Elsevier, Oxford, 142 pp
- Ma X (2015) Geoinformatics in the semantic web. In: Schaeben H, Delgado RT, van den Boogaart KG, van den Boogaart R (eds) *Proceedings of IAMG 2015*, Freiberg, Germany, pp 18–26

- Ma X, Asch K, Laxton JL, Richard SM, Asato CG, Carranza EJM, van der Meer FD, Wu C, Duclaux G, Wakita K (2011) Data exchange facilitated. *Nat Geosci* 4(12):814
- Ma X, Beaulieu SE, Fu L, Fox P, Di Stefano M, West P (2017) Documenting provenance for reproducible marine ecosystem assessment in open science. In: Diviacco P, Glaves HM, Leadbetter A (eds) *Oceanographic and marine cross-domain data management for sustainable development*. IGI Global, Hershey, PA, USA, pp 100–126
- Ma X, Carranza EJM, Wu C, van der Meer FD (2012) Ontology-aided annotation, visualization and generalization of geological time scale information from online geological map services. *Comput Geosci* 40(3):107–119
- Ma X, Fox P, Tilmes C, Jacobs K, Waple A (2014a) Capturing and presenting provenance of global change information. *Nat Clim Change* 4(6):409–413
- Ma X, Zheng JG, Goldstein J, Zednik S, Fu L, Duggan B, Aulenbach S, West P, Tilmes C, Fox P (2014b) Ontology engineering in provenance enablement for the National Climate Assessment. *Environ Model Softw* 61:191–205
- Ma X, Hummer D, Hazen RM, Golden JJ, Fox P, Meyer M (2016) Showing co-relationships between elements and minerals in a three-dimensional matrix. *GSA Annual Meeting*, Denver, Colorado, USA
- McNutt M, Lehnert K, Hanson B, Nosek BA, Ellison AM, King JL (2016) Liberating field science samples and data. *Science* 351(6277):1024–1026
- Merriam D (2004) The quantification of geology: from abacus to pentium: a chronicle of people, places, and phenomena. *Earth Sci Rev* 67(1–2):55–89
- National Research Council (2008) *Origin and evolution of Earth: research questions for a changing planet*. National Academies Press, Washington, DC, 150 pp
- Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon T, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T (2015) Promoting an open research culture. *Science* 348(6242):1422–1425
- OGC (2017). OGC Geoscience Markup Language 4.1 (GeoSciML). Open Geospatial Consortium, 234 pp. <http://www.opengeospatial.org/standards/geosciml>
- Peters SE, Zhang C, Livny M, Ré C (2014) A machine reading system for assembling synthetic paleontological data-bases. *PLoS One* 9:e113523. <https://doi.org/10.1371/journal.pone.0113523>
- Peters SE, McClennen M (2015) The paleobiology database application programming interface. *Paleobiology* 42(1):1–7
- Ramachandran R, Rushing J, Li X, Kamath C, Conover H, Graves S (2006) Bird's-eye view of data mining in geosciences. In: Sinha AK (ed) *Geoinformatics: data to knowledge*, geological society of America special papers, 397. The Geological Society of America, Boulder, CO, pp 235–247
- Reitsma F, Albrecht J (2005) Modeling with the semantic web in the geosciences. *IEEE Intell Syst* 20(2):86–88
- Schutt R, O'Neil C (2013) *Doing data science: straight talk from the frontline*. O'Reilly Media, Inc. Sebastopol, CA, 375 pp
- Sen M, Duffy T (2005) GeoSciML: development of a generic geoscience markup language. *Comput Geosci* 31(9):1095–1103
- 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 2014 IEEE 30th international conference on data engineering (ICDE)*, Chicago, IL, pp 2–2
- Tilmes C, Fox P, Ma X, McGuinness D, Privette AP, Smith A, Waple A, Zednik S, Zheng J (2013) Provenance representation for the national climate assessment in the global change information system. *IEEE Trans Geosci Remote Sens* 51(11):5160–5168

- Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading, PA, 688 pp
- Varela S, González-Hernández J, Sgarbi LF, Marshall C, Uhen MD, Peters S, McClennen M (2015) paleobioDB: an R package for downloading, visualizing and processing data from the Paleobiology Database. *Ecography* 38(4):419–425

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

