Acta Geologica Sinica (English Edition), 2019, 93(supp.1): 17–18

Disruption in Biogeosciences: Conceptual, Methodological, Digital, and Technological

Peter FOX*

Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY 12180,USA

Citation: FOX Peter, 2019. Disruption in Biogeosciences: Conceptual, Methodological, Digital, and Technological. *Acta Geologica Sinica* (English Edition), 93(supp.1): 17-18

Over four years ago, a group of investigators came together to determine if Big Data approaches (specifically data mining, machine learning and analytics in general) might provide insight into some of the grand challenges in Earth's history: evolution of minerals, rise of oxygen, life, influence of super continental cycles, quantifying the magnitude of extinction events, and more. As a result, the team of mineralogists, petrologists, geochemists, paleontologists, marine microbiologists, and data scientists tackled the problem of defining and implementing a deep-time data (science) infrastructure (DTDI) and methodology to study the co-evolution of minerals, fossils, and the role of proteins in Earth's evolution. Discovery and exploration without a clear idea of what we would uncover was a primary and somewhat unnerving goal to members of the team – a potentially very disruptive approach to multi-disciplinary science (Ma et al., 2017a). Not everyone knew how we would make the journey and what we might find, but the outstanding group of early career scientists showed no fear.

At the very beginning it was clear that: a) data was essential to addressing the big science questions, b) much of the data needed was "dark" or "grey" and c) that a means for integrating mineral, fossil and protein data representations of how the geosphere and biosphere co-evolved, was needed. The investigator team had access to, or knowledge of data sources (grey and dark data respectively). The effort for a) and b) was substantial from both the conceptual, methodological and technical aspects. Literature searches, hand entering of data in many cases but also semi-automated optical character recognition and data extraction from tables all contributed to the data corpora assembled (Zhong et al., 2017; Morrison et al., 2017). In conceptualizing the means for integration very diverse data sources; ones that were never conceived or structured to be integrated with other data, several attributes of the data became clear, i.e. Big Data is not just volume. The data-types were numerous and often confounding. One pivotal example for minerals was their "location". Across many of the data sources, e.g. mindat.org, rruff.info, locations varied from a latitude-longitude, to an outcrop name, to a town to a region, or a mine, and the list goes on. Over Earth's history, location becomes a greater consideration, e.g. at continental margins. This space and indeed time became "where" and "when". Much of the ancillary data (often known as metadata) became data. The expanding dimensions and type of the data assembled alone became a major disruption for the domain scientists. To the rescue, however, was the data science and computer science expertise to accommodate integrative data structures and common-place machine learning and data analytics methods such as multi-dimensional scaling. However, this step was just a humble beginning.

Almost all the primary data were resident in tabular formats that was not conducive to the types of integration and analyses needed. The team moved quickly to an underlying graph representation (based on the W3 Resource Description Framework; RDF). However, in migrating the data into to graph form, what became immediately clear was that the visual and analytic tools did not sufficiently render the graphs nor operate on them in analytic sense (Fox et al., 2017), especially in a manner that was accessible to domain scientists. The team explored a slightly less formal representation of the graphs in the form of networks, with initial analogies to citation and social networks popularized in the last decade. Network science (Fox et al., 2018) is an academic field which studies complex networks considering distinct elements or actors represented by nodes (or vertices) and the connections between the elements or actors as links (or edges). Network Analysis methods have been successfully applied in statistical physics, particle physics, computer science, biology, economics, finance, climatology and sociology. But these methods still had not been fully leveraged in many areas of Earth Science especially in visual representations (Ma et al., 2017b, Prabhu et al., 2017). Using networks can help view existing geological and biological information systems from a purely mathematical perspective, and infer new relationships or new information about existing relationships (Muscente et al., 2018, 2019). From a technology perspective facilitating interaction and collaboration both in face-to-face but also remote settings, the team used Jupyter notebooks extensively (http://www.jupyter.org) and the open-source languages and library packages coded in Python and R.

Among the many new discoveries (Morrison et al., 2017; Muscente et al., 2018, 2019), there were yet new questions generated. In geology: What geological characteristics most influence mineral occurrence? In biology: Is the distinctive large number of rare events distribution (not discussed here) of Earth's minerals a biosignature? In mathematics and statistics: What distribution functions best characterize locality/ mineral data? And in computer science: can we create interactive 3D (holographic) mineral networks?

We claim that the end product of the project and collaboration was "resilient disruption". On both sides, domain and data science, there needed to be adaptation. Data scientists continuously fight the 80-20 "rule" - 80% of their time on handling / reworking data and 20% on doing the data science. We

* Corresponding author. E-mail: pfox@cs.rpi.edu

http://www.geojournals.cn/dzxbcn/ch/index.aspx
https://onlinelibrary.wiley.com/journal/17556724

claim that after our disruption that the rule was changed and that 20% went to data handling and 80% to science and discovery. We have yet to quantitatively evaluated this claim.

This presentation conveys what data was needed, the barriers we encountered and the digital and technical solutions/ disruptions that led to discoveries from myriad sources in contrast to conventional approaches in prior discipline science, and more importantly how we spanned discipline boundaries.

Key words: mineral ecology; mineral evolution; fossil networks; mass extinctions; network analytics; visual analytics

Acknowledgments: This work was supported by the WM Keck Foundation, the AP Sloan Foundation and an anonymous foundation. The work of the DTDI team was essential.

References

Fox, P.A., Eleish, A., Li, C., Pan, F.F, Prabhu, A., Zhong, H., 2017. Heterogeneity and Heterarchy: How far can network analyses in Earth and space sciences take us? *EOS*, IN33B-0125.

Ma, X., West, P., Zednik, S., Erickson, J., Eleish, A, Chen, Y., Wang, H., Zhong, H., Fox, P., 2017a. Weaving a knowledge network for deep carbon science. *Frontiers in Earth Science*, 5: 36. DOI: 10.3389/feart.2017.00036.

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., 2017b. Using visual exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research. *ISPRS International Journal of Geo-Information*, 6(11): 368.

Morrison, S.M. Liu, C., Eleish, A., Prabhu, A., Li, C., Ralph, J., Downs, R.T., Golden, J.J., Fox, P., Hummer, D.R., Meyer, M.B., and Hazen, R.M., 2017. Network analysis of mineralogical systems, *American Mineralogist*, 102: 1588-1596.

Muscente, A.D., Prabhu, A., Zhong, H., Eleish, A., Meyer, M.B., Fox, P., Hazen, R.M. Knoll, A.H., 2018. Quantifying ecological impacts of mass extinctions with network analysis of fossil communities. *Proceedings of the National Academy of Sciences*, 115 (20): 5217-5222.

Muscente, A., Bykova, N., Boag, T., Buatois, L., Mangano, G., Eleish, A., Prabhu, A., Pan, F.F., Meyer, M., Schiffbauer, J., Fox, P., Hazen, R., and Knoll, A., 2019. Ediacaran biozones identified with network analysis provide evidence for pulsed extinctions of early complex life. Paper #NCOMMS-18-29230B (in press).

Prabhu, A., Fox, P.A., Zhong, H., Eleish, A., Ma, X.G., Zednik, S., Morrison, S.M., Moore, E.K., Muscente, D., Meyer, M., Hazen, R.M.,2017. Visualizing Complex Environments in the Geo- and BioSciences. *EOS*, IN31D-02.

Zhong, H., Ma, X., Prabhu, A., Eleish, A., Pan, F., Parsons, M., Ghiroso, M., West, P., Zednik, S., Erickson, J.S., Chen, Y., Wang, H., and Fox, P., 2017. Thermodynamic Data Rescue and Informatics for Deep Carbon Science. *EOS*, IN23D-0115.

About the first author

Peter FOX is Tetherless World Constellation Chair, Professor of Earth and Environmental Science, Computer Science and Cognitive Science, and Director of the Information Technology and Web Science Program at Rensselaer Polytechnic Institute. Fox has a B.Sc. (hons) and Ph.D. in Applied Mathematics (physics and computer science) from Monash University. Fox research includes computational and computer science; ocean and environmental informatics; and distributed semantic data frameworks, with applications to large-scale distributed data science investigations. http://tw.rpi.edu/web/person/PeterFox.