THE 4D INITIATIVE: DEEP-TIME DATA-DRIVEN DISCOVERY

VISION FOR A 4D INITIATIVE

In this document we explore the proposed 4D Initiative—an international collaboration for scientific discovery in the field of planetary evolution based on the analysis and visualization of multi-disciplinary, multi-dimensional data. The organization and activities of the 4D Initiative will differ from those of traditional academic departments or data science centers. Thousands of discipline scientists from hundreds of collaborating institutions will take the lead, employing data science methods to foster discovery through cross-disciplinary, multi-dimensional strategies. We propose to harness the power of abduction to accelerate the discovery of what we don't know we don't know.

ABDUCTION: A STRATEGY FOR DATA-DRIVEN SCIENTIFIC DISCOVERY

Discoveries in Earth, space, and life sciences rely to a significant degree on induction and deduction—classic approaches to reasoning that focus on the observation, modeling, and ultimately predictive explanations of known patterns and phenomena in nature. Knowledge emerges from targeted measurements and observations that are made in the context of established principles or testable, predictive hypotheses about the natural world. These powerful methods have proven successful in documenting and comprehending many aspects of the natural world, but they are inherently inefficient at discovering new complex patterns that increasingly involve multi-parameter or multi-dimensional analysis of large datasets or synthesis of diverse types of data. Consequently, recognition of such gradual global processes as biological evolution by natural selection (*1,2*), continental evolution by plate tectonics (*3,4*), atmospheric and ocean oxygenation by photosynthesis (*5,6*), and climate change (*7,8*) required decades of integrated data synthesis by experts preceding discovery and acceptance of critical Earth phenomena.

What are the next great conceptual breakthroughs awaiting discovery? How can we accelerate discovery of as yet hidden patterns in nature? Today, Earth, space, and life sciences benefit from ever-expanding data resources in numerous disciplines—data that for the most part serve the needs of focused communities of researchers. However, the potential now exists for a revolutionary integration and synthesis of these diverse data resources, leading to an alternative "abductive" approach to investigate Earth's co-evolving geosphere and biosphere. A growing number of scientists thus advocate a systematic, data-driven quest for the accelerated discovery of hidden multi-dimensional patterns in data resources from varied, interconnected disciplines (*9-14*). Today's scientific enterprises generate terabytes per day of new data, yet these extensive resources are woefully underutilized because they are not linked into a single platform (see, however, *12*). The 4D Workshop, as recorded in this document and the Workshop website (www.4d-workshop.net), examined needs and opportunities for accelerating scientific discovery by expanding and linking existing and new data resources, as well as employing methods for coupling integrated data platforms with statistical analysis and visualization capabilities. Thus, we envision a new kind of open-access "scientific instrument" that will transform what we know about Earth in its cosmic context.

Deduction, Induction, and Abduction: Most scientific discoveries arise from two complementary modes of logical reasoning. On the one hand, deductive reasoning begins with a general premise that is asserted to be true, and then draws specific inferences from that generalization that must also be true. Thus:

- Earth's atmospheric oxygenation influenced the partitioning of redox-sensitive elements.
- Molybdenum, rhenium, nickel, and cobalt are redox-sensitive elements.
- Therefore, we conclude by deduction that atmospheric oxygenation must have influenced the partitioning of Mo. Re, Ni, and Co.

In deduction, specific conclusions represent a subset of the initial general premise. Studies of the partitioning of redox-sensitive elements are thus conducted in the context of well-established physical and chemical principles and are not expected to yield surprising or anomalous results that contradict the original premise. Such efforts are critical to providing a solid foundation for scientific progress by filling in gaps in what we know we don't know, but they do not usually represent the most efficient path to discovering fundamentally new phenomena.

The complementary inductive mode of reasoning begins with observations of particular instances of a generalization, which then lead to predictions of further instances of the generalization (or to the generalization, itself). Thus:

- Each of the last 5 supercontinent cycles led to episodes of enhanced mineralization during intervals of continental convergence.
- B, Be, Hg, and Mo are mineral-forming elements.
- Therefore, we predict by induction that B, Be, Hg, and Mo minerals will display enhanced mineralization during intervals of continental convergence.

Unlike deduction, the specific predictions of induction are not necessarily contained within the initial premise and thus they cannot follow with certainty. Because one starts with instances of a generalization, and not an established premise, opportunities for discovering unexpected or anomalous patterns may be enhanced. Thus, for example, an anomalous absence of Hg mineralization during the assembly of the Mesoproterozoic Rodinian supercontinent at 0.9 to 1.3 Ga (*15*) parallels emerging data from other studies (e.g., *16,17*). The tradition in Earth sciences (and crime novels) of collecting data to discriminate amongst multiple working hypotheses (*18*) is inherently inductive in nature, and remains a powerful strategy for discovery.

Most research in the Earth, space, and life sciences is firmly grounded in deduction and/or induction. Most investigators, most of the time, start with an established deductive premise or an inductive generalization consistent with observations about known phenomenon, and then collect new data to test the validity of one or more explanatory hypotheses or to develop new hypotheses.

These deductive and inductive efforts stand in contrast to "abduction", which is a form of logical inference that begins with the accumulation of reliable data independently of a premise or generalization. Analysis of these data, including statistical "data mining" approaches, then point to previously unrecognized patterns and correlations, often in multi-dimensional space not easily represented on an "X-Y" plot, and ultimately to the development of potentially new hypotheses to explain those patterns. Discoveries that lead to "paradigm shifts"—for example, James Hutton's recognition of gradual geological change and deep time (*19*), Charles Darwin's

elucidation of evolution by natural selection (*1*), and the collective development of the concept of plate tectonics (*3*)—tend to be intrinsically abductive in character, even if the initial collection of data was motivated in a deductive/inductive context. Each of these transformative discoveries required synthesis and integration of vast amounts of diverse data resources accumulated over decades to articulate a new framing of the natural world. Abduction thus provides a pathway to discovering what "we don't know we don't know."

Data-Driven Discovery

For most of the history of science, abduction has proven a difficult and time-consuming path to discovery. A lifetime of meticulous data collection and thoughtful synthesis, at times amplified by creative intuition or blind luck, may be required to recognize previously hidden higher-dimensional patterns in diverse data. Only through decades of intimacy with observations, and recognition of subtle quirks and idiosyncrasies in data, will some significant correlations emerge from the noise. Such abductive discoveries do not easily come to the impatient or distracted researcher, which serves as an important justification for the support of dedicated specialists who devote their lifetimes to a focused scientific pursuit.

The development of large and expanding data resources, coupled with powerful computation methods, has the potential to change the nature of abductive scientific discovery. Advances in cyberinfrastructure are poised to integrate data from numerous sources into semantically cohesive data platforms (*10, 20-25*). Furthermore, new and widely available statistical methods and visualization procedures are providing the means to interrogate these data resources in varied ways and thus to tease out subtle correlations that are otherwise inherently invisible to the human brain (26-30).

Such data mining and discovery efforts that exploit large databases and enhanced data interrogation techniques to seek patterns are much in the news, particularly with respect to investment (*31*) and national security (*32*) applications. In science and medicine, new data resources also have the potential to reveal previously unrecognized phenomena. For example, seismological data (from nuclear test ban verification efforts), coupled with ocean floor topography, geochronology of ocean basalts, and paleomagnetism data, were critical in the discovery of patterns that elucidated mechanisms of plate tectonics (e.g., *3*). Today, analyses of hidden patterns in genome databases to map viral evolution (*33,34*) and statistical exploration of medical records to find potential causal factors in pervasive diseases (*35,36*) represent growing applications of abductive strategies.

The premise of this 4D prospectus is that similar opportunities await researchers in the Earth, space, and life sciences, particularly in the context of the complex co-evolving geosphere and biosphere. For example, scientists who focus on Earth materials have accumulated vast amounts of data on rocks, minerals, and geofluids, including their major element, minor element, and isotopic compositions; optical, electrical, magnetic, elastic, and other physical properties; atomic structures, as well as the variations of those structures with pressure, temperature, and composition; petrologic context and associated minerals; their ages; thermochemical parameters and phase relations; tectonic settings and geologic context; and even their mineral-hosted microbial ecosystems. These data are complemented by resources on the evolution of

paleoatmospheres and paleooceans, geomicrobiology, paleontology, genomics and proteomics, paleotectonics, paleomagnetism, and observations of other terrestrial planets and moons, as well as their varied host stars. The ultimate goal of data-driven discovery is to capitalize on and enhance the numerous efforts underway toward a small set of interoperable platforms that offer access to multiple, heterogeneous dimensions in a new "data space." Armed with such resources, we can envision a time when integrated data resources provide the key for discovering, exploring, and understanding numerous complementary aspects of Earth's evolution in space and time.

In spite of the promise of data-driven discovery, pitfalls abound. Data resources must be approached with a firm grounding in chemical and physical principles; an awareness of the meaning, quality, and sources of the data employed; and a keen sense of intuition. Synthesis of unreliable or biased data from varied sources may lead to false or misleading trends. For example, geochemical data employing different analytical instruments or standardization procedures may display subtle systematic differences (*37,38*). Other sources of bias reflect logistical factors: Recent studies of mineral distributions in space and time (e.g., *15,39*) are invariably biased by the proximity of the most scrutinized deposits to major academic institutions, as well as the preferential description of colorful and well crystalized minerals of economically valuable elements. Therefore, any interrogation of integrated data resources must be undertaken within a framework of established deductive and inductive pursuits.

Brute-Force Use Cases: Integrated data resources for abductive discovery in planetary evolution do not yet exist. However, based on recent "brute-force use cases," we can be confident that previously unrecognized patterns and correlations will emerge from the thoughtful integration and evaluation of reliable data. Brute-force use cases involve time-consuming, manual accumulation of relevant data, either through literature searches or acquisition of new measurements. Such efforts have been undertaken in many facets of the Earth, space, and life sciences. Consider a recent example from the field of "mineral evolution"—studies of variations in the diversity and distribution of the minerals of beryllium and mercury through deep time, which demonstrate the potential of this concept as a means to recognize tectonic patterns; search for critical resources; generate insights regarding the evolution of ocean and atmospheric chemistry; and document subtle ongoing feedbacks among terrestrial life, weathering, soils, and climate.

Hazen et al. (*15*) surveyed 128 mercury mineral localities, including the earliest known occurrences for 89 of the 90 known Hg species—a study that required examination and synthesis of data in more than 400 references in a dozen languages. This brute-force effort led to the discovery of two unexpected and previously hidden patterns. First, the ages of almost all Hg mineral localities correlate with 4 episodes of supercontinent assembly. This dataset, assembled with minimal preconceptions about what trends might emerge, revealed four significant episodes of Hg mineralization at approximately at 2.69 ± 0.04, 1.81 ± 0.05, 0.53 ± 0.05, and 0.32 ± 0.07 Ga. Similar trends are now emerging from data-intensive studies of granite pegmatites (*39,40*), as well as the ages of many thousands of individual detrital zircon crystals—large datasets that add robustness to the interpretation of episodic mineralization over at least the past 3 billion years (*41-47*). However, an as yet unexplained billion-year gap in Hg mineralization occurred between 1.8 and 0.8 Ga, an interval that includes the assembly of the Rodinian

supercontinent. This interval may correlate with the innovation of microbial Hg methylation, perhaps coupled with changes in ocean chemistry (*48*). Alternatively, these data on Hg minerals may contribute to growing evidence that the tectonic setting of Rodinian assembly differed from that of other supercontinents (*17,49,50*).

Second, the largest known Hg deposits, formed at ~0.3 Ga, are coeval with Carboniferous coal measures, suggesting co-burial of organic carbon and Hg, with subsequent hydrothermal mobilization and re-deposition. Thus, the mineralization of Hg—a rare element with no biological function—has been shown to be coupled to the evolving terrestrial biosphere.

This study was based entirely on published and web resources, yet it consumed more than 1 person-year of effort, mostly devoted to locating and evaluating previously published data in sometimes obscure sources, as well as integrating those data with lists of dozens of minerals approved by the International Mineralogical Association (*51*; rruff.info/ima) and thousands of Hg mineral locality data (principally in mindat.org).

This and other abductive studies, though focused on single rare elements with relatively few localities, demonstrate the untapped potential for a new strategy of discovery based on development and mining of enhanced data resources. Such brute-force studies of Earth's near-surface rocks and minerals are by no means limited to the temporal occurrence and aerial distribution of mineral species. For example, Farquhar et al. (*52-54*) collated extensive data on sulfur isotopic fractionation in varied lithologies versus age. They found remarkable mass-independent effects, a presumed consequence of upper atmosphere photolysis of sulfur compounds, that are largely constrained to formations > 2.25 Ga. They ascribed this finding to enhanced ozone shielding—a conclusion heralded as the "smoking gun" for the timing of the Great Oxidation Event and its irreversible, biologically-induced transformation of atmospheric chemistry (*6*).

Large-scale community data resources, such as EarthChem/PetDB (*55,56*), are especially relevant for discoveries in statistical petrology, geochemistry, and mineralogy (see http://www.earthchem.org/citations/petdb for examples). In one such effort, Keller and Schoene (*11*) employed a database of 70,000 analyses of continental igneous rocks to discover evidence for significant lithospheric disruption at ~2.5 Ga—a time just prior to the Great Oxidation Event. Their comprehensive overview of secular variations in major and incompatible elements in basalt reveals a significant decrease in mantle melt fraction at that time—a trend not obvious without a large and relatively unbiased data set. Keller and Schoene concluded that atmospheric oxidation may be linked in part to redox changes associated with crustal evolution. The availability of large-scale, community supported, persistent, and quality-controlled data resources is critical to the success of such endeavors.

Accumulations of mineral data also point to Earth's gradual subsurface oxidation. Golden et al. (*57*) gathered new and published trace element analyses of the rhenium content of 422 molybdenite (MoS_2) specimens from 135 localities with known ages from 2.91 billion years to 6.3 million years. Rhenium is a redox-sensitive element that is mobilized in its Re^{7+} form only under relatively oxidized subsurface conditions. This brute-force data effort revealed two statistically significant trends: (1) Systematic increases in average and maximum trace concentrations of Re in molybdenite since 3.0 Ga point to enhanced oxidative weathering by subsurface fluids, and (2)

episodic molybdenum mineralization correlates with five intervals of supercontinent assembly from ~2.7 Ga (Kenorland) to 300 Ma (Pangaea).

These and other examples demonstrate that brute-force methods have the potential to reveal hidden correlations among diverse data in multiple dimensions. Numerous other trends in Earth, space, and life deep-time data are undoubtedly awaiting discovery. However, brute-force data recovery methods are inherently time-consuming and correspondingly inefficient. A far better strategy is to develop and further enhance community supported data resources.

NEEDS AND OPPORTUNITIES IN DEEP-TIME DATA-DRIVEN DISCOVERY

Participants in the 4D Workshop enthusiastically agreed that we are poised to benefit from data-driven discovery, both by focusing on seldom explored interfaces among traditional science disciplines and by exploiting analytical and visualization methods that interrogate data in multiple dimensions. In this pursuit we are faced with many needs and opportunities. The critical recurrent theme of our discussions was the recognition that data-driven discovery relies on promoting ways of thinking about natural systems by employing and analyzing many attributes simultaneously. Such higher-dimensional thinking is both an opportunity and a challenge. On the one hand, numerous as yet unrecognized patterns must lie within higher-dimensional Earth, space, and life science data; however, the human mind is not well suited to visualize patterns in systems much greater than three dimensions. Accordingly, the 4D Workshop identified many promising scientific challenges that might be elucidated by application of analytical and visualization methods that foster abductive discovery in higher dimensions.

A strategy for data-driven discovery is emerging (*58,59*). Three parallel types of advances are necessary to make significant progress. First, we need to identify and tackle key questions in the Earth, space, and life sciences that are inherently multi-dimensional in character. Second, we need to develop and implement powerful analytical and visualization methods to interrogate data in multiple dimensions. Third, we must build and sustain open-access data resources that inform these scientific questions. In the following sections we explore these three essential directions to advance deep-time data-driven discovery.

1. Identify Key Questions in the Earth, Space, and Life Sciences

Evolving natural systems are inherently complex, involving dynamic changes over immense spaces through deep time. The resulting patterns of planetary evolution are particularly suitable for study by the multi-dimensional approach associated with abduction. Participants of the 4D Workshop identified numerous opportunities for scientific investigation, though we recognized that specific research questions will evolve and multiply as each scientist and research team follows inherently unpredictable paths. The promise and power of abduction is that previously hidden correlations in the data will lead us in unanticipated directions as we discover new patterns in multi-dimensional data.

Earth sciences: Opportunities for abductive discovery in the Earth sciences rest in large measure on combining and comparing datasets within the Earth science communities, to tackle critical

questions about the 4.5 billion year evolution of our planet. Planetary evolution is an intrinsically multi-dimensional problem that can be revealed by analyzing numerous attributes of natural systems. For example, geochemical, geophysical, and geological data sets can be integrated and leveraged to address several key questions:

1) How can we recognize and characterize connections among Earth's varied spheres? For example, can we document and understand the intertwined redox evolution of the atmosphere, oceans, and crust?
2) Can we constrain the evolving timing, geometry, and rates of plate tectonics? When did modern-style plate tectonics begin and what was the history of continental crust formation?
3) Can we predict the occurrence of natural resources, for example by employing affinity analysis and data on the distribution of minerals and ores to identify the most promising sites for further exploration?
4) Can we predict the timing and severity, and mitigate effects, of natural hazards by analyzing geochemical, geophysical, and geological data for potential premonitory effects?

Tackling these and numerous other complex geoscience problems will require standardization of many types of data—seismic, geochemical, atmospheric, oceanic, magnetic, Earth materials, and more—as well as coupling these data at different spatial and temporal resolutions. Reliable interpretations will additionally require the development of new methods to quantify error in observational and model data. Armed with well-integrated data, we can begin to tailor further large-scale data collection efforts.

Life sciences: The quest to understand the complex origins and evolution of life on Earth is an inherently multi-dimensional pursuit in space and time, requiring the integration of richly varied physical, chemical, geological, and biological data. Identification of unifying themes in the biosciences is crucial to developing a multi-scale approach—one that facilitates application of data-driven tools across disciplines. Participants of the 4D Workshop pointed to varied unanswered questions that might be elucidated by the abductive strategy.

1) What are the most deeply rooted biochemical pathways—in essence the "fossil biochemistry" that reflects life's most ancient metabolic strategies?
2) How do microbes in consortia interact? Can we understand as yet poorly understood mechanisms, such as syntrophic interactions, and work towards greater understanding of the chemical and physical processes in complex microbial populations?
3) Can we identify a consistent set of thermodynamic rules and metrics to interpret interactions at molecular, cellular, organismal, and environmental scales?
4) To what extent do environments influence gene expression? What relationships exist among microbial metagenomes and associated environmental conditions, including temperature, pH, salinity, chemistry, and other critical parameters?

A recurrent theme in these questions is a critical need to explore microbial ecology, environments, biochemistry, and protein expression—aspects of life that suggest a wide range of research opportunities, including improved algorithms, enhanced metadata, and exploration of microbial interactions from model systems or extreme ecosystems.

Linking the Geosphere and Biosphere: Special opportunities in abductive discovery relate to the co-evolution of the geosphere and biosphere, which share the same physical and chemical environments but are rarely studied in tandem. These two spheres respond to the same thermodynamic rules, the same gradients in redox and temperature, and the same aqueous environmental conditions. Questions that will drive future research include:

1) What are the relative roles of accretionary delivery vs. chemical differentiation of Earth in the origins and evolution of life?
2) How have the sources, distribution, and cycling of Earth's volatiles varied through time, and how have these variations influenced, and been influenced by, life?
3) To what extent have trace metal availabilities affected the evolution of ocean chemistry, minerals, and metalloproteins?
4) How have redox gradients changed in space and time through Earth history and how has the cycling of biologically essential elements varied with time and space, at scales from cellular to global?

In spite of the promise for linking our understanding of the evolution of life and rocks, few examples of geo/bio coevolution (e.g., the Great Oxidation Event) have been explored in detail, owing in large measure to the lack of an accessible, comprehensive, and interoperable deep-time record of geosphere evolution (tectonics, climate, marine chemistry) and biosphere evolution (genome, diversity, ecology, paleontology). A major effort of data integration and curation is thus a key need and opportunity.

Planetary science: Data on planets beyond Earth are as yet sparse. We have modest chemical and physical data on a dozen bodies within our solar system, coupled with minimal information on thousands of exoplanets. Nevertheless, these data, when merged with our much more extensive documentation of Earth, have the potential to lead to important insights regarding planetary evolution. Among the enticing unanswered questions that might be addressed by abductive methods:

1) Can we establish a planetary taxonomy? In particular, does the observed range of planets represent a continuum of types, or are there distinct natural kinds of planets?
2) Do meteorites provide a picture of solar system evolution?
3) How do planets acquire, retain, and cycle volatile elements essential to life?
4) What planetary attributes contribute to habitability?

Fortunately, planetary data tend to be accessible and well curated, thanks to ongoing commitments by NASA and other international space agencies. We are poised to make discoveries as rapidly as new data become available—all the more reason, therefore, to establish open-access methods to explore and integrate diverse data resources.

An Abductive Use Case: Consider one example of a data-driven research project that emerged from the 4D workshop—an effort to reconstruct tectonic processes in deep time that is linked to many aspects of Earth-life coevolution. This approach involves exploiting the four-dimensional GPlates tectonic reconstruction software (GPlates.org), which models Earth's shifting tectonic plates over the past billion years. Our ambition is to link these plate motions to numerous data resources on fossil distributions, mineral and ore deposits, stratigraphy, petrology, geochemistry,

geomagnetics, and other deep-time data. This integration of many types of deep-time data to the familiar visualization platform of evolving Earth holds the potential to link many different types of biosignatures to geological evidence. In so doing, we will provide a paleoenvironmental context to early life, while documenting previously unrecognized patterns in the distribution and diversity of Earth materials. This type of collaborative effort epitomizes the promise of the 4D approach.

2. Identify Needs and Opportunities in Data Science

Data science is revolutionizing the way we approach all fields of science and business. Data scientists at the 4D Workshop recognized important needs and opportunities related to advancing the ways we analyze and visualize multi-dimensional data.

Visualization Methods: Data-driven discovery advances through the implementation of varied analytical and visualization methods. Visualization methods, in particular, allow us to comprehend interrelationships among as many as eight attributes in a single representation (*71*).

1) How can we indicate errors in complex multi-dimensional visualizations, for example uncertainties in plate tectonics reconstructions?
2) How can we exploit interactive three-dimensional representations to elucidate planetary evolution (e.g., *72*)?
3) How can we enhance use and discovery employing the GPlates platform (GPlates.org)?
4) What are optimal methods to visualize time series through animations?

Deep Learning: Hand-in-hand with data-driven discovery are a growing arsenal of machine-learning tools. A challenge is developing collaborations between discipline Earth-space-life scientists and experts in deep learning. On both sides there exist gaps in language and concepts that can hinder rapid progress. (And, as more than one attendee pointed out, "We can't compete with their salaries.") In any event, it is essential to engage fully data scientists, who must see themselves as collaborators, not technicians. Examples of needs and opportunities include:

1) Identify optimal problems in Earth, space, and life science for deep learning applications?
2) Explore optimal applications for deep neural networks (DNN), deep belief networks (DBN), deep convoluted neural networks (DCNN), and other forms of machine learning.
3) Provide examples of data-driven discoveries by machine learning.
4) Seek subtle patterns, and identify and understand "black swan" rare events.

"Small" and "sparse" data science: Many highly publicized advances in data science, including medical diagnoses, marketing strategies, and traffic avoidance apps, involve the analysis of "big data" with millions of integrated data. Many applications in Earth, space, and life sciences, by contrast, rely on relatively small data resources. "Small" can be defined in different ways across different research domains. In geochemical data, small may refer to a modest number of samples and/or a minimal number of attributes. A diamond dataset, for example, may consist of a few dozen samples, yet may lead to far reaching conclusions regarding Earth's deep interior (*69,70*). In biology, small may refer to physical sample size, low bio-mass, or low complexity metagenomes. Each example from different research fields suggests challenges in applying

traditional big data science techniques. Small data may be pivotal as a road map to larger scale data integration, but valid data use requires particular techniques, such as weighting, cross-sample comparisons, meta-analysis, creation of small subsets from massive data, and networks of data. Accordingly, we identified a number of needs and opportunities in small data science:

1) Grow and scale small data into big data by developing methods and standards for building aggregations of small data. Repository advances are vital for improved data management and storage, explicit representation of relationships, systems for cleaning data, functionality comparison across datasets, intra-disciplinary validation, and cross-disciplinary validation.
2) Develop tools for specific application to different sizes of data resources.
3) Implement methods for data scaling and aggregation (data filtering, bootstrapping, and wrangling) through extensive data sharing within and across research domains.
4) Recognize the importance of metadata in developing small data resources.

3. Build and Sustain Earth-Space-Life Data Infrastructure

Underlying all of our discussions of Earth, space, life, and data science opportunities was the recurrent theme that we must individually and collectively foster a cultural shift to support the evolution of data sharing and data science education. To this end, perhaps the largest single challenge in establishing the goals of the 4D Workshop remains building and sustaining data infrastructure. Data-driven discovery relies on the availability of comprehensive, reliable, accessible, and interoperable open-access data resources. 4D participants cited numerous examples of promising database developments in Earth, space, and life sciences, but concerns regarding the diligent maintenance, professional vetting, potentially biased content, lack of interoperability, and especially sustainable institutional support (both in terms of infrastructure and finances) reflect potential roadblocks to progress.

Build a culture of data sharing: A number of critical needs remain unmet, largely because of the necessity for a culture shift in the way data science is approached and integrated into the scientific discourse. The Earth, space, and life science communities lack clear policies and standardized practices regarding the production, utilization, re-use, and sharing of data. Achieving the necessary culture shift will require the adoption of rules for reusability and reproducibility of data products and processes, the creation of incentives to data sharing, and the fostering of data literacy across all scientific disciplines. Reducing the entry barriers to collaboration between domain experts and data scientists is a top priority—a challenge that becomes more difficult as collaborative research groups become larger and more diverse. In these cases, the ability of scaling expertise is fundamental. A major focus is also needed in order to help bridge different communities and support data integration that will facilitate data-driven discoveries. Key opportunities in data science include:

1) Create documents describing examples of success stories in data-driven discovery to underscore the opportunities and encourage the cultural change needed to embrace data sciences in all disciplines.
2) Promote data literacy to bridge gaps between data scientists and domain experts in order to lower entry barriers to data-driven discoveries.

3) Encourage and support exploratory, high-risk/high-reward projects and multidisciplinary collaborations that leverage existing data.
4) Work toward building a large open-access data infrastructure that fosters open science principles, such as accessibility, reusability, and reproducibility.

Accelerate data infrastructure development: We agreed that these problems are exacerbated during a time of transition in both the implementation and culture of data science. Vast amounts of relevant data (and their associated metadata) are preserved only in the hardcopy pages of journals and books. More data are to be found in the disk drives and file drawers of scientists around the world. The culture of data sharing is changing rapidly—one can anticipate a time when it is routine to deposit all of one's data into official repositories, with citable digital object identifiers. In a sense, therefore, the transient 4D challenge is to retrieve as much of the less accessible "dark data" as efficiently as possible during this time of shifting cultures.

In this context, it must be acknowledged that some subcultures will be more resistant to change than others. For example, the vast stores of information retained by natural resource companies are unlikely to become open access anytime soon. Only when the power of "all the data" to inform and advance individual ambitions and priorities is recognized will such data sharing likely to become commonplace. Nevertheless, the opportunity to accelerate data-driven discovery by the focused and efficient entry of data from more public sources represents a priority in data-driven discovery. Accordingly, we propose the following five steps:

1) Place a priority on funding efficient approaches to data resource development that exploit machine-learning capabilities (i.e., DeepDive, Snorkle). These efforts should support the interaction of data scientists and discipline scientists.
2) Strive to create and access large digital repositories of publications, even if those publications are under copyright. Emphasize that copyright holders stand to benefit from having their published data tabulated and openly available, as users of data want to examine original publications for context.
3) Emphasize the importance of recording relevant metadata, including photographs, methods, contextual information for samples, and original references.
4) Create and sustain data journals with citable digital object identifiers (DOIs) for previously unpublished data resources. In the process, establish discipline-based standards for deposited data, including requirements for sample identification numbers (e.g., www.geosamples.org), definitions of units, and contextual metadata.

Establish support for existing and new data infrastructure: We fully understand that significant progress and transformation of the community will take time and resources. However, the small, often handcrafted artisanal data produced in many sciences, and development of data science techniques for mobilizing the value of small data, are integral to future breakthroughs in research on the complex and co-evolving geosphere and biosphere.

The most basic need and responsibility of any natural history discipline is the accurate, timely, comprehensive, and accessible archiving of data on species. No task is more fundamental to the long-term stability and integrity of a field, nor should such data management be left to the unfunded good will of individuals, no matter how skilled and well intentioned they may be. Some

data resources such as EarthChem and Volcanoes enjoy significant, if not guaranteed long-term, Federal support. However, it is astonishing that the official web-based list of approved mineral species (rruff.info/ima), which is freely available and widely used by the international community, has no long-term institutional home or financial support. And, until recently, the server for LEPR and MELTS—widely used open-access resources for thermochemical data and analysis of Earth materials—resided in the bedroom of its founder, Mark Ghiorso. The international community, therefore, needs to initiate action on several fronts:

1) Identify and encourage recovery of "dark data" resources, including unpublished hard copy and electronic format data accumulated by individuals. Encourage publication of these data resources through electronic data journals with digital object identifier (doi) information.
2) Create active and engaged user communities to ensure quality control of data resources, which must be properly vetted prior to incorporation into open-access sources.
3) Establish data publication procedures and data citation policies that ensure proper credit and motivation for data producers.
4) Identify and exploit sources of funding for work being done now and for long-term institutional support of critical data resources.

Note that several of these steps will require both institutional and cultural changes within the Earth, space, and life science communities. Effective change will take time and effort, but we can anticipate a future environment when the larger scientific community recognizes the critical importance of shared, high-quality data resources and underscores the responsibility of all researchers to contribute to this infrastructure.

A second facet of fostering data-driven discovery in Earth, space, and life science research involves integrating existing databases into a larger deep-time data infrastructure (http://dtdi.carnegiescience.edu). Ultimately, we can envision linking separately maintained data resources into a federated data framework in which diverse data resources are semantically compatible and linked (see, for example, *12*; earthcube.org). Integration of data from many complementary disciplines into a single interoperable data platform represents the greatest unfulfilled opportunity in our quest for an abductive strategy for scientific discovery.

To accomplish this vision, we need to identify and integrate key data resources, while providing computational tools that can be used to select, analyze, and visualize data. Ultimately, a comprehensive Earth-space-life data infrastructure could be linked to artificial intelligence and machine-learning capabilities to accelerate data-driven discovery. Cyber-infrastructure programs such as EarthCube, which enjoy significant community support as well as Federal funding, are moving scientific research in these bold new directions. However, the Earth-space-life research community needs to increase its commitment to these efforts if we are to take maximum advantage of emerging opportunities.

THE 4-D INITIATIVE: A NEW BEGINNING FOR THE OLDEST QUESTIONS

One of the many successes of the 4D Workshop was the development and synthesis of new research projects among small and large groups of workshop participants. Throughout the workshop, the vast scope of research ideas became apparent. The excitement at advancing a

broadly interdisciplinary scientific agenda was palpable. When participants asked each other "where do we see this going?" no single research objective emerged. Rather, we realized that the essence of deep-time data-driven discovery transcends space and time by combining data layers from many disciplines. The potential to link interdisciplinary data in new, multi-dimensional ways represents the clearest pathway to discovering what we don't know we don't know.

To achieve this vision, a new cohesive community organization is needed to allow continuous free flowing synthesis and integration of Earth science, space science, life science, and data science. Such a collaborative network will allow a wide range of individual, entrepreneurial researchers in the coalescing 4D community to collaborate and move freely across disciplines wherever the data leads them, rather than following the disciplinary constraints of most previous efforts.

To harness the growing momentum of collaborative Earth-space-life data-driven research as spearheaded by the 4D workshop, we propose the 4D Initiative, which will serve to coordinate a global effort to explore the rich tapestry of planetary evolution. The format of the 4D Initiative will be modeled after the Deep Carbon Observatory, a 10-year international effort with more than 1200 collaborators in 55 countries (see: deepcarbon.net). Thanks to core funding by the Alfred P. Sloan Foundation, the DCO is advancing our understanding of the physical, chemical, and biological roles of carbon in Earth. Perhaps the most lasting legacy of the DCO is the nurturing of more than 600 early-career scientists at more than 100 institutions worldwide. Many of the 4D Workshop attendees were from that exceptional group of young scholars.

The 4D Initiative, with its support of Earth, space, life, and data scientists at dozens of institutions, will continue the momentum of the 4D workshop and build on the growing movement of the scientific community for collaborative interdisciplinary research. New ideas and data syntheses will emerge spontaneously as 4D colleagues interact on many levels and at varied disciplinary boundaries; indeed, it is crucial that these opportunities are nurtured before they are lost as individuals return to their home institutions and established research priorities. The 4D Initiative will provide a framework for fostering and facilitating new pursuits and will encourage the addition of other domain experts as ideas develop and data syntheses grow.

The model for the 4D Initiative is streamlined, involving three primary budget categories:

1) Provide seed funding for early-career scientists to pursue training and research in data-driven discovery related to their chosen disciplinary specialties.
2) Distribute funds for travel to meetings of varied sizes, from small targeted workshops of a few participants to annual conferences similar to the inaugural 4D Workshop.
3) Establish a small coordinating office at the Carnegie Institution for Science that will maintain membership lists, coordinate funding opportunities, publish a website and newsletter, and organize 4D workshops and conferences.

As with the Deep Carbon Observatory, the 4D Initiative will not fully fund all related initiatives. Rather it will serve to leverage initial 4D funds by at least a factor of ten by supporting "start-up" efforts through grant-writing workshops and matching funds. Employing this model, the DCO has raised approximately $500 million over the past 10 years. With a modest initial investment, we anticipate a similar leveraged return in the pursuit of 4D science.

Metrics for Success: What do we hope to achieve with the 4D Initiative? The most profound measure of success will lie in the unanticipated discoveries that must follow from the thoughtful application of abductive strategies to large and growing data resources. However, we can anticipate achieving several quantitative goals:

1. Build a planetary materials data resource with more than 1 billion data fields (a "Census of planetary materials").
2. Coordinate and engage an international community of more than 2,000 scholars in varied disciplines from more than 50 countries.
3. Publish 5,000 peer-reviewed publications in Earth, space, life, and data sciences.
4. Raise more than $600 million in funds from private, corporate, and government sources.
5. Enlist and train more than 1000 early-career scientists in methods and application of data-driven discovery to their disciplines.
6. Transform the culture of science to make open access sharing of data and methods of data analysis and visualization the norm.

Conclusion

Thanks to a host of recent advances and insights in Earth, space, life, and data sciences, the promise of harnessing data simultaneously from many domains beckons. Thanks to the power of abductive discovery, we have the opportunity to embrace the multi-dimensional complexity of the universe as never before. In so doing, we are poised to enjoy a burst of discovery that will change our perspectives on our evolving world and our place in the cosmos.

REFERENCES

1. Darwin, C. (1859) *On the Origin of Species*. London: John Murray.
2. Beddall, B. G. (1968). Wallace, Darwin, and the theory of natural selection. *Journal of the History of Biology*, **1,** 261–323.
3. Wood, R.M. (1985) *The Dark Side of the Earth*. London: George Allen and Unwin.
4. Hazen, R.M. (2012) *The Story of Earth*. New York: Viking-Penguin.
5. Holland, H.D. (1984) *The Chemical Evolution of the Atmosphere and Oceans*. Princeton, NJ: Princeton University Press.
6. Canfield, D.E. (2014) *Oxygen: A Four Billion Year History*. Princeton, NJ: Princeton University Press.
7. World Meteorological Organization (1989) The Changing Atmosphere: Implications for Global Security, Toronto, Canada, 27-30 June 1988: Conference Proceedings. Geneva: Secretariat of the World Meteorological Organization.
8. Weart, S.R. (2008) *The Discovery of Global Warming: Revised and Expanded Edition*. Cambridge, Massachusetts: Harvard University Press.
9. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996) From data mining to knowledge discovery in databases. *AI Magazine*, Fall 1996, 37-54.
10. Hazen R.M., Bekker A., Bish D.L., Bleeker W., Downs R.T., Farquhar J., Ferry J.M., Grew E.S., Knoll A.H., Papineau D., Ralph J.P., Sverjensky D.A., and Valley J.W. (2011) Needs and opportunities in mineral evolution research. *American Mineralogist*, 96, 953-963.
11. Keller, B., and Schoene (2012) Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gya ago. Nature, 485, 490-493.
12. National Science Foundation (2012) *A Community Roadmap for Earthcube Data: Discovery, Access, and Mining.* Arlington, Virginia: National Science Foundation, 38 p.
13. Bolukbasi, B., Berente, N., Cutcher-Gershenfeld, J., Dechurch, L., Flint, C., Haberman, M., King, J.L., Knight, E., Lawrence, B., Masella, E., McElroy, C., Mittleman, B., Nolan, M., Radik, M., Shin, N., Thompson, C.A., Winter, S., Zaslavsky, Allison, M.L., Arctur, D., Arrigo, J., Aufdenkampe, A.K., Bass, J., Crowell, J., Daniels, M., Diggs, S., Duffy, C., Gil, Y., Gomez, B., Graves, S., Hazen, R., Hsu, L., Kinkade, D., Lehnert, K., Marone, C., Middleton, D., Noren, A., Paerthree, G., Ramamurthy, M., Robinson, E., Percivall, G., Richard, S., Suarez, C., and Walker, D. (2013) Open data: Crediting a culture of cooperation. *Science*, 342, 1041-4042. DOI:10.1126/science.342.6162.1041-b
14. www.earthcube.org
15. Hazen, R.M., Golden, J., Downs, R.T., Hystad, G., Grew, E.S., Azzolini, D, and Sverjensky, D.A. (2012) Mercury (Hg) mineral evolution: A mineralogical record of supercontinent assembly, changing ocean geochemistry, and the emerging terrestrial biosphere. *American Mineralogist*, 97, 1013-1042.
16. Huston, D.L., Pehrsson, S., Eglington, B.M., and Zaw, K. (2010) The geology and metallogeny of volcanic-hosted massive sulfide deposits: Variations through geologic time and with tectonic setting. *Economic Geology*, 106, 571-591.
17. Liu, C., Knoll, A.H., and Hazen, R.M. (2017) Geochemical and mineralogical evidence that Rodinian assembly was unique. *Nature Communications.* DOI: 10.1038/s41467-017-02095-x

18. Chamberlin, T.C. (1890) The method of multiple working hypotheses. *Science*, 15, 92-96.
19. Hutton, J. (1795) *Theory of the Earth; with Proofs and Illustrations*. Edinburgh: Creech. 2 volumes.
20. Berner-Lee, T., Hendler, J., and Lassila, O. (2001) The semantic web. *Scientific American*, 284(5), 34-43.
21. Hey, T., and Trefethen, A.E. (2005) Cyberinfrastructure for e-Science. *Science*, 308, 817-821.
22. Fox, P., and Hendler, J. (2009) Semantic eScience: Encoding meaning in next-generation digitally enhanced science. In Hey, T., Tansley, S., Tolle, K. [Editors] The Fourth Paradigm: Data-Intensive Scientific Discovery. Redland, WA: Microsoft External Research, pp. 145-150.
23. Hey, T., Tansley, S., and Tolle, K. [Editors] (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redland, WA: Microsoft External Research.
24. McGuinness, D.L., Fox, P.A., Brodaric, B., and Kendall, E. (2009) The emerging field of semantic scientific knowledge integration. *IEEE Intelligent Systems*, 24, 25-26.
25. Narock, T., and Fox, P.A. (2011) From science to e-science to semantic e-science: A heliophysics case study. *Computers & Geosciences*. doi: 10.1016/j.cageo.2011.11.018
26. Card, S.K., Mackinlay, J.D., and Shneiderman, B. (1999) *Reading in Information Visualization: Using Vision to Think*. San Francisco, California: Morgan Kaufmann.
27. Hammer, Ø., Harper, D.A.T., and Ryan, P.D. (2001) PAST: Paleontological statistical software package for education and data analysis. *Paleo-Electronica.org*, issue 1.
28. Perer, A., and Shneiderman, B. (2008) Integrating statistics and visualization: Case studies of gaining clarity during exploratory data analysis. ACM Conference on Human Factors in Computing Systems, Florence, Italy.
29. Fox, P., and Hendler, J. (2011) Changing the equation on scientific data visualization. *Science*, 331, 705-708.
30. Kim, J.D., Senn, S., Harel, A., Jelen, B.I., and Falkowski, P.G. (2013) Discovering the electronic circuit diagram of life: structural relationships among transition metal binding sites in oxidoreductases. *Philosophical Transactions of the Royal Society*, B 368, 20120257.
31. Kovalerchuk, B. and Vityaev, E. (2000) *Data Mining in Finance: Advances in Relational and Hybrid Methods*. New York: Kluwer Academic Publishers.
32. Gellman, B. and Poitras, L. (2013). U. S. intelligence mining data from nine U.S. Internet companies in broad secret program. The Washington Post, June 7, 2013, A1.
33. Holmes, E.C. (2007) Viral evolution in the genomic age. *PLOS Biology*, DOI: 10.1371/journal.pbio.0050278
34. Lam, T.T., Hon, C.C., and Tang, J.W. (2010) Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical Reviews in Clinical Laboratory Sciences*, 47, 5–49.
35. Clos, K.J. [Editor] (2001) *Medical Data Mining and Knowledge Discovery*. Dortrecht, Netherlands: Springer.
36. Berka, P., Rauch, J., and Djamel, A.Z. [Editors] (2009) *Data Mining and Medical Knowledge Management: Cases and Applications*. Hershey, Pennsylvania: IGI Global.
37. Pyle, J.M., Spear, F.S., and Wark, D.A. (2002) Electron microprobe analysis of REE in apatite, monazite and xenotime: Protocols and pitfalls. *Reviews in Mineralogy and Geochemistry*, 48, 337-362.

38. Donovan, J.J., Hanchar, J.M., Picollo, P.M., Schrier, M.D., Boatner, L.A., and Jarosewich, E. (2003) A Re-examination of the rare-earth-element orthophosphate standards in use for electron-microprobe analysis. *Canadian Mineralogist*. 41, 221-232.
39. Grew, E.S., and Hazen, R.M. (2014) Beryllium mineral evolution. *American Mineralogist*, **99**, 999-1021.
40. Tkachev, A.V. (2011) Evolution of metallogeny of granitic pegmatites associated with orogens throughout geological time. *Geological Society of London, Special Publications* 350, 7-23.
41. Valley, J.W., Lackey, J.S., Cavosie, A.J., Clechenko, C.C., Spicuzza, M.J., Basei, M.A.S., Bindeman, I.N., Ferreira, V.P., Sial, A.N., King, E.M., Peck, W.H., Sinha, A.K., and Wei, C.S. (2005) 4.4 billion years of crustal maturation: oxygen isotope ratios of magmatic zircon. *Contributions to Mineralogy and Petrology*, 150, 561-580.
42. Campbell, I.H., and Allen, C.M. (2008) Formation of supercontinents linked to increases in atmospheric oxygen. *Nature Geoscience*, 1, 554-558.
43. Rino, S., Kon, Y., Sato, W., Maruyama, S., Santosh, M., and Zhao, D. (2008) The Grenvillian and Pan-African orogens: world's largest orogenies through geologic time, and their implications on the origin of superplumes. *Gondwana Research*, 14, 51-72.
44. Hawkesworth, C.J., Dhuime, B., Pietranik, A.B., Kemp, A.I.S., and Storey, C.D. (2010) The generation and evolution of continental crust. *Journal of the Geological Society*, 167, 229-248.
45. Condie, K.C., and Aster, R.C. (2010) Episodic zircon age spectra of orogenic granitoids: The supercontinent connection and continental growth. *Precambrian Research*, 180, 227-236.
46. Condie, K.C., Bickford, M.E., Aster, R.C., Belousova, E., and Scholl, D.W. (2011) Episodic zircon ages, Hf isotopic composition, and the preservation rate of continental crust. *Geological Society of America Bulletin*, 123, 951-957.
47. Voice, P.J., Kowalewski, M., and Eriksson, K.A. (2011) Quantifying the timing and rate of crustal evolution: Global compilation of radiometrically dated detrital zircon grains. *Journal of Geology*, 119, 109-126.
48. Canfield, D.E. (1998) A new model for Proterozoic ocean chemistry. *Nature*, 396, 450-453.
49. Cawood, P.A., Kroner, A., Collins, W.J., Kusky, T.W., Mooney, W.D., and Windley, B.F. (2009) Accretionary orogens through Earth history. *Geological Society [London] Special Publication*, 318, 1–36.
50. Huston, D.L., Pehrsson, S., Eglington, B.M., and Zaw, K. (2010) The geology and metallogeny of volcanic-hosted massive sulfide deposits: Variations through geologic time and with tectonic setting. *Economic Geology*, 106, 571-591.
51. Downs, R.T. (2006) The RRUFF Project: an integrated study of the chemistry, crystallography, Raman and infrared spectroscopy of minerals. *Program and Abstracts of the 19th General Meeting of the International Mineralogical Association in Kobe, Japan*. O03-13.
52. Farquhar, J., Bao, H., and Thiemens, M.H. (2000) Atmospheric Influence of Earth's Earliest Sulfur Cycle. *Science*, 289, 756-758.
53. Farquhar, J., Savarino, I., Airieau, S., and Thiemens, M.H. (2001) Observations of wavelength-sensitive, mass-independent sulfur isotope effects during SO_2 photolysis: Implications for the early atmosphere. *Journal of Geophysical Research*, 106, 1-11.

54. Farquhar, J., Peters, M., Johnston, D.T., Strauss, H., Masterson, A., Wiechert, U., and Kaufman, A.J. (2007) Isotopic evidence for mesoarchean anoxia and changing atmospheric sulphur chemistry. *Nature*, 449, 706-709.
55. Lehnert, K.A., Su, Y., Langmuir, C.H., Sarbas, B., and Nohl, U. (2000) A global geochemical database structure for rocks. *Geochemistry Geophysics Geosystems* 1.
56. Lehnert, K.A., Walker, D., and Sarbas, B. (2007) EarthChem: A geochemistry data network. *Geochimica et Cosmochimica Acta*, 71, A559.
57. Golden, J., McMillan, M., Downs, R.T., Hystad, G., Stein, H.J., Zimmerman, A., Sverjensky, D.A. Armstrong, J., and Hazen, R.M. (2013) Rhenium variations in molybdenite (MoS_2): Evidence for progressive subsurface oxidation. *Earth and Planetary Science Letters*, **366**, 1-5.
58. Hazen, R.M., Cotrell, E., Downs, R.T., Fox, P., Ghiorso, M., Lehnert, K., Saxena, S., and Spear, F. (2013) Report of the Chair of the Ad Hoc Committee on Earth Materials Data, To the First 2013 Mineralogical Society of America Council Meeting, 23 April 2013, 6 pp.
59. Hazen, R.M. (2014) Data-driven abductive discovery in mineralogy. *American Mineralogist*, **99**, 2165-2170.
60. Lehnert, K.A., and Klump, J. (2008) Facilitating research in mantle petrology with geoinformatics. *9th International Kimberlite Conference Extended Abstracts*, 91KC-A-00250.
61. Downs, R.T., and Hall-Wallace, M. (2003) The American Mineralogist Crystal Structure Database. *American Mineralogist*, 88, 247-250.
62. Grazulis, S., Daskevic, A., Merkys, A., Chateigner, D., Lutterotti, L., Quiros, M., Serebryanaya, N.R., Moeck, P., Downs, R.T., and Le Bail, A. (2012) Crystallography Open Database (COD): an open-access collection of crystal structures and platforms for world-wide collaboration. *Nucleic Acids Research*, 40, D420-D427. doi:10.1093/nar/gkr900
63. Mutschler, F.E., Rougon, D.J., Lavin, O.P., and R.D. Hughes (1981) *PETROS version 6.1 Worldwide Databank of Major Element Chemical Analyses of Igneous Rocks*. National Geophysical Data Center, NOAA. doi:10.7289/V5QN64NM.
64. Ghiorso, M.S., and Sack, R.O. (1995) Chemical Mass Transfer in Magmatic Processes IV. A revised and internally consistent thermodynamic model for the interpolation and extrapolation of liquid-solid equilibria in magmatic systems at elevated temperatures and pressures. *Contributions to Mineralogy and Petrology*, 119, 197-212.
65. Ghiorso, M.S., Hirschmann, M.M., Reiners, P.W., and Kress, V.C. III (2002) The pMELTS: A revision of MELTS for improved calculation of phase relations and major element partitioning related to partial melting of the mantle to 3 GPa. *Geochemistry Geophysics Geosystems*, 3, 10.1029/2001GC000217.
66. Holland, T.J.B., and Powell, R. (1998) An internally consistent thermodynamic data set for phases of petrological interest. *Journal of Metamorphic Petrology*, 16, 309–343.
67. Stixrude, L., and Lithgow-Bertelloni, C. (2005) Thermodynamics of mantle minerals – I. Physical properties. *Geophysical Journal International*, 162, 610–632.
68. Stixrude, L., and Lithgow-Bertelloni, C. (2011) Thermodynamics of mantle minerals – II. Phase equilibria, *Geophysical Journal International*, 184, 1180–1213.
69. Smith, E.M., Shirey, S.B., Nestola, F., Bullock, E.S., Wang, J., Richardson, S.H., and Wang, W. (2016) Large gem diamonds from metallic liquid in Earth's deep mantle. *Science*, 354, 1403-1405.

70. Smith, E.M., Shirey, S.B., Richardson, S.H., Nestola, F., Bullock, E.S., Wang, J., and Wang, W. (2018) Blue boron-bearing diamonds from Earth's lower mantle. *Nature*, 560, 84-87.
71. Tufte, E. (2001) *The Visual Display of Quantitative Information, 2nd Edition*. Cheshire, CT: Graphics Press.
72. Ma, X., Hummer, D., Golden, J.J., Fox, P.A., Hazen, R.M., Morrison, S.M., Downs, R.T., Madhikarmi, B., Wang, C., and Meyer, M.B. (2017) Using visualized exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research. *International Journal of Data Science*, 6, 368, 11 p.