International Journal of
Geo-Information

MDPI

Article

Using Visual Exploratory Data Analysis to Facilitate Collaboration and Hypothesis Generation in Cross-Disciplinary Research

Xiaogang Ma [1,*], Daniel Hummer [2], Joshua J. Golden [3], Peter A. Fox [4], Robert M. Hazen [5], Shaunna M. Morrison [5], Robert T. Downs [3], Bhuwan L. Madhikarmi [1], Chengbin Wang [1,6] and Michael B. Meyer [5]

[1] Department of Computer Science, University of Idaho, 875 Perimeter Drive, MS 1010, Moscow, ID 83844-1010, USA; madh9981@vandals.uidaho.edu (B.L.M.); cwang@uidaho.edu (C.W.)

[2] Department of Geology, Southern Illinois University Carbondale, 1263 Lincoln Drive, Carbondale, IL 62901, USA; daniel.hummer@siu.edu

[3] Department of Geosciences, University of Arizona, 1040 E. 4th Street, Tucson, AZ 85721, USA; jgolden@email.arizona.edu (J.J.G.); rdowns@email.arizona.edu (R.T.D.)

[4] Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA; pfox@cs.rpi.edu

[5] Geophysical Laboratory, Carnegie Institution for Science, 5251 Broad Branch Road, NW, Washington, DC 20015, USA; rhazen@carnegiescience.edu (R.M.H.); smorrison@carnegiescience.edu (S.M.M.); mmeyer@carnegiescience.edu (M.B.M.)

[6] State Key Laboratory of Geological Processes and Mineral Resources & Faculty of Earth Resources, China University of Geosciences, 388 Lumo Road, Wuhan 430074, China

[*] Correspondence: max@uidaho.edu; Tel.: +1-208-885-1547

Received: 6 October 2017; Accepted: 15 November 2017; Published: 16 November 2017

Abstract: Massive open data resources are changing the way that people do science. To make use of those data resources, data science methods and technology can be leveraged by stakeholders of various disciplines. The objective of this paper is to present our experience of using visual exploratory data analysis as a method to facilitate collaboration and hypothesis generation in geoscience research. The research team consisted of both geoscientists and computer scientists. A use case-driven, iterative approach was applied to create a collaborative and communicative environment. Through several rounds of use case analysis and technological development, a data visualization pilot system was created for studying the co-relationships between chemical elements and mineral species. The exploratory data analyses conducted in those use case studies led to several research hypotheses for future work. This research illustrates the usefulness of exploratory data analysis for hypothesis generation in a data science process. Although the presented project is in geoscience, the discussed method and experience can also be translated into other disciplines.

Keywords: exploratory data analysis; data visualization; data science; geoinformatics; mineral ecology

1. Introduction

The open data movement is changing the way that people do science [1–3]. A conventional process of scientific research begins with background study and hypothesis generation. Then data will be collected in experiments and the results of data analysis will be used to approve or revise the hypothesis. With abundant datasets made freely accessible through the open data movement, researchers can now retrieve massive datasets from the open data environment on the Web [4]. However, researchers often struggle to develop hypotheses despite the abundance of data available to them. In this new era of science, methods and tools are desired to help researchers generate and test hypotheses.

Studies in data science can provide methods to address this challenge. Data science is the study of the generalizable extraction of knowledge from data [5]. The theoretical foundations of data science have strong connections to the disciplines of mathematics, statistics, computer science, and more [6]. In the field of statistics, the method of exploratory data analysis (EDA) is used as a step for hypothesis generation before the step of confirmatory data analysis (CDA) (i.e., statistical hypothesis testing) [7,8]. In recent years, EDA has been suggested by data scientists [3] as an effective step for pattern recognition and hypothesis generation in a data science process (Figure 1). The term "exploratory" represents the characteristics of the method: The EDA process is flexible and the result is uncertain, so it can be used to search for characteristics that are believed to be present or absent [9].

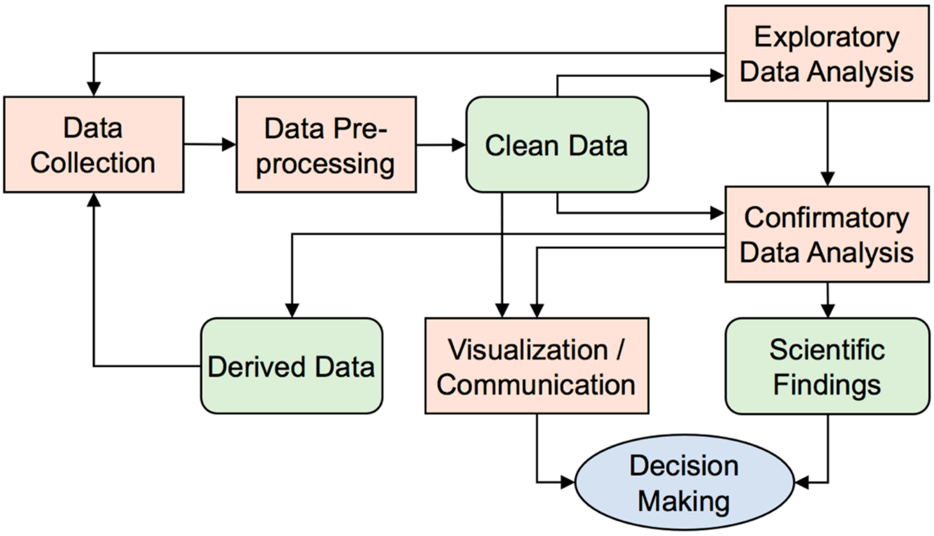


Figure 1. Key steps in a data science process (adapted from [3]).

The data science process featured in EDA is comparable to the approach of data-driven abductive discovery [10–12]. Abduction is defined as the formation of a plausible explanation for an observed phenomenon [13]. Charles S. Peirce (1839–1914) viewed abduction as the first stage of scientific reasoning, i.e., to create a hypothesis [14]. Following abduction, deduction is carried out to refine the hypothesis based on other plausible premises and to determine the specific evidence needed to support the hypothesis. Finally, induction is used to extrapolate a general rule or principle from the findings. Abduction and deduction are a part of the conceptual understanding of a phenomenon, and induction is the quantitative verification. Ho [10] used a short sentence to summarize the interactions among the three concepts: "Abduction creates, deduction explicates, and induction verifies". This process fits well with the two steps of EDA and CDA in the data science process (Figure 1). For a domain-specific study that aims to leverage abduction and the data science process, Hazen's summary [11] might also be useful: deduction and induction are to discover what we know we do not know, and abduction is to discover what we do not know we do not know.

Data visualization is an efficient way to display the results of a data science process [15,16]. In recent years, researchers have also proposed that data visualization should be applied in each step of the data science process rather than only for the end product [17]. In EDA for statistics, data visualization is an essential part of quantitative datasets. Many visualization techniques have already been developed, such as scatter plots, box plots, histograms, stem-and-lead plots, and more. For EDA in cross-disciplinary studies, other types of visualization techniques may also be applied, such as mind maps [18], conceptual maps [19], and workflow visualizations [20]. Through the usage of those visualization techniques, researchers from different backgrounds can quickly obtain an overview of the subject under study, gain insights about the datasets, and discuss hypotheses for the focused work of the next step [21,22].

The domain of Earth and space sciences, like other disciplines, faces opportunities raised by open data, and requires methods and technologies to help transform massive amounts of data into meaningful information [23,24]. The objective of this paper is to present our experience of applying visual EDA to facilitate cross-disciplinary research collaboration and hypothesis generation. A few use cases of applying a three-dimensional matrix to show co-relationships among chemical elements and mineral species will be used to demonstrate the collaborative process. The increasing complexity of datasets and research discussions along with those use cases also reflects the effectiveness of this method for formulating hypotheses. The presented use cases are from studies of mineral evolution in the Deep Time Data Infrastructure (DTDI) [25], a research initiative that joins data science with geoscience and bioscience to study the co-evolution of Earth systems. The remainder of this paper is organized as follows: Section 2 describes the data sources of this study, the methods of visual EDA, and the design of a workflow; Section 3 presents a demonstration system that implements the designed workflow, and also demonstrates the usefulness the system through a few focused use cases; Section 4 highlights several research topics in the data science process and lists a few topics for future work; and, finally, Section 5 concludes the paper.

2. Datasets and Methods

More than 5000 mineral species have been discovered on Earth. Each mineral species is a natural chemical compound characterized by a definite crystalline structure. Through the studies of chemistry, physical properties, crystal structure, and geographical distribution of those mineral species, the geoscience community has built many reusable data resources. For example, the database of Raman spectroscopy, X-ray diffraction and chemistry of minerals (RRUFF) [26] aims at creating and sharing a complete set of high-quality spectral data from well-characterized minerals. The collected data [27] provide a standard for structural, spectroscopic, and chemical mineral identification, and can be used in studies of Earth and other planets. RRUFF also hosts a continually updated list of mineral names that are officially accepted by the International Mineralogical Association (IMA) and the detailed source information of those minerals [28]. The website of the IMA mineral list [29] provides an interactive user interface that allows users to search the list and the source information in various ways and download for research uses. Another useful data resource is Mindat [30], a crowd-sourced website that collects and shares information about mineral species, their properties, and their geographic distribution on Earth.

The abundant datasets about minerals and their properties have initiated new ideas and studies in recent years. DTDI is an integrated program that leverages various existing data sources to discover patterns in the evolution of Earth's environment, including the geosphere and biosphere. One of the umbrella research themes in DTDI is mineral evolution—the mineralogy of terrestrial planets and moons evolves as a consequence of a range of physical, chemical, and biological processes that lead to the formation of new mineral species [31]. In the past few years, several new findings have been reported. One of them is the pattern of Large Number of Rare Events (LNRE) in the frequency distribution of mineral species [32]. By extrapolations from the LNRE model, researchers can predict how many new mineral species can be discovered at an assumed larger observation size. Going further from that work, studies on the population probabilities of all mineral species have led to the characterization and comparison of Earth-like planets [33,34].

To leverage more studies with those open data mineral resources, we designed and developed a pilot system that can be used to support EDA in the multidisciplinary data science process (Figure 1) of the above-mentioned mineral evolution research. Our idea was to construct a three-dimensional (3D) matrix to visualize co-relationships among mineral-forming chemical elements and mineral species found on Earth. The three axes in this matrix, X, Y, and Z, were identical lists of arranged chemical elements. A simple example was to list 30 key mineral-forming elements along each axis. This 30 × 30 × 30 3D matrix resulted in 27,000 cells, in which we could assign different values, such as the raw number of minerals in which elements X, Y, and Z co-exist. If each cell was rendered in a color according to the value of the number inside it, then the 3D matrix could reveal patterns in the co-relationships among elements and minerals, such as clusters of high mineral species numbers for the element triplets F-Si-O, Na-Si-O, Mg-Si-O, Al-Si-O, F-Al-O, Na-Al-O, and Mg-Al-O. We also developed functions to manipulate the matrix, so that a user could rotate the matrix, zoom in and out, select and highlight certain cubes or patterns, and slice one or more two-dimensional planes out from the matrix to see patterns of interest. Those detected patterns may lead to the formation of research hypotheses for further works, such as why oxygen has the highest number of mineral species among all mineral-forming elements. Such a visualization system is easy to understand and operate for both geologists and data scientists. It lowers the barrier of communication between collaborators, and facilitate discussion on research topics.

The multidisciplinary collaboration in DTDI follows the data science steps shown in Figure 1. The pilot system played an important role in the EDA step. Before carrying out EDA, data collection and data pre-processing were conducted by DTDI team members who were familiar with the subject, structure, and format of datasets in RRUFF, the IMA mineral list, and Mindat. The resulting clean data were well-organized in a sample structure, which saved a great deal of time for data science team members when it came to loading and visualizing the data in the 3D matrix of the pilot system. A few meaningful visualization outputs from the EDA could be published as research results directly. The case studies in the next section will illustrate a few visualization outputs of this kind, such as the co-relations between primary and secondary cobalt minerals shown in Figure 5. Another relevant DTDI research of network analysis and visualization [35] also revealed a similar EDA approach but applied different techniques. Those visualization results and recognized patterns were used in research discussions and to support decision-making. Derived datasets could be published, shared, and reused in other research (i.e., another round of the data science process).

3. Implementation and Case Studies

The team that conducted this research consisted of geoscience and computer science researchers with complementary academic backgrounds in minerology, paleontology, data management, data visualization, and data analysis. A use case-driven iterative approach [36,37] was applied throughout the whole work to facilitate the interactions among team members. Several use case studies were conducted in this research with an iterative process. For computer scientists, this iterative approach helped refine the functions of the developed pilot system because each use case had unique datasets and data visualization requirements. For geoscientists, the information revealed through the EDA of each use case was meaningful and led to the discussion of more research topics. We used comma separated values (CSV) as the file format for the dataset, and reused a JavaScript library three.js [38] to develop the visualization. The current pilot system was made accessible online [39]. The source code and datasets of the demo system were shared on Github [40].

Our first use case was the co-existence of key elements in minerals. The objective of this case study was to examine the correlation between triplets of elements by counting the number of mineral species in which those three elements co-exist. By plotting the same list of 30 key mineral-forming elements along each axis of a 3D coordination system, we constructed a 30 × 30 × 30 matrix. We then referred to the RRUFF and the IMA mineral list to find the numbers of minerals in which elements X, Y, and Z coexist, and filled those numbers into the corresponding 27,000 cells in the 3D matrix. Subsequently, we developed a color spectrum according to the range of the numbers in the matrix, and applied the spectrum to the matrix to render each cell with a color. Figure 2a shows an initial output from the first use case. Subsequently, the geoscientists team members offered suggestions on how to make the visualization outputs easier to operate and more meaningful from the geoscience perspective. By using the developed functions, geoscientists could manipulate the 3D matrix, sliced out planes, conduct transformations, and observed the clustering patterns (Figure 2b–e). Through this use case, the basic visualization toolkits were set up. Although the clusters of high values in the matrix clearly demonstrate there are more minerals for certain elements, the mineralogists in the research team wanted a deeper view of the relationship. This led to the second use case.

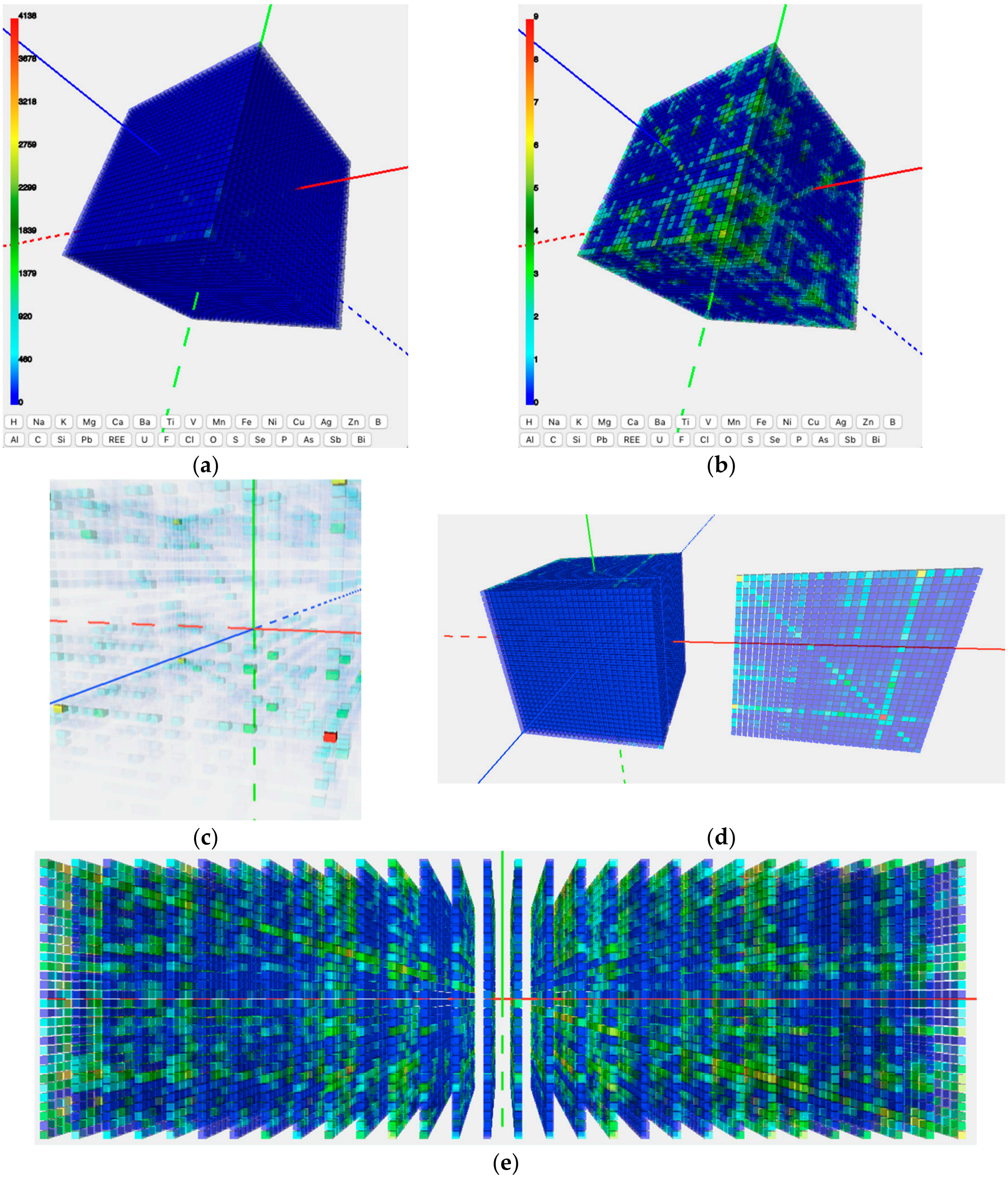
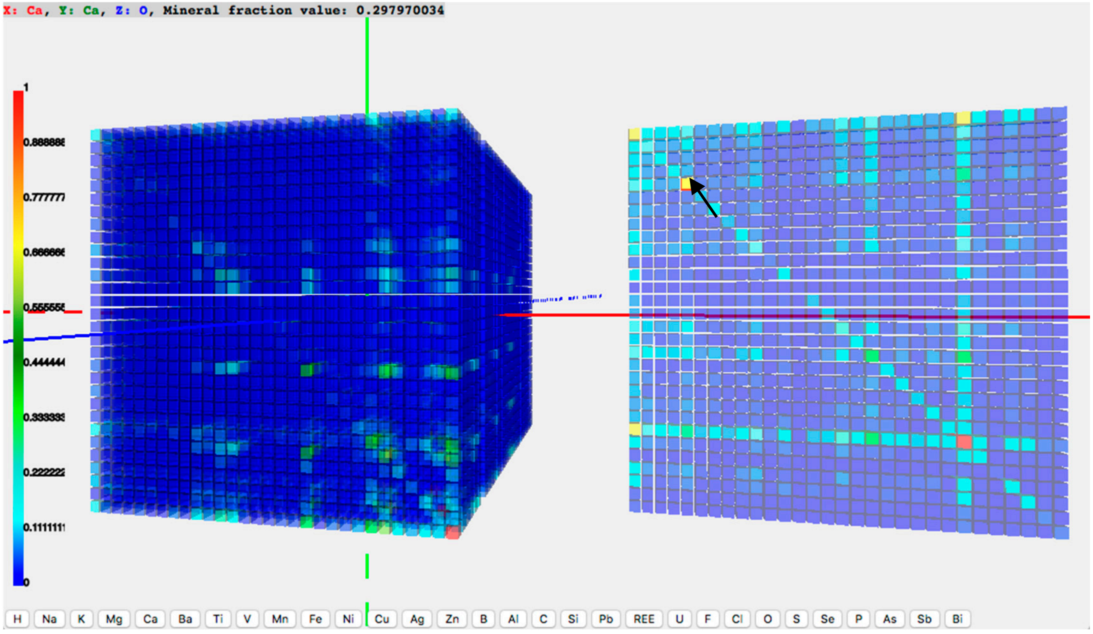

4138
3678
3218
2759
2299
1839
1379
920
460
0
H Na K Mg Ca Ba Ti V Mn Fe Ni Cu Ag Zn B
Al C Si Pb REE U F Cl O S Se P As Sb Bi
(a)
9
8
7
6
5
4
3
2
1
0
H Na K Mg Ca Ba Ti V Mn Fe Ni Cu Ag Zn B
Al C Si Pb REE U F Cl O S Se P As Sb Bi
(b)
(c)
(d)
(e)

Figure 2. Pilot system for the exploratory data analysis of co-relationships among elements and minerals. (**a**) An initial output by visualizing the raw mineral counts; (**b**) output after taking a logarithmic calculation of the mineral counts in each cell; (**c**) changes in the opacity of each cell based on the value of the mineral counts. The cell filled with solid red (lower right) has oxygen on all three axes. It has the highest mineral count, 4138, in the whole matrix; (**d**) sliced-out two-dimensional planes to see the patterns. Here it shows a plane for oxygen, i.e., oxygen is the element on the Z-axis; and (**e**) changing the distance between cells along one or more axes to see patterns in a two- or one-dimensional context.

The second use case was a small research topic initiated by the visualization output of the first use case. It had the same objective as the first use case to show the co-existence of elements in mineral species, but with updated datasets. In the first use case, the value in each matrix cell was the mineral counts. In the second use case, the dataset was replaced by one in which the cell values represented the fraction of minerals containing an element on the Z-axis that also contain both X- and Y-axes. A new function developed in the pilot system was to show attributes of a matrix cell when the cursor is placed over it. In Figure 3, the plane of oxygen is sliced out such that oxygen is the Z element for all cells on the plane. When a user moves the cursor over the cells in the plane, the cell below the cursor is highlighted and the attributes of that cell will be shown on top of the 3D matrix in the browser window. In Figure 3, the shown attributes read 'X: Ca, Y: Ca, Z: O, Mineral fraction value: 0.297970034'. This means that about 29.8% of minerals containing oxygen also contain calcium. After finished the first two use cases with the 30 key mineral-forming elements, the research team decided to expand the scope of the dataset, and move on to all 72 mineral-forming elements.



```
X: Ca, Y: Ca, Z: O, Mineral fraction value: 0.297970034
```

Figure 3. Use a 'mouse over' operation to see attributes of a matrix cell. The cell below the cursor is highlighted, and the attributes of the cell is shown on the top of the window. The value '0.297970034' means that about 29.8% of minerals containing oxygen also contain calcium.

In the third use case, we expanded the dataset to cover all 72 mineral-forming elements. Correspondingly, a 72 × 72 × 72 matrix was constructed with a same list of 72 elements along each axis. Instead of filling raw mineral numbers, we used a chi-squared test to generate values in the 373,248 cells of the 3D matrix. The aim of those values is to answer the question 'Does the presence of element Z affect the correlation between elements X and Y in mineral species?' For example, in Figure 4 the rows of red and blue cells corresponding to the O-H plane highlight different elements' association with hydrated minerals. The Z axis, representing all the elements pairing with O and H, is shown in dark blue. Cells that are colored red represent elements that correlate strongly to O–H bearing minerals, and cells colored blue represent elements that are anti-correlated to O–H bearing minerals. These results indicate that some elements are very common in hydrated mineral species, while others are rarely found in hydrated minerals. This is an entirely new result gained from this use case, and leads geoscientists to new questions regarding what causes an element to associate with hydrated minerals.

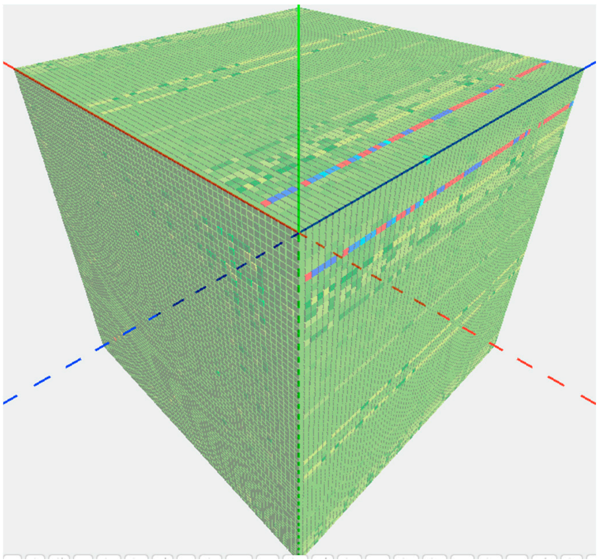
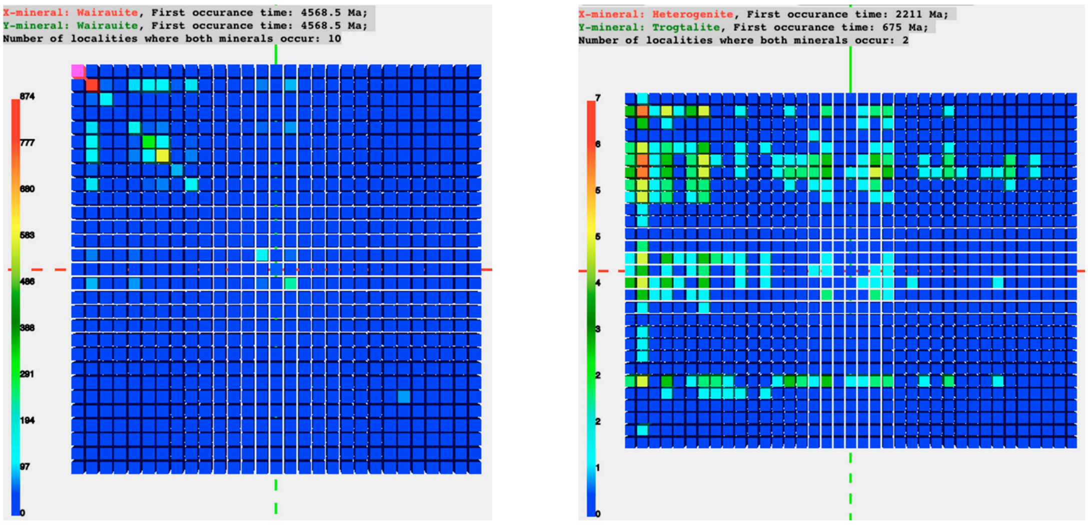


Figure 4. Visualization of a 72 × 72 × 72 matrix in the developed pilot system. The rows of red and blue cells are corresponding to the O-H plane, and they highlight different elements' association with hydrated (i.e., O–H bearing) minerals.

The above three use cases helped the team develop most of the functions in the pilot system. With minor adaption to the code, the system was also used to visualize and analyze datasets in a few other use cases. One of them was the study of co-relations between primary and secondary cobalt (Co) minerals. A primary mineral is any mineral formed during the original solidification (crystallization) of the host igneous rock. A secondary mineral is any mineral that forms later through processes such as hydrothermal alteration and weathering. In this use case, the raw datasets were collected from Mindat and the IMA mineral list, and were organized in a two-dimensional matrix. Figure 5 shows the visualization output from the pilot system. Rows of higher values in Figure 5b show a clear correlation of certain secondary Co minerals arising with certain primary Co minerals, and at certain geologic time. This type of previously unrecognized correlation is of great interest to geoscientists for further research.



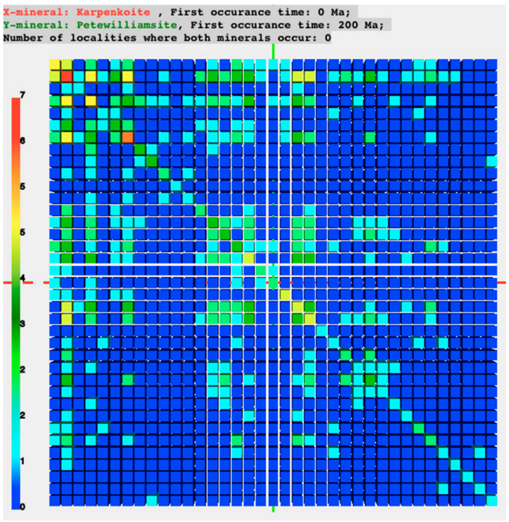
```
X-mineral: Wairauite, First occurance time: 4568.5 Ma;
Y-mineral: Wairauite, First occurance time: 4568.5 Ma;
Number of localities where both minerals occur: 10
```

(**a**)

```
X-mineral: Heterogenite, First occurance time: 2211 Ma;
Y-mineral: Trogtalite, First occurance time: 675 Ma;
Number of localities where both minerals occur: 2
```

(**b**)

Figure 5. *Cont.*



(**c**)

Figure 5. Visualization outputs showing co-relations between primary and secondary cobalt minerals. Minerals are arranged by their first occurrence time (old to young: left to right along the horizontal X-axis; top to bottom along the vertical Y-axis). The raw value in each cell represents the number of localities at which both the X and Y minerals occur. (**a**) Primary to primary (raw values); (**b**) primary (Y) to secondary (X), with logarithmic values; and (**c**) secondary to secondary, with logarithmic values.

4. Discussion

Our experience of developing and using the pilot system in the DTDI project demonstrates the usefulness of visual EDA for facilitating hypothesis generation in cross-disciplinary collaboration. Data visualization leverages human's visual ability to detect complex relationships in data that are difficult to reveal through numbers and text. Quick prototypes and outputs in the EDA increase the chance to see unexpected discoveries. Through the use case-driven, iterative approach, geoscientists and computer scientists from different disciplinary backgrounds had a context to communicate and could work together on focused topics. The accumulated small works could lead to useful technology or a tool that can be reused, such as the pilot system in our work [39] and the shared code and dataset [40]. With several rounds of EDA case studies, the research team obtained a better understanding of the underlying data structure and were able to choose appropriate models and plan future data collection.

The results of the presented use cases led to new questions and hypotheses for researchers in geoscience. For example, one use case revealed that elements are divided into those that strongly favor hydrated minerals versus those that do not. Since water is considered a volatile constituent in minerals, this result leads to questions about other volatile elements. Can elements be sorted into groups based on correlation or anti-correlation with fluorine? Can they be sorted based on correlation or anti-correlation with chlorine? Do these divisions tell us something new about the sorting of elements in geochemical environments? These are all examples of research hypotheses that arise from the EDA techniques presented here.

Although this research was in the domain of geoscience, the function of visual EDA in a data science process and the experience of the use case-driven, iterative approach can also be translated into other disciplines. The way people do science is being changed by massive open and/or proprietary data resources. Researchers of various disciplines can benefit from the visual EDA for hypothesis generation. In addition to the application in cross-disciplinary contexts, the visual EDA can also be applied to intra-disciplinary applications through a data science process. In general, data science helps transform raw data into meaning and understanding [41]. Small and focused use cases help researchers understand the datasets, choose the research question, and efficiently collaborate on data analysis [42]. In addition to the changes in hypothesis generation, the data science process in an open data environment also has a few other characteristics. Since the datasets are collected from different resources, there could be heterogeneities in the data format, conceptual structure, and even the terminology. A step of data pre-processing or data wrangling [43] is needed to transform the raw datasets into an organized form that is appropriate for EDA and CDA. Data pre-processing was a very important step in our research as we had raw datasets from three sources: RRUFF, the IMA mineral list, and Mindat. Several team members worked together to find connections among those datasets, build a data structure to host the connected records, and transform it into the CSV format. The EDA in our work focused on quick prototype and visualization output. The well-organized dataset from the pre-processing allowed quick visualization outputs through very easy operations, so the team could have more time to focus on analyzing patterns in the result. Some clues for the EDA were generated in the step of data pre-processing when we were discussing the connections between data resources. We then reflected the discussed idea in the data structure and visualized the dataset in the EDA step.

A few future research topics can be proposed. To facilitate more interactive collaboration in a cross-disciplinary research team, we can leverage virtual or augmented reality in the visual EDA, such as the Microsoft HoloLens or the Computer Animated Visualization Environment (mini-CAVE). The developed 3D matrix pilot system and the conducted use case studies also resulted in a few new research hypotheses. We can calculate the expected numbers of minerals with X + Y + Z based on average crustal abundances. By comparing the observed and expected numbers, we will be able to estimate the extent to which the element triplets occur with greater or lesser frequency than would be expected. In the 3D matrix, the arranged elements on each axis can have multiple associated parameters. For example, we can add data on atomic number, ionic radius, period, electronegativity, crustal abundance, and more. By using those parameters, we can order elements along the three axes automatically to test different clustering of elements. The value in each cell of the 3D matrix can also represent other properties besides the mineral counts. Furthermore, using cation and anion oxidation states instead of chemical elements on the axes may allow us to see dramatic correlations based on redox.

5. Conclusions

Earth and space science, like many other disciplines, are facing opportunities and challenges raised by the open data environment. The large and growing number of datasets freely accessible on the Web requires scientists to change their way of working. They need to deploy efficient methods for hypothesis generation to make better use of the open data. The step of exploratory data analysis in a data science process can be leveraged to meet that need. In this paper, we presented our experience of using visual exploratory data analysis to facilitate collaboration and hypothesis generation in a cross-disciplinary research project. The scientific topic of the research was the co-relationship among chemical elements and mineral species. The research team consisted of both geoscientists and computer scientists. The successful use case studies, as presented in the paper, show the effectiveness of the visual exploratory data analysis. Although our work is in the domain of geoscience, the discussed methods and experience can also be translated into other disciplines.

Acknowledgments: This work presented in the paper was partially supported by the W. M. Keck Foundation through the grant "The Co-Evolution of the Geo- and Biospheres: An Integrated Program for Data-Driven Abductive Discovery in Earth Sciences", the National Science Foundation (NSF) through the NSF Idaho EPSCoR Program (no. IIA-1301792), and the University of Idaho ORED Seed Grant. Additional support was provided by the Deep Carbon Observatory, the Alfred P. Sloan Foundations, an anonymous foundation, and the Carnegie Institution for Science. We thank four anonymous reviewers for their constructive comments on an earlier version of this manuscript.

Author Contributions: Xiaogang Ma, Peter Fox, Robert M. Hazen and Daniel Hummer designed the work plan. Joshua J. Golden, Daniel Hummer and Robert T. Downs contributed and pre-processed the datasets. Xiaogang Ma and Bhuwan L. Madhikarmi developed the demo systems. Shuanna M. Morrison, Chengbin Wang and Michael B. Meyer participated the discussion during the use case analyses. All authors contributed to the manuscript writing and revising.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cutcher-Gershenfeld, J.; Baker, K.S.; Berente, N.; Flint, C.; Gershenfeld, G.; Grant, B.; Haberman, M.; King, J.L.; Kickpatrick, C.; Lawrence, B.; et al. Five ways consortia can catalyse open science. *Nature* **2017**, *543*, 615–617. [CrossRef] [PubMed]
2. Kitchin, R. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*; Sage: London, UK, 2014; 222p.
3. Schutt, R.; O'Neil, C. *Doing Data Science: Straight Talk from the Frontline*; O'Reilly: New York, NY, USA, 2013; 406p.
4. Phethean, C.; Simperl, E.; Tiropanis, T.; Tinati, R.; Hall, W. The role of data science in web science. *IEEE Intell. Syst.* **2016**, *31*, 102–107. [CrossRef]
5. Dhar, V. Data science and prediction. *Commun. ACM* **2013**, *56*, 64–73. [CrossRef]
6. Drineas, P.; Huo, X. *NSF Workshop Report: Theoretical Foundations of Data Science (TFoDS)*; TFoDS workshop organizing committee: Arlington, VA, USA, 2016; 20p. Available online: http://www.cs.rpi.edu/TFoDS/TFoDS_v5.pdf (accessed on 26 September 2017).
7. Brillinger, D.R. Exploratory Data Analysis. In *International Encyclopedia of Political Science*; Badie, B., Berg-Schlosser, D., Morlino, L., Eds.; SAGE Publications: Thousand Oaks, CA, USA, 2001; Volume 1, pp. 530–537.
8. Cox, V. *Translating Statistics to Make Decisions*; Apress: New York, NY, USA, 2017; 324p.
9. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley: Reading, PA, USA, 1977; 688p.
10. Ho, Y.C. 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, LA, USA, 4–8 April 1994; 18p.
11. Hazen, R.M. Data-driven abductive discovery in mineralogy. *Am. Mineral.* **2014**, *99*, 2165–2170. [CrossRef]
12. Fox, P.; Hendler, J. The science of data science. *Big Data* **2014**, *2*, 68–70. [CrossRef] [PubMed]
13. Magnani, L. *Abduction, Reason and Science: Processes of Discovery and Explanation*; Springer: New York, NY, USA, 2011; 205p.
14. Miller, H.J. The data avalanche is here. Shouldn't we be digging? *J. Reg. Sci.* **2010**, *50*, 181–201. [CrossRef]
15. Kraak, M.-J. Exploratory visualization. In *Encyclopedia of GIS*; Shekhar, S., Xiong, H., Eds.; Springer: Berlin, Germany, 2008; pp. 301–307.
16. Tufte, E.R. *The Visual Display of Quantitative Information*, 2nd ed.; Graphics Press: Cheshire, CT, USA, 2001; 197p.
17. Fox, P.; Hendler, J. Changing the equation on scientific data visualization. *Science* **2011**, *331*, 705–708. [CrossRef] [PubMed]
18. Buzan, T. *Mind Map Handbook: The Ultimate Thinking Tool*; HarperCollins: Toronto, ON, Canada, 2005; 431p.
19. Novak, J.D.; Cañas, A.J. *The Theory Underlying Concept Maps and How to Construct and Use Them*; Technical Report IHMC CmapTools; Institute for Human and Machine Cognition: Pensacola, FL, USA, 2008. Available online: http://cmap.ihmc.us/docs/theory-of-concept-maps (accessed on 26 September 2017).
20. Mou, X.; Jamil, H.; Ma, X. VisFlow: A visual database integration and workflow querying system. In Proceedings of the 33rd IEEE International Conference on Data Engineering (ICDE 2017), San Diego, CA, USA, 19–22 April 2017; pp. 1421–1422.
21. Ma, X.; Chen, Y.; Wang, H.; Zheng, J.G.; Fu, L.; West, P.; Erickson, J.S.; Fox, P. Data visualization in the Semantic Web. In *The Semantic Web in Earth and Space Science: Current Status and Future Directions*; Narock, T., Fox, P., Eds.; IOS Press: Berlin, Germany, 2015; pp. 149–167.
22. Ma, X. Linked Geoscience Data in practice: Where W3C standards meet domain knowledge, data visualization and OGC standards. *Earth Sci. Inform.* **2017**, *10*, 429–441. [CrossRef]
23. Steed, C.A.; Ricciuto, D.M.; Shipman, G.; Smith, B.; Thornton, P.E.; Wang, D.; Shi, X.; Williams, D.N. Big data visual analytics for exploratory earth system simulation analysis. *Comput. Geosci.* **2013**, *61*, 71–82. [CrossRef]
24. Ma, X. Geoinformatics in the Semantic Web. In Proceedings of the 17th Annual Conference of the International Association for Mathematical Geosciences (IAMG 2015), Freiberg, Germany, 5–13 September 2015; pp. 18–26.
25. The Co-Evolution of the Geo- and Biospheres. An Integrated Program for Data-Driven, Abductive Discovery in the Earth Sciences. Available online: https://dtdi.carnegiescience.edu (accessed on 26 September 2017).
26. Lafuente, B.; Downs, R.T.; Yang, H.; Stone, N. The power of databases: The RRUFF project. In *Highlights in Mineralogical Crystallography*; Armbruster, T., Danisi, R.M., Eds.; De Gruyter: Berlin, Germany, 2015; pp. 1–30.
27. Database of Raman Spectroscopy, X-ray Diffraction and Chemistry of Minerals (RRUFF). Available online: http://rruff.info (accessed on 26 September 2017).
28. Rakovan, J. Words to the Wise—More than 4,000 To Be Exact. *Rocks Miner.* **2007**, *82*, 423–424. [CrossRef]

29. IMA Mineral List with Database of Mineral Properties. Available online: http://rruff.info/ima/ (accessed on 26 September 2017).
30. Mineralogy Database—Mineral Collecting, Localities, Mineral Photos and Data (Mindat). Available online: https://www.mindat.org (accessed on 26 September 2017).
31. Hazen, R.M.; Papineau, D.; Bleeker, W.; Downs, R.T.; Ferry, J.M.; McCoy, T.J.; Sverjensky, D.A.; Yang, H. Mineral evolution. *Am. Mineral.* **2008**, *93*, 1693–1720. [CrossRef]
32. Hystad, G.; Downs, R.T.; Hazen, R.M. Mineral Species Frequency Distribution Conforms to a Large Number of Rare Events Model: Prediction of Earth's Missing Minerals. *Math. Geosci.* **2015**, *47*, 647–661. [CrossRef]
33. Hazen, R.M.; Grew, E.S.; Downs, R.T.; Golden, J.; Hystad, G. Mineral ecology: Chance and necessity in the mineral diversity of terrestrial planets. *Can. Mineral.* **2015**, *53*, 295–324. [CrossRef]
34. Hystad, G.; Downs, R.T.; Hazen, R.M.; Golden, J.J. Relative Abundances of Mineral Species: A Statistical Measure to Characterize Earth-like Planets Based on Earth's Mineralogy. *Math. Geosci.* **2017**, *49*, 179–194. [CrossRef]
35. Morrison, S.M.; Liu, C.; Eleish, A.; Prabhu, A.; Li, C.; Ralph, J.; Downs, R.T.; Golden, J.J.; Fox, P.; Hummer, D.R.; et al. Network analysis of mineralogical systems. *Am. Mineral.* **2017**, *102*, 1588–1596. [CrossRef]
36. Fox, P.; McGuinness, D.L. TWC Semantic Web Technology. Available online: http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology (accessed on 26 September 2017).
37. Ma, X.; Zheng, J.G.; Goldstein, J.C.; Zednik, S.; Fu, L.; Duggan, B.; Aulenbach, S.M.; West, P.; Tilmes, C.; Fox, P. Ontology engineering in provenance enablement for the National Climate Assessment. *Environ. Model. Softw.* **2014**, *61*, 191–205. [CrossRef]
38. Three.js—JavaScript 3D Library. Available online: https://threejs.org (accessed on 26 September 2017).
39. Demo System for Exploring Co-relationships between Elements and Minerals. Available online: https://goo.gl/FAEepi (accessed on 10 November 2017).
40. Cube Matrix to Show Element-Mineral Co-relationships. Available online: https://github.com/xgmachina/3dcube (accessed on 10 November 2017).
41. Wickham, H.; Grolemund, G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*; O'Reilly Media: Sebastopol, CA, USA, 2017; 522p.
42. Bittner, K.; Spence, I. *Use Case Modeling*; Pearson Education, Inc.: Boston, MA, USA, 2003; 347p.
43. Kandel, S.; Heer, J.; Plaisant, C.; Kennedy, J.; van Ham, F.; Riche, N.H.; Weaver, C.; Lee, B.; Brodbeck, D.; Buono, P. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Inf. Vis.* **2011**, *10*, 271–288. [CrossRef]