NSF Directorate for Computer and Information Science and Engineering
and
NSF Directorate for Geosciences

2015 Workshop on
Intelligent and Information Systems for Geosciences

Final Workshop Report

November 30, 2015

March 26-27, 2015

Arlington, VA

http://www.is-geo.org

This workshop was sponsored by the Directorate for Computer and Information Science and Engineering and the Directorate for Geosciences of the National Science Foundation under grant number IIS-1533930.

This report can be cited as:

"Final Report on the 2015 NSF Workshop on Information and Intelligent Systems for Geosciences." Yolanda Gil and Suzanne A. Pierce (Eds). National Science Foundation Workshop Report, October 2015. This report is published as part of the National Science Foundation's (NSF) Directorate for Computer and Information Science and Engineering (CISE) report series and is available at http://dl.acm.org/collection.cfm?id=C13.

Table of Contents

Workshop Participants

Co-Chairs

Yolanda Gil, University of Southern California
Suzanne A Pierce, The University of Texas at Austin

Participants

Hassan Babaie, Georgia State University
Arindam Banerjee, University of Minnesota
Kirk Borne, George Mason University
Gary Bust, Johns Hopkins
Michelle Cheatham, Wright State University
Imme Ebert-Uphoff, Colorado State University
Carla Gomes, Cornell University
Mary Hill, The University of Kansas
John Horel, University of Utah
Leslie Hsu, Columbia University
Jim Kinter, George Mason University
Craig Knoblock, University of Southern California
David Krum, University of Southern California
Vipin Kumar, University of Minnesota
Pierre Lermusiaux, Massachusetts Institute of Technology
Yan Liu, University of Southern California
Deborah McGuinness, Rensselaer Polytechnic Institute
Chris North, Virginia Tech
Victor Pankratius, Massachusetts Institute of Technology
Shanan Peters, University of Wisconsin-Madison
Beth Plale, Indiana University Bloomington
Allen Pope, University of Colorado Boulder
Sai Ravela, Massachusetts Institute of Technology
Juan Restrepo, Oregon State University
Aaron Ridley, University of Michigan
Hanan Samet, University of Maryland
Shashi Shekhar, University of Minnesota
Katie Skinner, University of Michigan
Padhraic Smyth, University of California, Irvine
Basil Tikoff, University of Wisconsin-Madison
Lynn Yarmey, National Snow and Ice Data Center
Jia Zhang, Carnegie Mellon University – Silicon Valley

Cognizant NSF Program Officers

Dr. Héctor Muñoz-Avila, NSF CISE
Dr. Eva Zanzerkia, NSF GEO

Government Observers

Luciana Astiz, NSF GEO/EAR
Leonard Johnson, NSF GEO/EAR
Todd Leen, NSF CISE /IIS
Audrey Levine, NSF IIA/EPSCOR
Deborah F. Lockhart, NSF CISE/IIS
Marcia E. McNiff, U.S. Geological Survey
Stephen Meacham, NSF IIA
Frank Olken, NSF CISE/IIS
Aaron Rosenberg, NSF GEO/OCE
Jack M. Sharp, Jr., NSF GEO/EAR
Michael Sieracki, NSF GEO/OCE
Sylvia Spengler, NSF CISE/IIS
Jack Snoeyink, NSF CISE/CCF
Amy Walton, NSF CISE/ACI
Maria Zemankova, NSF CISE/IIS

Executive Summary

Synopsis: *This workshop's outcome is twofold. It highlighted potential breakthrough advances in geosciences resulting from computing research. It also revealed groundbreaking computing research challenges motivated by problems in geosciences. These outcomes call for synergistic research in computing and geosciences.*

The goal of the Intelligent Systems for Geosciences workshop was to identify avenues for future research on intelligent systems that will result in fundamental new insights in geosciences. Geosciences representatives brought requirements from Earth, ocean, polar, and geospace sciences. Participants from intelligent systems represented fields such as information integration, machine learning, knowledge representation, semantics and metadata, geospatial computing, robotics, visualization, and augmented reality. The workshop built on the momentum of the NSF EarthCube initiative for geosciences and was informed by ongoing cyberinfrastructure efforts.

Many aspects of geosciences research pose novel problems for intelligent systems research. Geoscience data is interesting to computer scientists because it tends to be uncertain, intermittent, sparse, multi-resolution, and multi-scale. Geosciences processes and objects often have amorphous spatio-temporal boundaries. The lack of ground truth makes model evaluation, testing, and comparison difficult. Overcoming these challenges would greatly benefit the geosciences and would require breakthroughs in intelligent systems.

Workshop participants agreed that in order to address these challenges new research is required in intelligent and information systems, including:

- **Knowledge representation**: Capturing scientific knowledge to represent our understanding of geoscience processes will push the limits of the state of the art.
- **Sensing and robotics**: New data collection capabilities that leverage scientific knowledge for optimized data collection and adaptive sampling.
- **Information integration**: Geosciences data and models need to be interconnected and easy to manipulate in an integrated system of systems.
- **Machine learning**: Algorithms need to identify and incorporate appropriate constraints as required by the governing geosciences processes.
- **Intelligent user interfaces**: Interactions must be guided by the geoscience questions that provide context for the content to be conveyed.

All these areas cannot be investigated separately as they are interdependent. For example, improvements in sensing will facilitate learning, deeper representations of data will facilitate information integration, and richer learning algorithms will lead to better interfaces.

Recommendations from the workshop include: 1) Interdisciplinary community building through sustained multi-year collaborations; 2) Educating and building awareness of computer scientists and geoscientists in each other's fields, and 3) Establishing direct partnerships between intelligent systems and geoscience researchers. Participants noted that these activities will result in innovations in both intelligent systems and geosciences.

1. Introduction

The goal of this workshop was to synthesize a vision and needs for intelligent systems research that will provide new capabilities to advance geosciences. In geosciences, the workshop identified starting requirements from Earth, ocean, polar, and atmospheric and geospatial sciences that would benefit from intelligent systems advances. In intelligent systems, the workshop included participants from fields such as information integration, machine learning, knowledge representation, social computing, visualization, and intelligent user interfaces. The workshop was informed by existing cyberinfrastructure efforts that support the geoscience community. The workshop served as a bridge to find areas of mutual benefit and to begin connecting these communities to explore collaborative research.

Participants discussed how to tackle problems in heterogeneous data integration and visualization (e.g., hand-made sketches, aerial imagery, field-data repositories, stakeholder interviews), ontological reasoning with scientific metadata and mathematical models (e.g., representing uncertainty, simulation predictions, evolving theories). Additionally, participants were asked to consider potential uses for intelligent assistants that make scientists more efficient by automating routine tasks that require some level of knowledge about the science context and that facilitate information sharing, collaborative workflow design and management to support data analytics, and sophisticated machine learning techniques to analyze geosciences data. Participants identified complexities and challenges in the application of intelligent systems to geoscience domains; addressing these challenges will require intensive collaborations including researchers from geosciences and information systems. The workshop catalyzed a community and research agenda in the emerging area of intelligent systems grounded on geoscience requirements.

The NSF EarthCube Initiative[1] presents an opportunity for collaborative research on novel information systems enhancing and supporting geoscience research efforts. EarthCube enables geoscientists to address the challenges of understanding and predicting a complex and evolving Earth system by fostering a community-governed effort to develop a common cyberinfrastructure to collect, access, analyze, share and visualize all forms of data and resources, using advanced technological and computational capabilities [Gil et al 2014].

Workshop participants were asked to contribute position papers prior to the workshop. The papers, which are included in the Appendix of this report, outlined a range of community activities, recurring events, and ongoing projects where geoscientists and intelligent systems researchers are already interacting. These activities, summarized in Section 2, illustrate how research advances in intelligent systems are already impacting the geosciences. Full position papers from each participant can be accessed online at the Web site for the workshop.[2]

[1] http://www.earthcube.org
[2] http://www.is-geo.org

To open the workshop, the workshop chairs summarized the position papers in the first session of the meeting. The emerging themes were organized around the phases of data lifecycle management: data collection, data integration, data analysis, data processing, and data visualization. This opening session began to draw connections between geosciences and intelligent systems research. To elaborate on these themes, the next two workshop sessions were organized as a "World Café".[3] The "World Café" approach is a method for promoting open dialogue among groups. Before the event a set of substantive questions was developed around themes from the participant position papers and the questions were distributed around tables in the meeting space. Each table was assigned a theme with the questions to seed entry points into conversations; a set of participants stayed at each table while others rotated to balance continuity and diversity in each discussion and to build relationships among participants. The first "World Café" session had tables with geosciences themes with a geoscientist facilitator while computer scientists rotated along with the non-facilitating geoscientists. In the second "World Café" session, intelligent systems themes were assigned to the tables and the geoscientists rotated among tables with non-facilitating computer scientists. The "World Café" design created opportunities for small group interaction across the disciplines and the discussions were lively and productive.

The initial sessions were followed by plenary discussions on the themes that arose from each of the "World Café" tables. The results of these discussions are summarized in Section 3 of this report. Participants then reflected on grand challenges in geosciences that cannot be addressed without significant innovations over existing capabilities. These are presented in Section 4 of this report.

As participants gained familiarity and shared understanding of the topics, themes, and opportunities, discussions and focus turned to formulating a research agenda for intelligent systems that would advance these issues. The participant consensus was that intelligent systems that incorporate existing scientific knowledge and the geoscientist's context enable new research challenges and opportunities for both arenas. This alone would enable novel forms of reasoning and learning about geosciences data. This research agenda is described in Section 5. Finally, participants discussed recommendations and next steps, which are included in the closing section of this report.

2. Existing Interactions Between Intelligent Systems and Geosciences

There are several existing threads of interactions between intelligent systems researchers and geoscientists that we summarize here. These activities were reported by workshop participants in their pre-workshop statements, and helped ground the discussions throughout the workshop.

[3] http://en.wikipedia.org/wiki/World_Caf%C3%A9_%28conversational_process%29

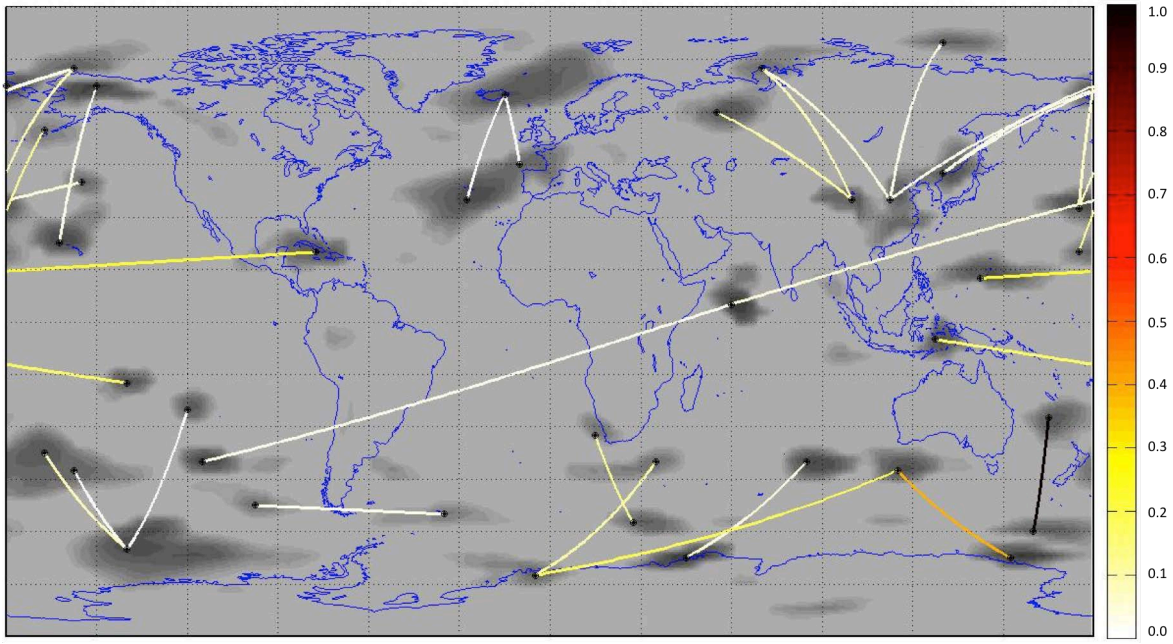

Figure 1. The novel Shared Reciprocal Nearest Neighbor (SRNN) method [Kawale et al. 2013] identifies locations with negatively correlated temporal behavior, shown using black dots, shading, and edges (lines) that link them. Here, the quantity being evaluated is the global monthly mean Hadley Centre sea level pressure (SLP). Region-pairs represent climate dipoles, and are key to understanding climate variability. The SRNN method was able to identify most of the already-known prominent dipoles in SLP data, and was also able to suggest new dipoles such as the black edge shown on the east side of Australia. This newly discovered dipole was found to represent circulation patterns that are associated with long-term droughts over Australia [Liess et al. 2014]. The lines are colored to indicate correlation of this dipole with other dipoles around the globe. Poor correlations (as reflected in light colored edges) suggest that the newly discovered dipole is unique. This method, unlike previous approaches, is able to identify weak and strong dipoles simultaneously, and thus enables comprehensive study of behavior, interactions, and dynamics of various dipoles.

2.1 Workshops and Other Community Activities

Several prior community activities have led to fruitful interactions in areas of common interest to geoscientists and intelligent systems researchers. NSF funded a Discovery Informatics workshop in 2012, which was broader in scope than geosciences, and was followed by a series of workshops on intelligent systems for scientific discovery[4]. An ongoing Climate Informatics workshop series[5], focused on machine learning in climate research, is also partially funded by NSF. The NSF-funded Computing Research Association Computing Community

[4] http://www.discoveryinformaticsinitiative.org
[5] http://www.climateinformatics.org

Consortium ran a Visioning Workshop on Spatial Computing[6]. The Polar Cyberinfrastructure division of NSF's Division of Polar Sciences supported a workshop reviewing the state and direction of cyberinfrastructure in support of that geoscience community (Pundsack et al. 2013).

2.2 Large Cross-Disciplinary Projects

Two major NSF Expeditions awards are of note. Expeditions projects have been funded in recent years by the NSF CISE directorate as major initiatives to catalyze new research communities. One NSF Expeditions award focuses on "Machine Learning for Climate Research"[7]. This has been a very successful project in generating advances in machine learning while making contributions to climate research. Figure 1 shows a major result that is illustrative of the kind of research taking place in this project.

A second NSF Expeditions award of note is "Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society"[8]. This project has catalyzed a research community on computational sustainability, and has spawned a series of special tracks on that topic at major AI conferences. The focus is more on biodiversity and environmental biology than geosciences proper.

Other major efforts resulted from the NSF ITR program. These include "GEON: A Research Project to Create Cyberinfrastructure for the Geosciences"[9], and "The SCEC Community Modeling Environment: An Information Infrastructure for System-Level Earthquake Research"[10]. The NSF ITR initiated several programs within Geoscience that later were continued under the NSF and AFOSR DDDAS programs and other initiatives.

2.3 EarthCube

The NSF EarthCube initiative has funded several projects that use intelligent systems methods to address particular aspects of geosciences research. Relevant EarthCube-funded projects include:

- BCube uses semantic technologies to map terms and data across diverse resources for interdisciplinary data discovery and access;
- Earth System Bridge is developing ontologies and representations of essential variables and assumptions to map across model representations;
- EarthCollab is using semantic and linked data technologies to represent and connect people, projects, data, and documents;
- GeoDeepDive is using natural language processing techniques to extract structured geo-located information from published articles;

[6] http://www.cra.org/ccc/visioning/visioning-activities/spatial-computing
[7] http://climatechange.cs.umn.edu/
[8] http://computational-sustainability.cis.cornell.edu/
[9] http://www.geongrid.org/
[10] http://scec.usc.edu/research/cme/

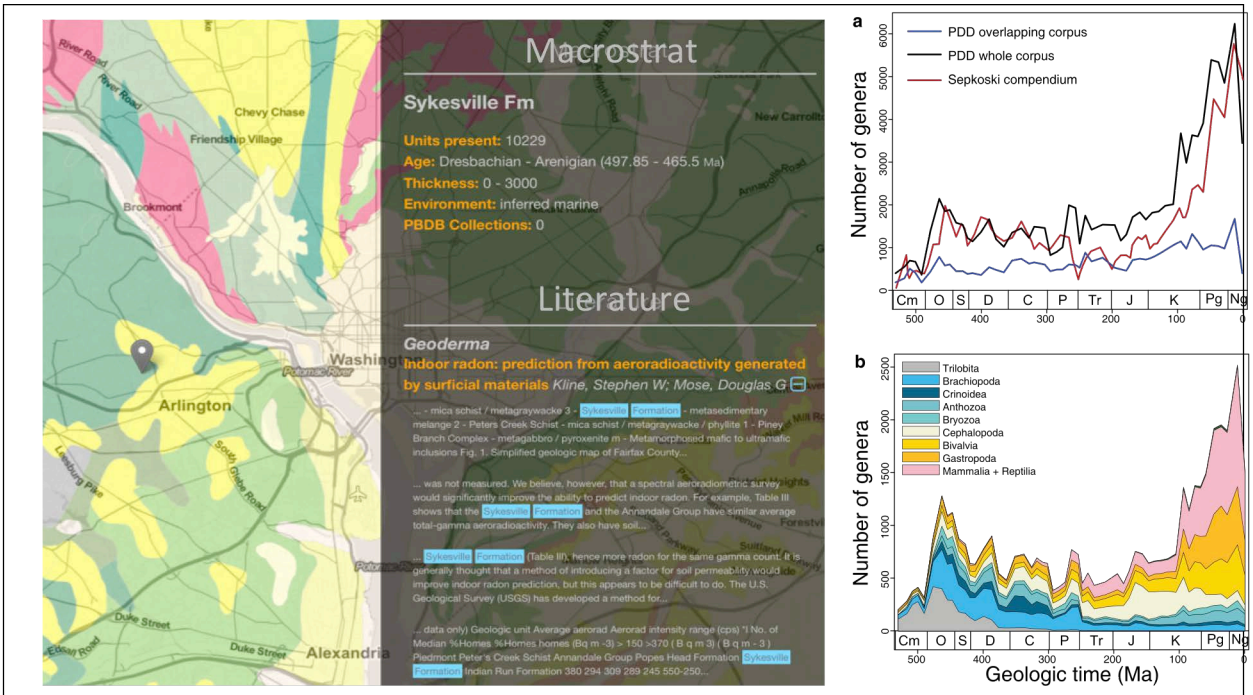

Figure 2. A machine reading system that extracts space-time indexed information from the literature and integrates it with existing paleogeosciences data focusing on the rock record. GeoDeepDive uses natural language processing techniques for parsing, document layout, and font recognition over a high-throughput computing infrastructure [Peters et al. 2014]. (a) Results from GeoDeepDive compared to results achieved by humans manually entering data from the published literature. Two different machine compilations are shown: a PDD overlapping corpus consisting of documents also read by humans and entered into the Paleobiology Database (http://paleobiodb.org), and a PDD whole corpus that comprises a larger set of documents. The latter extracted more information than what was obtained through a compendium manually assembled by Jack Sepkoski over the course of more than 10 years. (b) The machine-generated genus diversity history for select class-level taxa.

- GeoLink is creating a linked open data repository that uses semantic web representations to create open web objects with geoscience data;
- GeoSemantics is using a graph knowledge base to connect and reason about geoscience objects;
- OntoSoft is using ontologies to describe scientific software metadata, and intelligent user interfaces to assist users to publish and describe their software.

More details about these and other EarthCube activities can be found at the EarthCube Web site[1].

Perhaps the longest running of these projects is GeoDeepDive, illustrated in Figure 2. GeoDeepDive is a machine reading system that uses natural language processing techniques to extract large amounts of geoscience data that currently reside in the text, tables, and figures of scientific publications. Many important science questions require synthesizing legacy data that

is only available in the published literature, and currently involve slow and costly manual extraction. GeoDeepDive demonstrates that data can be automatically extracted from the literature with results comparable to manual extraction.

In addition to these funded projects, a series of EarthCube End User Workshops in different areas of geosciences occurred between 2012 and 2014. A majority of reports documenting these workshops consistently point to requirements in semantic metadata standards and interactive 4D visualizations. These requirements indicate that the geosciences community considers knowledge representation and visualization as important priorities in their research, and that the EarthCube geoscience community is poised for the kinds of interactions and collaborations that this workshop aimed to foster.

2.4 The Challenges in Crossing Paths

The opportunities for geoscientists to interact with intelligent systems researchers are not many. First, many geoscientists are in institutions with no significant intelligent systems researcher capabilities and are unaware of the possible benefits of such collaborations. This includes not only academia but also many government agencies. Second, there are not many community events devoted to synergistic work between geosciences and intelligent systems. The workshops and activities mentioned above are either very general and encompass many sciences beside geosciences, or they are very specific to an area in geosciences and/or intelligent systems. Third, there are no broadcast channels to disseminate successful uses of intelligent systems in geosciences. This is all in stark contrast with biomedical research, where cross-disciplinary forums abound in the form of events (e.g., the Intelligent Systems for Molecular Biology conference), journals (e.g., Bioinformatics), and community infrastructure (e.g., the National Center for Biomedical Ontologies). In geosciences, events such as the American Geophysical Union (AGU) meetings may include relevant sessions (e.g., on ontologies), but intelligent systems researchers do not typically attend these meetings. Conversely, events such as the Association for the Advancement of Artificial Intelligence include special tracks (e.g., sustainability) but they are not forums where geoscientists participate. The opportunities for exploring relevant research, formulate collaborations, and learning about success stories are very few, but offer strong evidence of successful results to provide the confidence for increasing the collaborative scope.

Figure 3 illustrates an example of an area of research with enormous potential but with little or no awareness in geosciences. The advent of low-cost virtual reality devices opens new possibilities for scientists to experience different locations and timeframes, to explore datasets and annotate findings and possible hypotheses.

In summary, although there have been significant and beneficial interactions between the intelligent systems and geosciences communities, the potential for synergistic research in intelligent systems for geosciences is largely untapped.

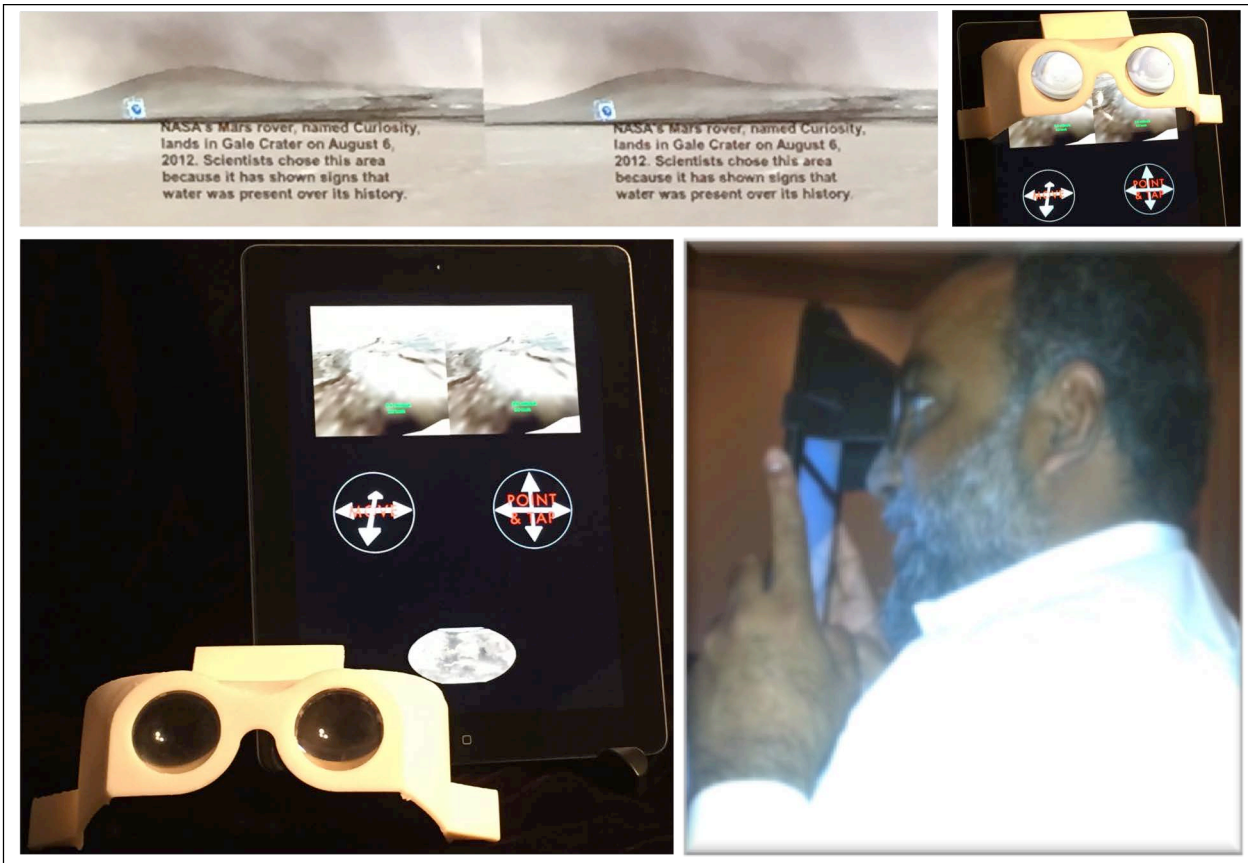

Figure 3. A low-cost virtual reality device on a tablet can be used to navigate and annotate a site. The Martian landscape shown here was created from data provided by NASA's Jet Propulsion Laboratory. Placing specially designed goggles over a tablet, scientists can explore the Mars the landing site for Curiosity Rover at Gale Crater. The site was chosen because it has shown signs that water was present over its history. They can survey the geological features in stereoscopic 3D by turning their heads and look up and down. Controls on the surface of the tablet allow users to travel across the terrain, visit different sites of interest, and annotate with text or narration. (Source: David M. Krum, University of Southern California's Institute for Creative Technologies).

3. Initial Synthesis of Emerging Themes

A major challenge in cross-disciplinary workshops is the diversity of viewpoints, vocabulary, and research interests represented by participants. To create common ground across disciplines, an initial synthesis of emerging themes was done through discussions around five aspects of the data lifecycle that were familiar to all participants: data collection, data integration, data processing, data analysis, and data visualization. These discussions were conducted in two phases. The first phase was driven by geoscientists and focused on science drivers, and the second phase driven by intelligent systems researchers who examined those drivers and derived the required capabilities for intelligent systems.

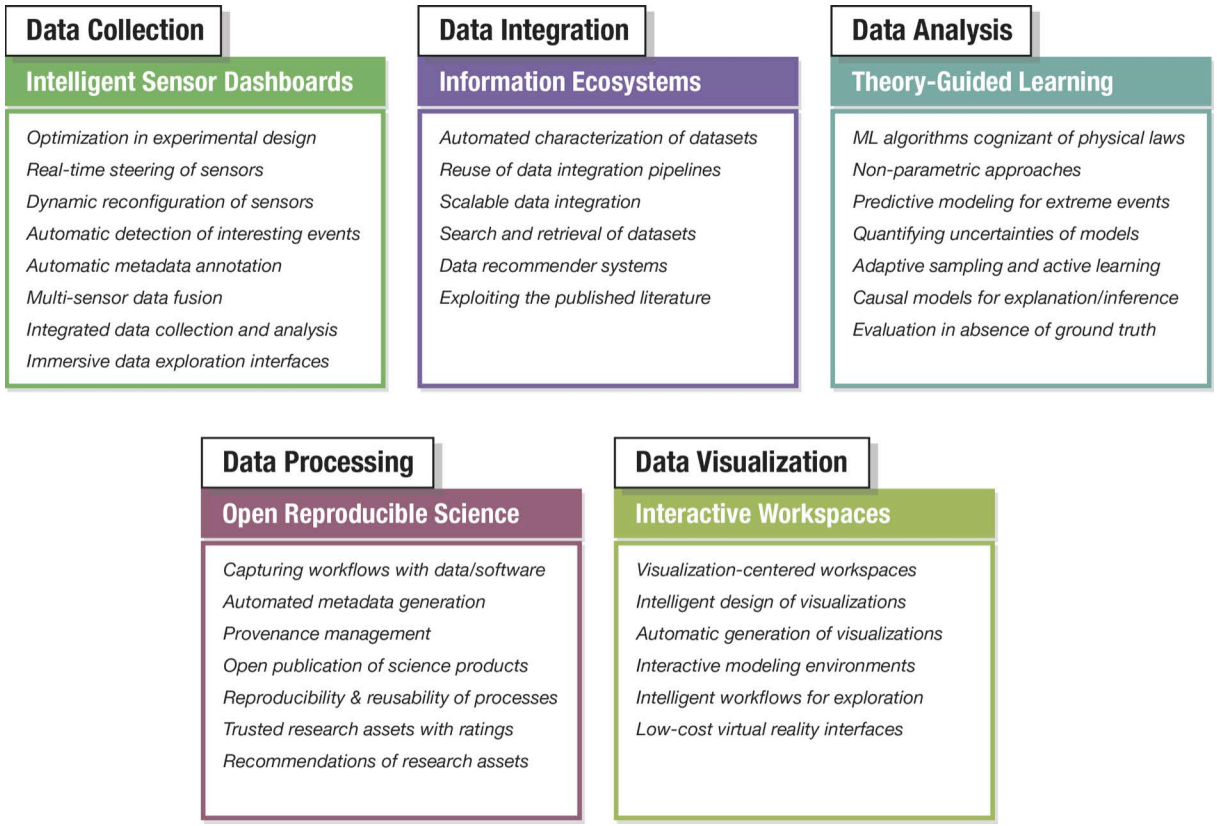

Figure 4. Common ground across disciplines was facilitated through discussions of particular aspects of the data lifecycle in science: data collection, data integration, data processing, data analysis, and data visualization. For each of these topics, workshop participants articulated science drivers, ongoing research in intelligent systems with potential to support geosciences, and additional requirements given the longer-term science drivers. New required capabilities for intelligent systems are summarized here for each topic.

Figure 4 shows an overview of the topics for each of the aspects of the scientific data lifecycle discussed, which are presented in the rest of this section. For each aspect of the data lifecycle, we discuss the science drivers followed by the required capabilities for intelligent systems.

3.1 Data Collection

Collecting observations for all physical parameters everywhere and all the time would be ideal, but is logistically impractical given resource and instrumentation constraints. The goal instead is to amplify how much science is made possible within those constraints, which means increasing the sophistication of existing approaches to data collection.

3.1.1 Science Drivers

Although highly desirable, the collection of data for all observable parameters and at all temporal and spatial scales is not possible. Scientists are always limited by practical resource constraints, and some natural phenomenon will likely remain unobservable. Further, practical constraints affect the ability of almost every instrument to collect observations about various parameters in a time period. For example, an instrument may be positioned to collect an observation at a given time and take too long to reposition to collect a different observation soon after that. Other constraints include the cost of maintaining and operating instruments, battery life, data storage and transmission limits, harsh natural conditions, remote locales, international policies, and other factors that limit the temporal and spatial density of data collection.

"Although highly desirable, the collection of data for all observable parameters and at all temporal and spatial scales is not possible, as scientists are always limited by practical resource constraints. […] The goal is to amplify how much science is made possible within those constraints, which requires increasing the sophistication of existing approaches to data collection."

Resource constraints can be mitigated through optimization theory and statistical experimental design. These techniques lead to data collection approaches that obtain the most information while respecting practical observation constraints and at the minimum cost without compromising the needs for appropriate bounds on uncertainty. Ultimately, the parameters observed, as well as their temporal and spatial scales on which they are sampled, need to match the needs of the prioritized science question(s).

A complementary approach is dynamic sensor steering. Advances in sensor technologies and connectivity offer the ability to make sophisticated real-time decisions regarding how and when to collect and deliver geoscience-related measurements. These include unmanned vehicle sensor platforms, low-cost embedded processors (such as BeagleBone, Intel Edison, and Raspberry Pi), as well as multicore mobile devices to effectively communicate remotely with sensors. Real-time steering of instrument sensors or activation of complex sensor networks could depend on observations, event triggers, and/or model analysis and predictions. Many such techniques are currently being investigated and in some cases used in geoscience, including active and selective heuristic sampling approaches, dynamic sensor steering, dynamic integration from heterogeneous sensors, closed loop synthesis from modeling to sensing in both real world and observing system simulation experiments.

Dynamic sensor steering can increase resolution for data collection specifically when interesting events happen (e.g., a sudden solar flare) while saving precious resources (e.g., energy, bandwidth, storage space, manual instrumentation access), which are scarce in field deployments. The classical pipeline of data collection, data processing, data analysis could be modified to include additional automation and establish feedback between different stages to steer online decisions on data collection and thus provide finer-grained observations when needed.

Adaptive steering sensor configurations combined with model-based sensitivity analysis would provide greater insight into data collection and data processing infrastructures while data collection is ongoing. Advances will include integration of related data collection efforts around the world, including interaction with simulated results of related systems. One example is using sensors for on-the-fly sensitivity analysis, optimization, uncertainty analysis, visualization of data, and modeling the results. These advances will produce more relevant, efficient, integrated data collections.

Of particular interest is the automatic annotation of data during collection with descriptive and contextual metadata. Metadata and provenance annotations enable others to understand the data and reuse them. Moreover, they are instrumental to extend the usability of the data for research problems beyond the original intent. Mobile devices are already being used as cost-effective ways to provide metadata information (e.g., GPS coordinates, camera attributes, etc.) in diverse file formats. Sensor platforms need to be extended to incorporate similar capabilities to enable automated metadata capture. Measures of data error and bias are often underrepresented in metadata, and yet are critical to enable data reuse.

Finally, data integration remains very challenging. Sensors and instruments collect increasingly more sophisticated kinds of data, and a broader range of observations. Geoscientists coming from different field traditions maintain different methodological, semantic, and data management practices. When captured at all, metadata are non-standard and inconsistent. Scientific learning often happens at the point of data aggregation and comparison, and is needed before integration can be successful. This complexity pushes the limits of current data fusion algorithms and automated approaches, and is exacerbated by the lack of metadata on data error and bias.

3.1.2 Required Capabilities for Intelligent Systems

Collaborative research between intelligent systems researchers and geoscientists is needed to provide new capabilities for more sophisticated real-time decisions regarding how and when to collect geoscience-related measurements. Desirable capabilities include:

- *Optimization in experimental design* could be improved through statistical and dynamical techniques such as adaptive sampling and active learning
- *Integrated data collection and analysis* would target more responsive and adaptive scientific data processing approaches
- *Real-time steering of heterogeneous sensors* depending on observations or model output is underway in many geoscience fields but stronger algorithmic underpinnings would be desirable
- *Dynamic reconfiguration of sensor networks and instruments* would enable targeted data collection depending on the science questions

- *Automatic detection of interesting events* requires improved image and pattern recognition approaches to enable dynamic redirection of data collection strategies
- *Automatic metadata annotation* at the collection point would require more sophisticated approaches to provenance and representations of sensors, instruments, platforms, etc.
- *Multi-sensor data fusion* remains very challenging as instruments increase in sophistication and coverage
- *Immersive data exploration interfaces* could be designed to highlight to scientists observation patterns and unusual trends

The vision underlying these capabilities is the autonomic management of sensory environments through dynamically integrating data and models. An *intelligent sensor dashboard* could provide a scientist with the ability to steer data collection to suit the goals of their study using: 1) A high-level real-time view of the observations 2) Integration of preliminary analytic results with applicable reference data and/or models; and 3) The ability to steer data collection assets to suit the goals of their study.

3.2 Data Integration

Earth systems are integrated, but current geoscience datasets and models are not. Our ability to understand the Earth system is heavily dependent on our ability to integrate geoscience data and models across time, space, and discipline.

3.2.1 Science Drivers

Access to integrated data is central to research questions in many areas of the geosciences. Methods and tools that support the integration of geoscience data will accelerate the pace of research and make some advances possible that would not be feasible without wholistic, integrated data and analysis spanning the entire range of geoscience disciplines.

"Earth systems are integrated, but current geoscience data and models are not. Our ability to understand the Earth system is heavily dependent on our ability to integrate geoscience data and models across disciplines."

Integrating geosciences data requires handling intermittent and multi-scale data as well as sparse data since Earth science data derives from a variety of sources with differing coverage, scale, and resolution. Individual geosciences researchers may collect their data using idiosyncratic formats and frequencies, often inconsistently annotated and rarely shared. (Semi-)automated characterization of datasets would lower the barriers to data sharing by enabling data pipelines that could readily transfer data to other users. Identifying and facilitating the use of appropriate

standard representations would enable the integration and discovery of a wealth of very valuable data. Any standardization would need to maintain enough flexibility to accommodate the creativity needed to study new phenomena and data types.

Data integration also requires harmonization of data and model components from across geoscience domains that do not routinely interact with one another (e.g., geological sample-based data sets and geophysical results; see EarthScope initiatives and others for small-scale, regional examples). Mapping terminology and annotations across research teams and domains requires developing shared understanding and corresponding terminology to establish a common interdisciplinary knowledge foundation. Capturing data integration pipelines would enable others to reuse them and integrate new data more consistently. The propagation of metadata over these pipelines would enable derived characterization of the integrated datasets. Scalability becomes an issue given the ever-increasing volumes of data that need to be integrated in geosciences. Given the fundamental nature of most geoscience datasets, space and time are likely to serve as primary organizing principles for integration, though handling of data outside of this framework will be needed in some circumstances.

In addition, researchers often do not know what data are available to integrate with the data that they collect on their own. While open data sharing is increasing, finding data at the desired scales is often a challenge. Connecting scientists with data related to their topical interests or the locations they are investigating is needed to make these connections more efficient and the scientific investigations more productive.

Finally, a great deal of data collected in the past is only available in technical papers. To create a useful dataset from disparate observations reported in different papers, scientists have to identify the papers and extract the data from tables and charts by hand. This is very time consuming and often too costly to do. Methods for finding and delivering data from previously published resources are starting to be developed (see Figure 2), but efforts are in their infancy.

3.2.2 Required Capabilities for Intelligent Systems

Fully automated data integration is far beyond the state of the art. AI researchers would call it an "AI-hard" problem, meaning that it would require solving the central problem of making computers as intelligent as people. While full automation is out of reach, there is much room for improvement over current data integration tools which require a significant investment in time and effort to learn and apply before there is any scientific payoff. More proactive data integration techniques are needed to automate often basic integration tasks that take up valuable researcher time. Desirable capabilities include:

- *(Largely) automated recognition of dataset mappings* would require systems that have knowledge about how data are collected, processed, and represented across scales in different areas in geosciences, including initial scientific purpose and uncertainty estimates as well as precise conceptual mappings across scientific domains
- *Automated reuse of data integration pipelines* would require standardized data sets and systems that capture data integration workflows and reapply them to new datasets

- *Automated metadata generation* during data integration would ensure that data are properly interpreted and would additionally manage metadata created during data collection to avoid loss
- *Scalable data integration* approaches that would require algorithms and hardware interoperability for real-time data integration for large and possibly distributed datasets
- *Search and retrieval of datasets* across disciplines including methods for location- and time-based data retrieval and subsetting
- *Data recommender systems* that can analyze the data that researchers use and suggest relevant datasets and publications as well as connect scientists working in related areas
- *Exploiting the published literature* by discovering, locating, extracting, and integrating knowledge from text, tables, and figures of disparate papers

The results of data and model integration would be fully interconnected, open data sources to support geoscience research. This integrated information ecosystem would also provide fully coupled Earth system models that could be efficiently calibrated with and tested against empirical data from the range of geosciences. Ideally, an integrated ecosystem for data and models would provide a *playground for researchers to support exploration and hypothesis development*, enabling them to better understand not just their own areas, but how their expertise and data sets fit into a larger integrated model of the evolving Earth system. In addition, the data and model integration system would provide a basis for ranking and evaluating incoming data and automatically push data and model results to relevant researchers. Such a system would also facilitate the dynamic tracking and discovery of new research communities that are producing or using related data pertaining to a common problem, a geographic region, or a geological time interval.

3.3 Data Analysis

Complex geoscience research falls into an interesting situation where it can be both data rich and data poor. That is, while it may be possible to collect very large amounts of data about a phenomenon, the data information content may be insignificant compared to that needed to characterize the phenomenon for scientific purposes or practical applications. Scientists need new approaches that supplement data with already existing knowledge about the underlying processes. This would augment researchers' ability to make use of whatever data is available.

3.3.1 Science Drivers

The current scientific understanding of various phenomena and processes in geosciences is limited in part by the methodologies used to create models for such phenomena. Scientists have refined models based on disagreement between model projections and true observations, but the scale and complexity of geoscience phenomena make a manual exercise of model refinement

based on data difficult and time consuming. It also lacks the reproducibility needed to identify advantageous and disadvantageous methods of data-model integrations and model development. Data analysis methods that encompass refinement of models can expedite the process and make it testable, and this synergistic approach is expected to surpass the capabilities of purely model-driven or purely data-driven approaches.

A unique and challenging aspect of geosciences phenomena is that they are extremely high-dimensional, involving thousands of variables along with their spatiotemporal or other types of complex interactions. Data analysis for high-dimensional problems is challenging, although the underlying physics governing the variables substantially constrains the variables involved, leading to low intrinsic dimensionality, e.g., when viewed as a manifold. Further, such physical phenomena often exhibit multi-scale behaviors, which need to be suitably captured to improve our understanding.

"Complex geoscience phenomena fall into an interesting situation where […] while it may be possible to collect very large amounts of data about a phenomenon, the information content may be tiny compared to that needed to characterize the phenomenon... Scientists need new approaches that supplement data with already existing knowledge about the underlying processes, which would augment their ability to exploit whatever amount of data is available."

For predictive modeling purposes, one often has a small number of observations from phenomena of interest, which leads to a `high dimensions, small sample' regime for predictive modeling. Classical statistical machine learning methods (e.g., least squares regression, logistic regression, etc.) assume that the number of samples is much larger than the ambient dimensionality. For the `high dimensions, small samples' regime, available methods include, for example, using field data to produce simple models, using expert knowledge to introduce prior information terms dominant to attain tractable problems, and producing a range of results based on unresolvable variability. The data integration and reproducibility envisioned in this report would allow these methods to be analyzed and compared in ways that no scientist has been able to do up to now. There is also need for new approaches, such as non-parametric approaches that consider the geometry of the problem induced by the physics. For example, greater integration of nonstationary behavior that is often averaged to obtain values to compare with measured values, long-memory processes, nonlinear dynamical systems, and phase transitions. There is also a need to characterize atypical behavior, i.e., behavior in the tails of distributions. This is useful, for example, to understanding extreme precipitation, heat waves, etc.

Physical phenomena in the context of geosciences often exhibit variability due to complex interactions and nonlinear dynamics. Improved methods for quantifying the uncertainties associated with such variability would help improve the scientific understanding of such phenomena. Further, being able to decompose the overall uncertainty into natural/inherent components and structured components may lead to improved models. Uncertainty usually

comes due to limited availability of data, metadata, or both. Reusing models for new locations without quantifiable data is also a source of uncertainty, since we cannot assume that the data and models will be appropriate in the new location.

Directing data collection is needed to sample the space and gather data needed to guide model development. Active sampling, balancing the needs of field science and modeling, would enable the collection of datasets that are most effective for advancing multiple science questions while maintaining realistic costs.

Probabilistic models of high-dimensional physical phenomena can encode the known dependencies among the variables involved, and can then be used to do inference, perform simulations for what-if scenarios under suitable conditioning, attribution of observed phenomenon, among others. Being able to explore the dependency structure of such graphical models can also be potentially leveraged to build causal models of the underlying physical processes explaining the observations.

Finally, there is limited availability of usable ground truth data for many modeling problems. This limits our ability to evaluate and improve models using traditional methods.

3.3.2 Required Capabilities for Intelligent Systems

Over the past two decades, a significant focus for data analysis and machine learning has been on applications arising out of Web applications, such as text, image, and video analysis, recommendation systems, and ad placement among others. While considerable progress has been made on these problems, the models and methods are not directly transferable to data analysis problems in geosciences because of a one key reason: variables and associated phenomenon in geosciences are governed by precise physical laws which need to be taken into consideration for data analysis model and method development. This observation illustrates the need for an entirely new paradigm of data analysis, where the models and methods are cognizant of the physical constraints while developing predictive capabilities or understanding structure from data. Bridging statistical machine learning models and physical characterizations, say based on partial differential equations, will lead to new challenges in estimating the parameterization in the hybrid models, leading to new questions in numerical methods and optimization. Further, models capturing dynamics have to also consider stability of the system, leading to additional constraints under which such numerical methods have to operate. The resulting models also need to be studied from the perspective of predictive skill, i.e., the ability to model observed phenomena accurately, predictability and reliance on initial conditions and other parameters, and overall sensitivity analysis. Desirable capabilities include:

- *Machine learning algorithms cognizant of physical laws* that can use physics to constrain high dimensionality problems
- *Non-parametric approaches* that consider physical constraints and dynamic processes
- *Predictive modeling for extreme situations*, to handle the behaviors in the tails of distributions

- *Quantifying uncertainties of models* based on available data and physics metadata
- *Adaptive sampling and active learning* techniques to steer data collection to sampling the space where a model has large uncertainties
- *Manifold learning techniques for model reduction, uncertainty quantification, compensation for model error.*
- *Information theoretic learning for dealing with non-Gaussian, high-dimensional inference.*
- *Causal models* that can explain observations and can be used to make inferences
- *Methodologies for evaluation in the absence of ground truth data*

3.4 Data Processing

Studying geoscience phenomena requires the ability to process data that is incredibly diverse, encompassing multiple disciplines, scales, and methodologies. Geoscientists need data processing frameworks that minimize human effort to harness this diversity.

3.4.1 Science Drivers

"Studying geoscience phenomena requires the ability to process data that is incredibly diverse, encompassing multiple disciplines, scales, and methodologies. Geoscientists need data processing frameworks that minimize human effort to harness this diversity."

Data processing and data analysis should be as efficient as possible. A geoscientist may run the same algorithms multiple times by tuning different parameters to explore various possible paths and conduct correlations, until reaching satisfactory results. Typically, data analytics is a multi-step procedure, often part of a *scientific workflow*. A scientific workflow refers to a formal way of defining, automating, and repeating such multi-step computational procedures. To enable the reproducibility and reusability of scientific workflows, a wholistic collection of metadata has to be captured. Metadata should include scientific intent, algorithms and models, ordering of algorithm and model application, workflow versioning, input datasets, how data were collected, parameter settings, boundary conditions, and any other relevant aspects that document the provenance of new results.

Access to provenance of data, including where and how they were collected, under what conditions, for what intent, by whom, and application of any processing from raw values, is of critical importance in geoscience research. Correctly representing, integrating, and propagating provenance is an important precursor to data sharing and reuse. For example, knowing the weather conditions in which sensor measurements were taken or are about to be taken is critical for cleaning and selecting the data. Understanding the fidelity and limitations of a particular sensor platform is vital for appropriately representing and handling the inherent uncertainty.

Provenance is also needed for citizen science and crowdsourcing. Given the availability of provenance standards, including W3C PROV and ISO 19115, provenance information should be routinely recorded in a structured, searchable way by geoscientists.

The scientific community is increasingly moving towards open publication, and when a research article is published all related metadata are published as well. The provenance and workflow that are often implicit in traditional research papers should be published explicitly in an open and accessible manner, using standards that maximize their dissemination and reuse. Reproducibility is a necessary quality of scientific research. In order to verify any new results, other geoscientists should be able to re-run experiments.

The amount of data available to scientists continues to grow, but data are in diverse formats and repositories, available in non-standard projections and scales, have inconsistent documentation, and cross-disciplinary boundaries with conceptual and semantic shifts. Therefore it takes a lot of effort to both articulate science needs and to them find relevant data to explore the problem at hand. Recommender systems that understand what scientists are trying to do could anticipate what data, metadata, models, workflows, and people may be useful for their task. Even when relevant resources are found, it is often difficult for a scientist to determine if that resource is trustworthy. The ability to access usage and relevance meta-information would greatly improve the confidence and trust on the resources found.

3.4.2 Required Capabilities for Intelligent Systems

Desirable capabilities include:

- *Capturing workflows and associated data and software,* so that they can be automatically discovered, processed, repeated, and reused by others
- *(Largely) automated creation of metadata for new datasets as well as for legacy data* would require new approaches for recognizing common types of data in geosciences and understanding what metadata are needed to answer different geoscience questions
- *Provenance management,* so that provenance records can be searched and analyzed
- *Open publication of science products*, which includes representing metadata for new data products including their associated workflows and all the software and data sources used
- *Reproducibility and reusability* of data analytics processes
- *Recommender systems* that link appropriate people, data, metadata, models, and workflows with context awareness
- *Trusted data*, so that data can be evaluated based on its provenance, its quality, and its utility

3.5 Data Visualization

Geoscientists are visual thinkers and visualizations are key aids to understanding very complex phenomena. Traditionally, visualizations are used to render the results of a model or data analysis process. Instead, visualizations should be thoroughly embedded in all science interfaces, and advance approaches for visualizing data should include visualization of models and other relevant science contexts.

3.5.1 Science Drivers

> ***"Geoscientists are visual thinkers […], visualizations should be thoroughly embedded in all science interfaces, and go beyond visualization of data to include visualization of models and other relevant science context."***

Visualizations should be ubiquitous throughout the entire science process, from data collection and processing, through model building, to publication and presentation. Interactive workspaces that would allow geoscientists to manipulate visualizations of data and models would offer different perspectives and improve their understanding of the underlying phenomena. Current practice is to use visualizations at the end of the process to show results, but facilitating the pervasive use of visualizations throughout intermediate stages of the process would be very beneficial. One barrier is that generating visualizations is often difficult and time consuming. Intelligent assistance to create visualizations that are appropriate for given data or process would make them more ubiquitous and effective. Automating the generation of visualizations using spatial or temporal references would enable scientists to get an integrated view of diverse information.

Geoscientists typically create models and validate them with data, then iterate through this process to improve the models. This process is currently not interactive, that is, it takes time to go from a refinement in the model to seeing its impact on the quality of the model. Geoscientists would greatly benefit from interactive model building, refinement, and validation. By interacting with models and data simultaneously, they would gain novel perspectives on comparing models, specifying knowns and unknowns, identifying model biases, and testing hypotheses.

It may be hard for a geoscientist to figure out what visualizations to use, what strategy to follow to explore datasets, and how to combine visualizations with data analysis. To assist scientists in exploring and processing data, interactive systems could provide guidance based on the steps in common scientific workflows.

Finally, virtual reality interfaces are ideal for navigation of spatio-temporal information. The advent of low-cost virtual reality devices can make these systems easier to come by. More direct interaction with information and direct manipulation to get alternative perspectives can help a geoscientist gain a better understanding of physical phenomena.

3.5.2 Required Capabilities for Intelligent Systems

Intelligent user interfaces and visualization systems must be designed to support geoscientists in understanding the complex phenomena they study. The following capabilities would fundamentally change how scientists interact with data:

- *Visualization-centered workspaces* that allow assist scientists to manipulate visualizations throughout the entire science process
- *Intelligent design of visualizations* that can generate rich multi-dimensional or multi-scale visualizations that fit a given scientist's problem
- *Automatic generation of visualizations* grounded on spatial and temporal coordinates
- *Interactive modeling environments* that integrate models with data, model parameters, model results, and hypothesis specifications
- *Intelligent workflow systems* that can guide scientists to explore and analyze data
- *Low-cost virtual reality interfaces* that allow scientists to explore 4D datasets routinely and interactively

4. Geoscience Challenges Requiring Innovations in Intelligent Systems

The challenges in geoscience research have been reviewed and described in several recent reports [NSF 2014; NRC 2014a; NRC 2014b; NRC 2013]. Geosciences encompasses and describes the vast scales of temporal and spatial systems of Earth. Concomitant with these scales come a notable diversity of data, knowledge, and scientific approaches. Geoscience problems do not typically adhere to simple and symmetric models. Earth systems phenomena are non-linear, heterogeneous, and highly dynamic. Geosciences research will also be challenged by extreme events and long-term shifts in Earth systems [NSF 2014]. In addition, recent unprecedented increases in data availability together with a stronger emphasis on societal drivers emphasize the need for research that crosses over traditional knowledge boundaries.

This section reflects on geosciences challenges that could be tackled through new capabilities resulting from future innovations in intelligent systems. We describe the needs and potential impact at several scales:

- **Site-level needs**, where new research in intelligent sensors poses new opportunities, particularly in hard to access areas,
- **Regional-level needs**, where efficient techniques are needed to integrate data from disparate locations, data types, and collection efforts within a wide area,
- **Global-level needs**, where pattern recognition and analytical tools need to be applied to data and models to study wholistic phenomena of the Earth system, and
- **Layered needs**, where interactive workspaces will accelerate research by supporting synthesis of information and knowledge across geosciences disciplines.

Figure 5. Needs and potential impact at different scales at which significant new avenues of research in geosciences would be open through advances in intelligent systems, illustrated with case examples. From left to right: 1) site-scale, 2) local scale, 3) global scale, and 4) layered wholistic scale.

Figure 5 illustrates the challenges for each of these scales, which will be described in more detail in the rest of this section by presenting specific case instances followed by a discussion of how these challenges are actually common among geosciences sub-disciplines.

4.1 Polar Sciences

The rapid changes occurring in polar regions present the need to understand a broad range of complex processes and interactions. These processes and interactions span multiple domains and research communities, from oceanography and ecology to large-scale atmospheric modeling. Beyond the basic science needs, multiple stakeholder groups and participants bring further complexity by adding inherent connections to the social realm. Knowledge needs to be easily and quickly translated across multiple community boundaries without losing accuracy given the increasing speed and importance of decisions at the poles. Compounding the domain and stakeholder complexities, polar regions are vast, challenging, and expensive places to work. However, with recent advances in observational, logistical, and analytical technologies, polar and cryospheric sciences are quickly progressing. Polar research addresses many of the priorities identified across the geosciences. Major research frontiers in polar sciences include [NSF 2014]:

- Understanding sea level rise, particularly changes related to melting and loss of major ice sheets
- Predicting Arctic sea ice extent and thickness changes with impacts for shipping, geopolitics, and possibly storm patterns
- Improving water resources considerations with understanding of seasonal patterns for important components of the cryosphere, such as snowpack
- Modeling ice-ocean-atmosphere ecosystem interactions and processes
- Assessing impacts of the changing Arctic environment on indigenous communities including but not limited to erosion prediction, permafrost degradation, and subsistence hunting

4.1.1 Exemplifying Site-Level Needs: Forecasting Rates of Sea Level Change

Polar scientists, along with atmospheric and ocean scientists, face an urgent need to understand sea level rise around the globe. To do this, sensing in remote or extreme locations can provide critical information about conditions and relationships across settings. For example, conditions along and under ice shelves are extremely challenging to observe but provide critical data to understand changes to the ice caps. Yet ice shelf environments represent extreme environments for sampling and sensing. Current efforts to collect sensed data are limited and use tethered robots with traditional sampling frequency and collection limitations. New research on intelligent sensors would support selective data collection, on-board processing, and adaptive sensor steering. New submersible robotic platforms could detect and respond to interesting situations while adjusting sensing frequencies that could be triggered depending on the data being collected in real time.

With autonomous sensors, Polar geoscientists would be able to identify when the platform arrives at a transition zone and collect information that is key to understand what is happening under ice shelves. The ability to collect extensive data about conditions at or near the ice shelves will inform our understanding about changes in ocean circulation patterns, as well as feedbacks with wind circulation. Achieving these outcomes requires advances in data collection to observe and characterize complex physical processes that combine turbulence, dispersion, diffusion, etc. This includes:

- Robust sensor platforms to collect data in extreme and difficult environments
- Automatic detection of interesting events
- Real-time steering of sensors
- Adapt sampling through dynamic reconfiguration
- Integrated data collection and analysis

The benefits of intelligent sensing would accrue rapidly on important issues, like understanding sea level rise processes and feedbacks, and would enable new geosciences research to characterize and quantify rates of change with data from transition zones across numerous geologic settings.

4.2 Earth Sciences

The Earth Science community focuses on understanding the dynamics of the Earth. Studies of the interior of the Earth, or deep Earth, include wide-ranging topics such as tectonics, seismology, magnetic or gravity fields, and volcanic activity. Studies of the near-surface Earth largely focus on the critical zone that is the constantly changing layer where rock, soil, water, air, and living organisms come into contact. It includes most of the hydrologic cycle, the carbon cycle, the food production cycle, and the energy cycle.

All Earth Sciences research uses complex geological data to address a wide range of temporal and spatial scales, to characterize the complex three-dimensional geometry of some geological structures, and to make temporal inferences from spatial observations. Frequently, Earth Sciences inquiry involves geologic hazards or integrated problems related to risks and resources that are key to human systems and societal concerns. Contemporary societal interests can include topics such as the formation of minerals or water resources central to modern life and the geologic events like earthquakes or volcanic activity. Examples of major research frontiers in Earth Science include [NSF 2014; NRC 2012; 2012a]:

- Understanding when and why the Earth's core formed and how the geo-dynamo originated
- Understanding how the origin of life was constrained by the timing and nature of early Earth's atmosphere, oceans, and tectonics
- Predicting geohazards such as earthquake events and volcanic eruptions

- Modeling water cycle processes in relation to mass and energy transport with attention to rates, patterns, distribution, and impacts of human behavior
- Enhancing societal resilience and decision making, particularly in relation to the availability of resources and the processes that influence or perturb dynamics in near surface systems, such as ecosystems, watersheds, coastal systems, and urban environments.

4.2.1 Exemplifying Wide-Area Needs: Unlocking Deep Earth Time

Earth Science opportunities are broad and deep. Earth Science researchers are frequently faced by data-sparse situations and problems. While collecting data from the field is done by individuals in select locations, the problems under consideration cover spatially vast regions of the planet. Moreover, scientists have been collecting data at different times in different places and reporting results in separate repositories and often unconnected publications. This has resulted in a poorly connected collection of information that makes wide-area analyses extremely difficult and is impossible to reproduce.

To unravel significant questions about topics, such as Deep Earth Time, geoscientists need mechanisms and tools to enhance the interconnections among previous and future studies. Effective data integration approaches can result in highly interlinked information ecosystems that would enable great advances in our understanding of Deep Earth Time. These information ecosystems would enable:

- Rapid sharing and discovery for many sites and data types, using data collection and provenance annotations that enable integration of observations from the field directly into reusable information repositories
- Intelligent curation support and knowledge representation to suggest interconnections among single-investigator data
- Interactive mapping aides that identify and suggest geologic features and interpretations based on repositories of known relationships (e.g. geophysical, stratigraphic, paleoenvironmental)
- Recommending potentially relevant data and information given the location and scientific goals of a study
- Event and pattern detection in geo-referenced data to aid understanding of the signals and structural relationships that may explain how plate tectonics start
- Tracing events from early planet formation using integrated field observations to characterize how tectonic processes work

An integrated information ecosystem can support connections among scientists within and across disciplines (e.g. biology and geosciences data) to augment work that is site specific but also elevate research to broader and more generalized questions.

4.3 Atmospheric and Geospace Sciences

Atmospheric and geospace science research aims to improve understanding of the Earth's atmosphere and its interdependencies with all of the other Earth components, and to understand the important physical dynamics, relationships and coupling between the incident solar wind stream, and the magnetosphere, ionosphere and thermosphere of the Earth. Atmospheric research investigates phenomena operating from planetary to micro spatial scales and from millennia to microseconds.

Atmospheric and geospace science research will benefit from developments in sensing including appropriate utilization of autonomous and unmanned sensor platforms, efficient utilization of crowd-sourced data, observing system design, inference methodology, sampling strategies, extracting structure and composition, data analytic tools, and in diagnosing and compensating for model error, in addition to interactive visualizations that enable data fusion and views of multi-dimensional information. Geospace sciences, as generally framed within the NSF programs, are focused on advancing our understanding of near-Earth space using two complementary research areas. First is the basic physical understanding of the geospace environment. Second is understanding, predicting and mitigating negative impacts of space weather upon society. These two areas share many of the same research challenges and advances in understanding the geospace environment are requisite to understand implications of space weather. Researchers use observational, theoretical, computational, and laboratory capabilities to advance understanding of these dynamics, over the entire range of spatial (1000s of kilometers to centimeters)-temporal (years to micro-seconds) scales of interest, as well as over the entire range of geophysical conditions (solar cycle, seasons, time of day, magnetic activity). Because of the coupling between different physical systems, the external forcing and the spatio-temporal scales involved, geospace physics can benefit from intelligent systems research at every step along the chain of analysis, understanding, and construction of new knowledge about geospace science systems.

Major research frontiers in atmospheric sciences and geospace sciences are driven by scientific goals [NRC 2013] that include:

- Characterize the physics, chemistry, and dynamics of the earth's upper and lower atmosphere and its interactions with land-ocean-hydrosphere-ice systems
- Understand the impact of climate processes and variations on diverse applications such as ecosystem productivity and biodiversity
- Characterize natural and anthropogenically-perturbed local, regional and global cycles of gases and particles in the earth's atmosphere
- Integrate improved observational and modeling capabilities across relevant temporal and spatial scales to address societally relevant issues that affect public safety and the national economy now and in the future.
- Explore the processes that occur within the heliosphere
- Understand the origins and variations of Sun activity with predictive capacity

- Build understanding of the Sun, Earth, and heliosphere as coupled systems
- Expand knowledge of how the Sun interacts with the solar system and interstellar medium

4.3.1 Exemplifying Global Needs: Predictive capacity for critical events

Machine learning techniques have been applied in atmospheric and geospace research with great success (see Figure 1 for an example). However, atmospheric and space sciences are tackling challenges that go beyond existing machine learning techniques. There is a need for machine learning techniques that better account for nonlinearity and high dimensionality, and that are robust to non-Gaussian uncertainties for applications ranging from weather to climate scale phenomena.

Although the data collected is very large, it is miniscule given the complexity of the phenomenon and existing machine learning techniques are not very effective. New machine learning algorithms could augment the data available with knowledge about physical laws underlying the phenomena to generate effective models. Advanced nonlinear decomposition approaches could be used to define principal modes of transient or localized phenomena, from hurricanes to the El Niño Southern Oscillation. New techniques could aid in the diagnosis and correction of model error, and characterization of limitations of parameterization (e.g. convective parameterization). The goal would be to reduce or eliminate error propagation or assure that uncertainty within derived measures is flagged and managed. Machine learning advances are needed to assess long-term risk in a changing climate and plan mitigation strategies with robust uncertainty quantification methods. Similarly, geospace researchers would benefit from machine learning paired various workflows, such as intelligent sensing and sampling tools to recommend adequate sampling regimes, or assist with estimating state variables with active learning algorithms to assimilate and integrate heterogeneous measurements. To advance machine learning for atmospheric and geospace science applications will require:

- Novel frameworks for combining and integrating models would advance atmospheric and geospace research, but present many research challenges in machine learning.
- Bayesian inference of stochastic dynamical models would enable comparing scientific hypotheses and models in a rigorous fashion, and ranking and combining models or suggesting better models.
- Frameworks for coupling models between different physical regions, scales and systems will enable the dynamical analysis of complex event patterns and phase transitions.

4.4 Ocean Sciences

The ocean programs strive to enable and support research across the physical, chemical, biological, and geological processes that operate throughout the global ocean [NSF 2014]. Science priorities in ocean research are focused on understanding processes and responses related to sea level change, coastal and estuarine ecosystems, climate systems, marine resilience

and food webs, ocean basin formation, and characterization of geohazards and subseafloor environments [NSF 2014]. Major research frontiers in ocean sciences include:

- Understanding coastal ecosystems: research is needed to model, quantify, and discover fundamental ecosystem functioning, including their interactions with oceanic flows, transports and mixing, and coastal aquifers
- Predicting sea level variability: understanding the factors driving sea-level changes and predicting their impacts, both globally and regionally, as well as their association with melting and loss of ice sheets
- Effects of land-ocean-atmospheric-ice interactions: research on the multi-scale and nonlinear interactions at varying time and space and scales that occur at the interface
- Response to anthropogenic activity: effects in marine ecosystems of fisheries and other resource consumption, discharges, and coastal urbanization

4.4.1 Exemplifying Layered Needs: Ocean-Land-Atmosphere-Ice Interactions

The synthesis of models from vast amounts of multidisciplinary data in geosciences represent a culminating challenge to scientific research communities. Current research practice has achieved significant advances, but approaches to tackle broader scales and work across disciplines will usher in a new era of geoscience inquiry. This is the case with the analysis of the interactions between ocean, ice, atmosphere, and land phenomena.

Global drivers and shifts in the rates of change and response in Earth systems are quickly outstripping our ability to translate lessons from the past into expectations for future behavior and response in Earth systems. Interactive workspaces are needed to support synthesis and integration of information and knowledge across research disciplines and sectors and they are part of the transition towards interactive and collaborative science. This includes:

- Interactive exploration of interlinked data, models, and context
- Collaborative workflows for joint exploration and coupling of separate models, phenomena, regions, and events
- Automated generation of visualizations
- Low-cost virtual reality environments and visualization-centered workspaces to integrate models.

These capabilities will enable ocean sciences research directed towards identifying global drivers in ocean biogeochemical and physical processes, and their interactions with other processes, to understand shifts and rates of change.

4.5 Enabling Wholistic Research on the Earth as a System

The pace of geosciences investigations today can hardly keep up with the urgency presented by societal needs to manage natural resources, respond to geohazards, and understand the long-t

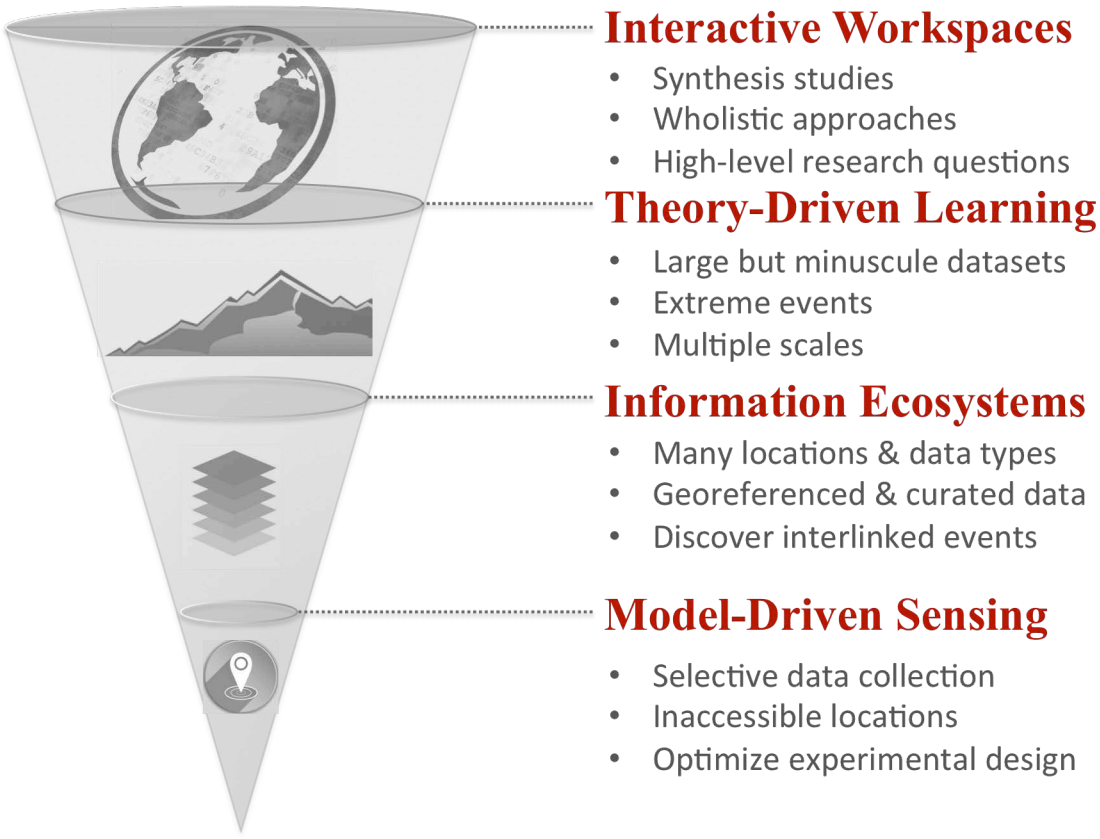

Figure 6. Significant new avenues of research in geosciences would be open through advances in intelligent systems through: 1) model-driven sensors that collect data selectively at hard to access locations in optimal patterns; 2) information ecosystems that integrate data from many locations and types to enable detection of interlinked phenomena; 3) theory-driven learning to supplement limited data with knowledge about physical laws; 4) interactive workspaces that enable geoscientists to synthesize diverse models and explore more wholistic scenarios across many layers of information at diverse scales.

erm effects of human activities on the planet. Studying the Earth as a system requires scaling up our ability to collect data where and when it matters, to integrate isolated observations into broader studies, to create models in the absence of comprehensive data, and to synthesize models from multiple disciplines and scales. Advances in intelligent systems to develop more robust sensor platforms, more effective information integration, more capable machine learning algorithms, and intelligent interactive environments have the potential to significantly transform geosciences research practices and expand the nature of the problems under study. Collaborations between intelligent systems and geoscience researchers can be logically organized at various scales as shown in Figure 6 as follows:

Table 1. An overview of new capabilities for geosciences research that would result from innovations in intelligent systems.

Capabilities	**Knowledge & Capture**	**Robotics & Sensing**	**Information Integration**	**Machine Learning**	**Intelligent User Interfaces**
Exploit sensing and sampling tools to collect data more frequently at hard to reach sites	✓	✓	✓	✓	
Enable adaptive collection rates and interleave processing methods based on predictive dynamics	✓	✓	✓	✓	
Design of optimal sampling strategies for high value data collection at critical locations and phenomena	✓	✓		✓	
Synthesis of diverse subjective single-investigator observations and interpretations based on field data	✓		✓	✓	✓
Interactive mapping aides that identify and suggest geologic features and interpretations based on repositories of known relationships (e.g. geophysical, stratigraphic, paleoenvironmental)	✓		✓	✓	✓
Machine learning algorithms for dynamic non-linear processes, sensitivity analysis, adaptive estimation	✓			✓	
Explore remote field sites in virtual and augmented reality environments, creating digital field access	✓	✓	✓		✓
Interactive exploration of coupling of different physical regions, scales and systems with visualizations representing high dimensionality information (e.g. 4D or more)	✓	✓	✓	✓	✓
Adaptive computational methods and systems that evolve based on the dynamics and measurements collected, using continuous updates to resolution and models	✓	✓	✓	✓	✓
Optimal sensing with collaborative swarms of heterogeneous autonomous platforms (e.g., AUVs, gliders, ships, moorings, remote sensing) with context knowledge about environment, uncertainties, and implications for sensing tasks	✓	✓	✓	✓	✓
Scientific support systems to compare hypotheses, understand complex data and models, provide model ranking, or suggesting better models	✓		✓	✓	✓
Automated and interactive visualizations of multivariate datasets, assisted discovery and extraction of complex features, understanding of uncertainties and probability density functions	✓		✓	✓	✓
Characterization of uncertainties, assessment of uncertainty sources and propagation	✓		✓	✓	✓

- Model-driven sensing: to support data collection in extreme or remote environments, difficult to monitor conditions, or phenomena with limited observability. Optimization strategies for data collection and adaptive sampling regimes would result in data of maximal utility to scientists.
- Information Ecosystems: to support data and model integration across different repositories across disciplines. Improved information discovery capabilities would further inform systems-level inquiry and leverage previous research particularly to support information aggregation through georeferenced and curated datasets and integration across scales.
- Theory-driven learning: to support geoscience researchers working on generating models from datasets that are large and yet insufficient due to their high dimensionality or multi-scale nature, so they need to be combined with theories about physical laws and other geosciences knowledge. In addition, new machine learning methods would be needed to study hard to observe or extreme events.
- Interactive Workspaces: to support exploration and hypothesis development, enabling researchers to better understand not just their own areas of specialization but expand into larger integrated models of the evolving Earth system. Such frameworks would also facilitate the dynamic tracking and discovery of new research communities that are producing data or models pertaining to a common problem or geological time interval. Layered needs, where interactive workspaces will accelerate research by supporting synthesis and integration of information and knowledge across research disciplines and sectors.

Table 1 summarizes major aspects of the above capabilities, together with an indication of the major areas of intelligent systems where we anticipate that research advances will be required.

5. A Roadmap for Intelligent Systems Research with Benefits to Geosciences

With the unprecedented increase in observational and model data being collected about physical processes on the Earth, geosciences is rapidly transcending from a small data to a big data era. This has been made possible through advancements in data collection technologies and through growing access to computing resources. The growing availability of Earth system data offers an immense opportunity for intelligent systems research to accelerate advancements in geosciences and vice versa.

The promise of intelligent systems research for geosciences is heightened by the recent success of traditional intelligent systems methods in several commercial domains involving massive datasets, such as product recommenders and advertising. However, geoscience datasets exhibit a variety of unique characteristics that differentiate them from big datasets in commercial domains. These are summarized in Figure 7. Geoscience datasets are immensely heterogeneous,

A new generation of geoscience-aware intelligent systems with novel forms of reasoning and learning will be needed to address the challenging features of geoscience research questions:

- Spatio-temporal structure
- Intermittent, sparse data
- Heterogeneous and dispersed data
- Small sample size
- Multi-resolution, multi-scale observations
- High dimensionality
- Tolerance of measurement
- Uncertainty at all stages
- Process-centered models
- Combine diverse (physical, geological, chemical, biological, ecological, anthropogenic) phenomena
- Objects and processes with amorphous spatial/temporal boundaries
- Contextualized by rich background knowledge
- Hard to understand information
- Lack of ground truth

Figure 7. Geoscience data exhibits a variety of differentiating and challenging characteristics that require new research to extend the traditional intelligent systems approaches that are successfully used in commercial domains involving massive datasets.

are usually spatio-temporal, and the phenomena or objects of interest do not have crisp boundaries. For example, ocean eddies and hurricanes have amorphous spatio-temporal boundaries that appear as patterns in continuous variables, such as the sea surface height. Geoscience datasets capture information about both well-known and little understood physical processes and relationships, which show varying characteristics in different regions of the world due to differences in geographies, climatic conditions, seasonal cycles etc. Even the relatively homogeneous 'big data' from remote sensing suffer from a high degree of uncertainty, incompleteness, and lack of comprehensive tools for easy use. These characteristics restrict the immediate application of existing intelligent systems approaches in geosciences, as they have been traditionally developed for relatively noise-free datasets. An additional challenge in the use of traditional intelligent systems approaches is the small sample-size problem that appears frequently in geoscience applications, due to the paucity of reliable ground truth or the unavailability of observations, e.g. in paleo-climate studies. This problem is especially severe when the geophysical phenomena are complex and non-linear, requiring large sample sizes for significance testing. Additionally, geoscience data are frequently collected to find or better understand a natural phenomena or relationship; the resulting data are observations, but it isn't always clear what those observations reveal. The geosciences are sciences of discovery, which means requirements, rules, and relationships among different systems are constantly changing.

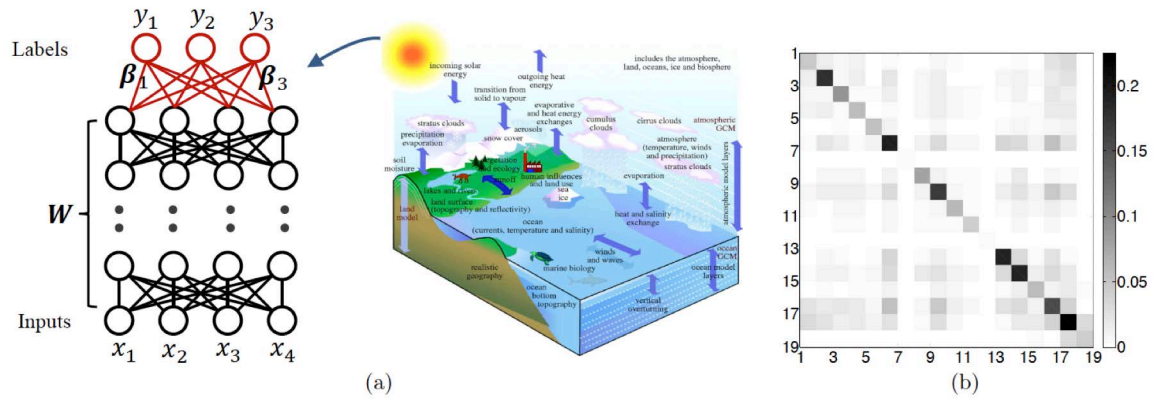

Figure 8. An example of how geoscience processes could be represented and augmented to improve machine learning methods [Che et al 2015]. A miniature illustration of a deep network with the regularization on predicted labels in geo-applications through a Laplacian prior constructed (a) based on physical processes and (b) based on the location adjacency matrix. The regularization is applied to the output layer of the network so that the physical processes that are highly correlated with each other or the locations adjacent to each other can share similar latent layers. Incorporating this knowledge about physical processes into the machine learning models results in better prediction performance (at least 3-5% improvement), especially when the amount of observation data is limited or when there are many missing observations.

Geosciences algorithms, workflows, and standards are an inherently moving target, adding the need for ongoing technical change as well.

Another fundamental property that differentiates geosciences from commercial domains where 'big data' approaches have been used with great success is the fact that geoscience processes are strongly guided by scientific principles. These scientific principles, available in the form of domain knowledge, can guide the process of knowledge discovery from geoscience datasets. For example, encapsulating knowledge about the physical processes governing Earth system datasets can help constrain the learning of complex non-linear relationships in geoscience applications, ensuring theoretically consistent results. What is needed is an approach that leverages the advances in data-driven research yet constrains both the methods and the interpretation of the results through scientific principles that govern the domain.

In order to address geosciences challenges involving complex multi-scale, multi-process phenomena, scientists will need intelligent systems incorporating cutting edge technology and the scientist's expertise, context, and experiences. Intelligent systems need to incorporate process-centered geoscience knowledge about phenomena that combine physical, geological, chemical, biological, ecological, and anthropomorphic factors. This will result in a new generation of knowledge-rich intelligent systems enabling novel forms of reasoning and learning about geosciences data.

Figure 8 shows an example of representations of physical processes can guide machine learning (from [Che et al 2015]). In this example, a multilayer neural network was constructed

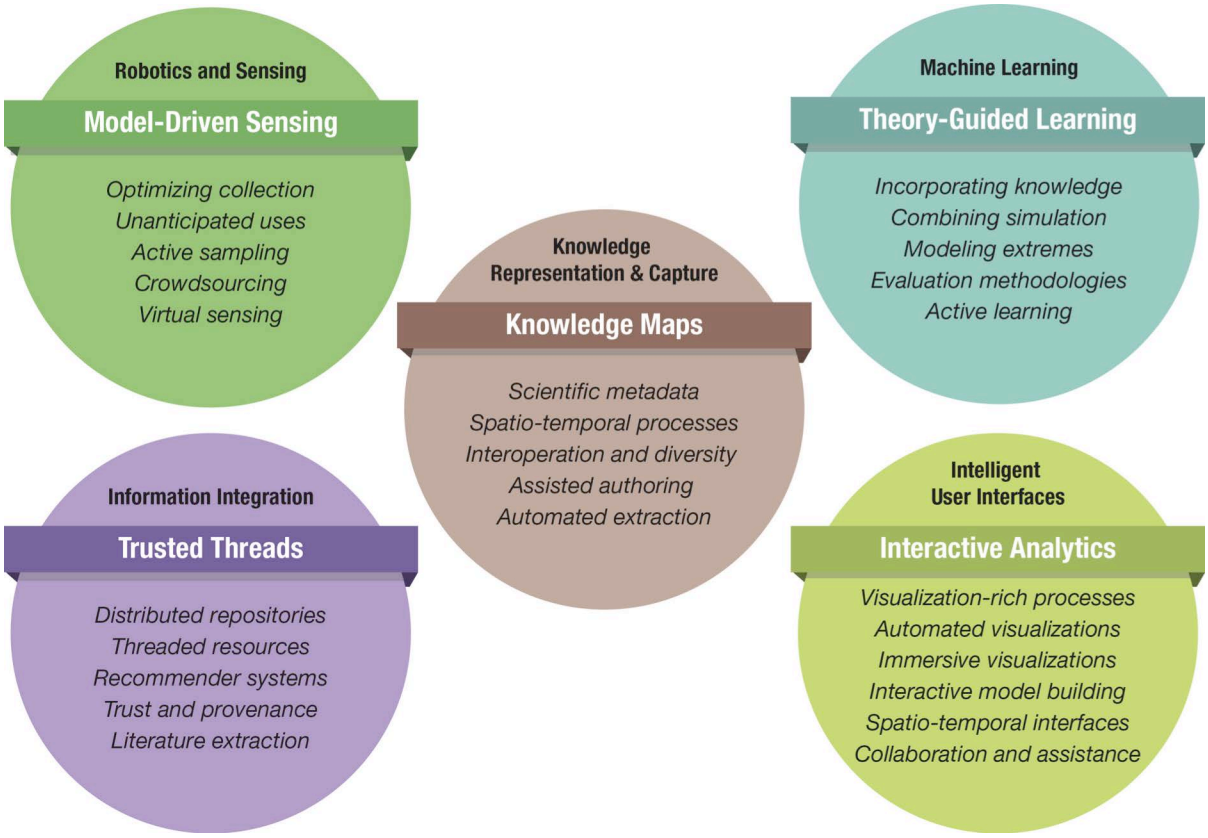

Figure 9. Major areas for intelligent systems research that could have major impact in geosciences. New knowledge representations could capture scientific knowledge that could be used for sensors and robotics, information integration, machine learning, and intelligent user interfaces.

with a Laplacian regularizer. The Laplacian regularizer is constructed from the adjacency matrices based on the similarity between physical processes or location adjacency. In this way, the neural networks can automatically learn the appropriate weights so that the physical processes that are highly correlated with each other or the locations adjacent to each other can share similar latent layers.

These "**geoscience-aware intelligent systems**" pose novel problems for intelligent systems researchers. Workshop participants converged on five major research areas:

1. *Knowledge Representation*: Capturing scientific knowledge in the form of geoscience processes (physical, geological, chemical, biological, and ecological and anthropomorphic) will push the limits of the state of the art.
2. *Sensing and Robotics*: Scientific knowledge should be used to guide what data needs to be prioritized and collected.

3. *Information Integration*: Geosciences processes need to be represented in a "geophysical system of systems" where all data and knowledge are interconnected.
4. *Machine Learning*: Algorithms need to be enriched with models of the relevant geoscience processes.
5. *Interfaces and Interactive Systems*: Knowledge-rich user models should provide context for the interactions.

All these areas cannot be investigated separately as they are interdependent. For example, improvements in sensing will facilitate learning, richer representations will facilitate information integration, and knowledge-guided learning algorithms will lead to better interfaces.

Figure 9 shows an overview of the major research areas presented in the rest of this section.

5.1 Knowledge Representation and Capture

In order to create geoscience-aware intelligent systems, scientific knowledge relevant to those geoscience processes must be explicitly represented, captured, and shared.

5.1.1 Research Directions

Representing Scientific Metadata. Geoscientists are collecting more data than ever before, but raw data sitting on isolated servers is of little utility. Recent work on semantic and Linked Open Data standards enables publishing datasets in Web standard formats with open access licenses, and describing their semantics with metadata that maps the data to an ontology that describes domain concepts. They also enable creating links among datasets to further interoperability. While semantics, ontological representations, scientifically accurate concept mappings across domains, and the application of Linked Open Data are all areas of active research in the data curation and informatics communities, these techniques have the potential to improve open data reuse and access. Progress in data curation, data management, informatics, and cyberinfrastructure will open the door to new approaches for automatically integrating data across sources and perform analysis on the data without a great deal of manual effort. Also needed are new techniques to automatically infer semantic structure from raw data that can be used to integrate, analyze, and visualize large datasets. Novel techniques that map across information spaces will enable linking across research disciplines.

Capturing Scientific Knowledge. An even greater challenge is representing the ever-evolving, uncertain, complex, and dynamic aspects of scientific knowledge and information. While ontologies are growing in use to state basic relations between objects, existing ontologies need to be extended to represent geoscience processes with buy-in from many diverse communities and capabilities of documenting, versioning, and representing various forms, such as spatio-temporal processes interacting with each other and multi-scale phenomena. These representations can be broadly linked to existing data and ontological concepts with actionable authority. Important

challenges will arise in representing mathematical concepts, dynamic processes, uncertainty, and other aspects of a constantly growing scientific knowledgebase. These representations need to be expressive enough to capture complex scientific knowledge, but they also need to support scalable reasoning that integrates disparate knowledge at different scales, and scientists need to understand the representations enough to trust the outcomes.

Interoperation of Diverse Scientific Knowledge. Scientific knowledge comes in many forms that use different tacit and explicit representations: hypotheses, models, theories, equations, assumptions, data characterizations, etc. Certainly, the scientists themselves construct and encode the various representations and forms to disseminate understanding. All these different perspectives come together with personal expertise to allow scientists to analyze different aspects of complex phenomena. However, these representations are all interrelated, and it should be possible to translate knowledge fluidly as needed from one representation to another. A major research challenge is the seamless interoperation of alternative representations of scientific knowledge, from descriptive to taxonomic to mathematical, from facts to interpretation and alternative hypotheses, from small to larger scale, and from isolated processes to complex integrated phenomena.

Authoring Scientific Knowledge Collaboratively. Formal knowledge representation languages, especially if they are expressive and complex, are not easily accessible to scientists for encoding understanding. A major challenge will be creating authoring tools that enable scientists to create, interlink, reuse, and disseminate knowledge about geoscience processes. Scientific knowledge needs to be updated continuously, allow for alternative models, and separate facts from interpretation and hypotheses. These are new challenges for knowledge capture and authoring research. Finally, scientific knowledge often needs to be created collaboratively, allowing different contributors to weigh in based on their diverse expertise and perspectives.

Automated Extraction of Scientific Knowledge. Not all scientific knowledge needs to be authored manually. Much of the data known to geoscientists is stored in semi-structured formats, such as spreadsheets or text, and is inaccessible to structured search mechanisms. Automated techniques are needed to make use of this vast store of existing knowledge by identifying and importing the data into structured knowledge bases. This will involve the development of geoscience-aware semantic and natural language analysis approaches. Another important research thread is the application of techniques from machine learning towards the problem of generating metadata from both large and small datasets. With these types of tools, scientists will be able to focus more time and attention on discovery from data, and less on data discovery.

5.1.2 Research Vision: Knowledge Maps

We envision rich knowledge graphs that contain explicit interconnected representations of scientific knowledge linked to physical time and space. These would form *knowledge maps* in five dimensions (3D + time + knowledge annotations). Interpretations and assumptions will be well documented and linked to observational data and models. Today's semantic networks and knowledge graphs link together distributed facts on the Web (eg, Wikidata[11]), but they contain simple facts that lack the depth and grounding needed for scientific research. Knowledge maps will have deeper representations about spatio-temporal processes and will be grounded in the physical world, interconnecting the myriad models of geoscience systems.

5.2 Robotics and Sensing

Data collection is a ubiquitous task across the geosciences. Through intelligent sensing and knowledge-informed data collection, sensing and robotics research has great potential to impact the geosciences.

5.2.1 Research Directions

Optimizing Data Collection. Geoscience data is needed across many scales, both spatial and temporal. Since it is not possible to monitor every measurement at all scales all of the time, there is a crucial need for intelligent methods of sensing. New research is needed to estimate the cost of data collection prior to sensor deployment, whether that means storage size, energy expenditure, or monetary cost. A related research challenge is the tradeoff analysis between the cost of data collection versus the utility of the data to be collected. Since the cost of collecting as much data as possible is impractically high, optimization theory and statistical experimental design approaches are particularly relevant to facilitate obtaining the most amount of information using the least amount of data at the minimum cost without compromising the needs for appropriate bounds on uncertainty.

Unanticipated Uses of Collected Data. A potential challenge for adaptive sensing informed by a specific physical process is in the reuse of data in applications dissimilar to that of original intent during data collection. Geoscience systems involve highly heterogeneous processes and data, which each play a role in modeling phenomena. A major research challenge is to collect data in a way that facilitates its use for unanticipated purposes. Additionally, the need to explicitly express the limits of collected data is crucial to the integrity of the data set.

Active Sampling. Geoscience knowledge can be exploited to inform autonomous sensing systems to not only enable long-term data collection, but to also increase the effectiveness of

[11] http://www.wikidata.org

sensing through adaptive sampling, resulting in richer data sets at lower costs. In the oceanographic community, hybrid autonomous underwater vehicle-gliders have the potential to extend vehicle endurance by combining active thrust and buoyancy. In an adaptive sensing scheme, autonomous vehicles with an embedded decision architecture assimilate data to generate and continuously update an environmental model that is guided by geoscience knowledge to provide prior predictions and estimations to the sensing system. Interpreting sensor data onboard allows vehicles to make decisions guided by real-time variations in data, or to react to unexpected deviations from the current physical model. For example, an AUV could assess the dynamics of observations to track patterns in the environment such as plumes of chemicals or oil spills. Active sampling methods require real-time validation, verification, and calibration of incoming data through analysis and reanalysis of observations to assess alignment with expected physical models, in order to make decisions that inform or adapt the sampling heuristic of an active sampling platform. This leads to models that are not only physically-derived or data-driven, but a combination of both.

Crowdsourcing Data Collection. Another means of gathering large volumes of data required by the geosciences is through crowdsourcing. Citizen scientists can contribute useful data (e.g., collected through geolocated mobile devices) that would otherwise be very costly to acquire. One challenge in data collection through crowdsourcing is in ensuring high quality of data required by geoscience research. A potential opportunity for intelligent systems in the geosciences is to develop improved methods of evaluating crowdsourced data collection empirically, and to gain an understanding of the biases involved in the collection process.

Virtual Sensing. One form of virtual data collection may be through real-time navigation through a virtual model of the area to be observed, to allow measurements to be taken remotely. Already existing repositories could be better leveraged through virtual reality and augmented reality user interfaces to enable "virtual data collection" by navigating and selecting data of interest. One mode of virtual reality data collection would be a visualization of available datasets, exploiting a highly interactive virtual reality platform to sort through available data. User queries and filtering mechanisms would need to be better integrated with navigation and visualizations.

5.2.2 Research Vision: Model-Driven Sensing

New research on sensors will create a new generation of devices that will contain more knowledge of the scientific context for the data being collected, they will use that knowledge to optimize their performance and improve their effectiveness in modeling the phenomena being studied. This will result new *model-driven sensors* that will have more autonomy and exploratory capabilities.

5.3 Information Integration

Data, models, information, and knowledge are scattered across different communities and disciplines, causing great limitations to current geosciences research. Their integration presents major research challenges that will require the use of scientific knowledge for information integration.

5.3.1 Research Directions

Integrating Distributed Repositories of Scientific Knowledge. The geosciences have phenomenal data integration challenges. One aspect of this is the inherently interdisciplinary nature of the fields. Most of the hard geoscience problems require that scientists work across sub-disciplinary boundaries, especially considering the increasing specialization of experts. Another facet of this issue is the sheer volume of data needed to accurately model phenomena in this domain. For instance, modern climate models leverage data collected on 1° grids. This results in a petabyte of data. It is no longer practical for each researcher to have her own local copy of all the data she needs to analyze. It is clear that developing tools and techniques for integrating, maintaining, and searching distributed repositories is critical for future progress. This will present enormous challenges. Geoscience data spans a wide variety of modalities and has greatly varying temporal and spatial scales. Research into how to appropriately represent and merge this data has already started. The cyberinfrastructure community is increasingly moving towards a distributed, decentralized, interconnected model for moving forward. Distributed data discovery tools, shared metadata records, metadata translators, and more appropriate and descriptive standards are emerging in this context. Open issues include: 1) Representing data using modeling concepts that are familiar and useful to domain experts and accurately translate across different domain experts, 2) Entity resolution and scientifically valid data linking, 3) Reward structures to encourage researchers to deposit their data with full and rich metadata and documentation, 4) Development of intelligent user interfaces to facilitate creation, search, and curation of geoscience data using the sematic web and other frameworks.

Threading Scientific Information and Resources. Scientific information and digital resources (data, software, workflows, papers, etc) should be interconnected and interrelated, enabling a rich threaded environment for science to occur. Research challenges include developing new knowledge networks that accurately and usefully link together people, data, models, and workflows. This research will be able to extend understanding of Earth science information interoperability and composition, and deepen our understanding of how collaborative expertise and shared conceptual models develop.

Knowledge-Rich Context-Aware Recommender Systems. Scientists would benefit from proactive systems that understand the task at hand and make recommendations for potential next steps, selection of datasets and analytical methods, and intelligent design of perceptually

effective visualizations. A major research challenge is to design recommender systems that appropriately take into account the complex science context of the geoscientist's investigation. Another research challenge for recommender systems is to carry out dynamic analyses of the interrelationships between the contexts of artifacts and scientists, and anticipate what elements will be relevant to the scientist before they even think of asking for them. Context for an artifact encompasses the environments within which the artifact can execute; context for a scientist describes the situated development environment. For example, a module that requires certain types of data for the best results can be suggested to a geoscientist when they are depositing that kind of data into a repository.

Scientific Discovery Processes and Provenance. Capturing complex data analysis processes as workflows facilitates reuse, scalable execution, and reproducibility. When these workflows are augmented with expert-grade rules to select and customize analyses for any given dataset, automation and validation become possible. The pace of research could be significantly accelerated by these intelligent workflow systems running on data repositories, and automating routine aspects of data analysis. Another research challenge is the analysis and comparison of large volumes of data from different experiments through automated workflows would facilitate progress on defining the bounds of reproducibility caused by the natural variability of boundary conditions that is frequently evident in geoscience fields. A major area of research is collaborative science, particularly supporting real-time co-design of data analysis processes, the ability to track how a workflow evolves over time based on changing designs contributed by multiple researchers, and the capability to capture and retrieve collaboration knowledge on workflow design, such as the discussions that lead to a particular design.

Provenance and trust. Incoming data to the integration process has to be analyzed for its fit and trustworthiness. The original sources must be documented, as well as the integration processes, in order for the information to be understood and trusted. The challenges are in developing appropriate models, and automating provenance/metadata generation throughout the integration process. Although there are standard models to represent provenance, they need to be augmented with geoscience-specific methodologies and scientific discovery processes. These representations need to support guided and natural interaction with scientists, and at the same time enable automated capture of geoscience-relevant provenance records.

Integrating Data From the Published Literature. Published literature is seeing a renaissance as a source of observations particularly as machine learning, text mining and natural language processing tools improve to where they can reliably extract scientific evidence from scanned articles. This information is contained not only in text, but also in tables of data; images of objects; and graphics. Important research challenges in this area include improving the quality of information extraction systems, minimizing the effort required to set up and train these systems, and making them scalable through the vast amounts of the published record. Another area of research is geo-referencing extracted facts, mapping entities that appear with different

labels (e.g., “United States” vs “USA”), and the integration of the information extracted with existing datasets.

5.3.2 Research Vision: Trusted Science Threads

The culmination of the proposed bi-directional, collaborative research program could result in a scientifically accurate, useful, and trusted landscape of data, models, information, and knowledge. The byproduct of scientific discovery includes integrated broad-scale data products derived from raw measurements. These products are described to explain the derivations and assumptions to increase understanding and trust of other scientists. These *trusted science threads* will be easily navigated, queried, and visualized.

5.4 Machine Learning

New machine learning approaches that incorporate scientific knowledge will be needed in order to address the challenges of analyzing sparse geosciences data and the complexity of the phenomena under study. In contrast to machine learning that is rooted in data-science the need here is for learning in the context of dynamic, adaptive integration of data and models, so that inferences can be obtained better than from either source alone. Thus, for example, in contrast to solving difficult extreme-value problems purely statistically, one can embody physics within the sampling process, and in contrast to using a purely numerical approach to forecasting one can integrate data-driven models within this process. Such integrated thinking is not just within the realm of machine learning per se but applies more generally to the development of computational intelligence for geosciences.

5.4.1 Research Directions

Incorporation of Geoscience Knowledge into Machine Learning Algorithms. Geoscience processes are very complex and high dimensional, and the sample size of the data is typically small given the space of possible observations. For those reasons, current machine learning methods are not very effective for many geoscience problems. A promising approach is to supplement the data with knowledge of the dominant geoscience processes. Examples from current work include the use of graphical models, the incorporation of priors, and the application of regularizers. Novel research is needed to develop new machine learning approaches that incorporate knowledge about geoscience processes and use it effectively to supplement the small sample size of the data. Prior knowledge reduces model complexity and makes it possible to learn from smaller amounts of data. Incorporating geoscience process knowledge can also address the high dimensionality that is typical of geoscience data. Prior knowledge constrains the possible relationships among the variables, reducing the complexity of the learning task.

Combining Machine Learning and Simulation Approaches. Machine learning offers data-driven methods to derive models from observational data. In contrast, geoscientists often use simulation models that are built. Process-based simulation approaches impose conservation

principals such as conservations of mass, energy, and momentum. Each approach has different advantages. Data-driven models are generally easier to develop. Process-based simulation models arguably provide reasonable prediction results for situations not represented in the model calibration period, while data-driven models are thought to be unable to extrapolate as well. Yet difficulties in the development of process-based simulation models, such as parameterization and the paucity of clear test results, can draw this claim into question. Intelligent Systems hold the promise of producing the evaluations needed to make the complex approaches used in data-driven and process-model simulation approaches more transparent and refutable. Such efforts will help to use these methods more effectively and efficiently. Novel approaches are needed that combine the advantages of machine learning and simulation models.

Modeling of Extreme Values. There are important problems in geosciences that are concerned with extreme events, such as understanding extremely high temperature or extremely low precipitation. However, existing simulation models are very sensitive to extreme values and therefore the results are not reliable. The heavy-tail property of the extreme value poses important challenges to machine learning algorithms. A major challenge is presented by the spatial-temporal nature of the data.

Evaluation Methodologies. Machine learning evaluation methodology relies heavily on gold standards and benchmark dataset with ground-truth labels. In geosciences, there are no gold standard datasets for many problems. It is unclear how to demonstrate the value of machine learning models. One possible approach is doing reanalysis, which involves making predictions, then collecting observations, and then adjusting the models. Holding data mining competitions using such data would be a very effective attractor for the machine learning community. A possible source of data could be the five standard datasets for climate reanalysis (era40, ncep, etc). We also encourage the creation of training datasets from simulations. Training datasets could be generated that would mimic real data but also have ground truth available, providing opportunity to rigorously train, test and evaluate machine learning algorithms.

Causal Discovery and Inference for Large-Scale Applications. Many geo-science problems involve fundamental questions around causal inference. For example, what are the causes of more frequent occurrences of heat waves? What could be the causes for the change of ocean salinity? There has been a series of recent breakthroughs in causal analysis by machine learning researchers, using methods such as generalization analysis of causal inference, causal inference in presence of hidden components, domain adaption and subsample data, Granger graphical models and causal discovery with probabilistic graphical models. These advances have lead to high efficiency algorithms that can handle large numbers of variables, deal with hidden common causes and incorporate prior knowledge (e.g. expert knowledge). While it is not possible to *prove* causal connections, it is possible to generate new (likely) *hypotheses* for causal connections that can be tested by a domain expert. These breakthroughs have lead to great advances in bioinformatics in recent years, but they are only just emerging in the geosciences.

Given the large amount of data available, we are in a unique position to use these advances to answer fundamental questions around causal inference in the geosciences.

Novel Applications of Advanced Machine Learning Methods to Geosciences. A wide range of advanced machine learning methods could be effectively applied to geoscience problems. This could not only produce novel results in geosciences, but also could result in new challenges for machine learning. Some examples include: 1) change detection algorithms could be applied to urban growth and landscape evolution problems, 2) ensemble methods could reduce climate model errors, and 3) pattern mining to monitor ocean eddies. Machine learning methods have already shown great potential in a specific geoscience application, but significant research challenges remain in order for those methods to be widely - and easily - applicable for other areas of geoscience.

Active Learning, Adaptive Sampling, and Adaptive Observations. Many geoscience applications involve learning highly-complex nonlinear models from data, which usually requires large amounts of labeled data. However, in most cases, obtaining labels can be extremely costly and demand significant effort from domain experts, costly experiments, or long time periods. Therefore, a significant research challenge is to effectively utilize a limited labeling effort for better prediction models. In machine learning, this area of research is known as active learning. Many relevant active sampling algorithms, such as clustering-based active learning, have been developed. New challenges emerge when existing active learning algorithms are applied in geosciences, due to issues such as high dimensionality, extreme events and missing data. In addition, in some cases, we may have abundant labeled data for some sites while being interested in building models for other locations (e.g., remote areas). Transfer active learning aims to solve the problem. It develops new active learning algorithms that can significant reduce the number of labeling requests and build an effective model by transferring the knowledge from areas with large amount of labeled data. The research on transfer active learning is still at early stage and many opportunities exist for novel machine learning research, and in particular to address geoscience challenges.

Interpretive models. In the past few decades, we have witnessed many successes of powerful but complex machine learning algorithms, exemplified by the recent peak of deep learning models. They are usually treated as a black box in practical applications, but have been accepted by more and more communities given the rise of big data and their modeling power. However, in applications such as geosciences, we are interested in both predictive modeling and scientific understanding which requires explanatory and interpretive modeling. A significant research area for machine learning is the incorporation of domain knowledge and causal inference to enable the design of interpretive machine learning approaches. This gives rise to new machine learning frontiers, such as (1) providing fundamental insights into complex approaches, for example to understand how and why deep learning works; and (2) designing surrogate models composed by simple interpretive algorithms to approximate the behaviors of complex uninterpretable models.

5.4.2 Research Vision: Theory-Guided Learning

Geosciences data presents new challenges to machine learning approaches due to the small sample sizes relative to the complexity and non-linearity of the phenomena under study, the lack of ground truth, and the high degree of noise and uncertainty. New machine learning algorithms will have to be developed to address these challenges by incorporating scientific knowledge. These new algorithms will result in a new research agenda of *theory-guided learning*, where knowledge about underlying geosciences processes will guide the machine learning algorithms in understanding complex phenomena.

5.5 Intelligent User Interaction

Scientific research requires well integrated user interfaces where data can easily flow from one to another, and that include and exploit the user's context to guide the interaction. New forms of interaction, including virtual reality and haptic interfaces, should be explored to facilitate understanding and synthesis.

5.5.1 Research Directions

Embedding Visualizations Throughout the Science Process. Visualizations can be graphical, cartographic, temporal, static, dynamic, and interactive. All types of visualization are useful in geoscience studies, yet visualization remains severely underutilized by the Earth science community. Pervasive use of visualizations and direct manipulation interfaces would allow scientists to experience data and models from completely new perspectives. These visualization-based interactive systems require research on the design and validation of novel interactive visual representations that effectively integrate the diverse forms of data associated with Geosciences. These include physically oriented data such as spatio-temporal data, multi-spectral data, and multi-scale data; and abstract forms of information including models, analytical results, hypotheses, provenance, uncertainty, and annotations.

Intelligent Design of Rich Interactive Visualizations. In order to be more ubiquitous throughout the research process, visualizations must be automatically generated and be interactive. One research challenge is to design visualizations that integrate diverse data in 2D, 3D, multi-dimensional, multi-scale, and multi-spectral views. Another challenge is the design of visualizations that fit the scientist's problem. An important area of future research is the interactive visualizations and direct manipulation interfaces would enable scientists to explore data and gain a better understanding of the underlying phenomena.

Immersive Visualizations and Virtual Reality. There are new opportunities for low-cost usable immersive visualizations and physical interaction techniques that virtually put geoscientists into the physical space under investigation, while also providing access to other related forms of data. This research agenda requires bridging prior distinctions in scientific visualization, information visualization, and immersive virtual environments.

Interactive Model Building and Refinement through Visualizations that Combine Models and Data. Interactive environments for model building and refinement would enable scientists to gain improved understanding on how models are affected by changes in initial data and assumptions, how model changes affect results, and how data availability affects model calibration. Developing such interactive modeling environments require visualizations that integrate data with models, ensembles of models, model parameters, model results, and hypothesis specifications. These integrated environments would be particularly useful for developing machine learning approaches to geosciences problems, for example in assisting with parameter tuning and selecting training data. A major challenge is the heterogeneity of these different kinds of information that needs to be represented. The complexity of the models and the model refinement process will also present challenges to the design of interfaces that can guide users through the process. A novel area of research would be interactive systems to support bidirectional workflows that would enable reverse reasoning, such as fixing algorithm outcomes in causality or sensitivity analyses to identify acceptable model parameters.

Interfaces for Spatio-Temporal Information. The vast majority of geosciences research is geospatially localized and with temporal references. Geospatial information requires specialized interfaces and data management approaches. New research is needed in intelligent interfaces for spatio-temporal information that exploit the user's context and goals to identify implicit location, to disambiguate textual location specification, or to decide what subset of information to present. The small form factor of mobile devices is also constraint in developing applications that involve spatial data.

Collaboration and Assistance for Data Analysis and Scientific Discovery Processes. Intelligent workflow systems could help scientists by automating routine aspects of their work. Because each scientist has a unique workflow of activities, and because their workflow changes over time, a research challenge is that these systems need to be highly flexible and customizable. Another research challenge is to support a range of workflows and processes, from common ones that can be reused to those that are highly exploratory in nature. Such workflows systems must enable collaborative design and analysis and be able to coordinate the work of teams of scientists. Finally, workflow systems must also support emerging science processes, including crowdsourcing for problems such as data collection and labeling.

5.5.2 Research Vision: Interactive Analytics

New research is required to allow scientists to interact with all forms of knowledge relevant to the phenomenon at hand, to understand uncertainties and assumptions, and to provide many alternative views of integrated information. This will result in a new generation of user interfaces focused on *interactive analytics*, where visualizations and manipulations will be embedded throughout the analytic process. These new intelligent user interfaces and interaction modalities should support the exploration not only of data but of the relevant models and knowledge that

provide context to the data. Research activities should flow seamlessly from one user interface to another, each appropriate to the task at hand and rich in user context.

6. General Findings and Recommendations

Intelligent systems have demonstrated significant transformative impact in the commercial sector. After decades of investment, a variety of artificial intelligence techniques have matured and given rise to product recommenders, ad placement systems, self-driving cars, speech-based interfaces, and web searches. Billions use these services in their daily lives to great benefit.

However, **these approaches are inadequate to meet the challenges presented by geosciences research**. First, using data alone is insufficient to create models of the complex phenomena under study. Second, geoscientists need to reach across disciplines to synthesize disparate data and models, which requires extensive qualification and context. Third, scientists need powerful partnerships with computers in order to explore complex hypotheses and understand how new findings relate to the existing body of knowledge.

A new generation of geoscience-aware intelligent systems must emerge with a much deeper understanding of the physical laws that provide context and structure to the data. Major investments in this area have the potential to have transformative impact in artificial intelligence research, and at the same time have transformative impact in geosciences as well as in other science disciplines and beyond into commercial world. These capabilities would support the integration of models from geosciences with other sciences to address the interactions among food, energy, and water resources. In engineering, more complex designs (Internet of Things, smart grid, smart cities) would be enabled by these capabilities to harness modeling, diagnosis, and prognosis to reduce costs and improve resilience.

The essence of these advances requires that intelligent systems and geosciences researchers work together to formulate knowledge-rich frameworks, algorithms, and user interfaces. Geoscientists need to take an active role in articulating the nature of their problems, and in working together with intelligent systems researchers to incorporate new approaches into their work and inform subsequent research directions. Unfortunately, these interactions are not likely to occur without significant facilitation.

This section presents four major general findings and recommendations to facilitate joint research between intelligent systems and geosciences. It also includes specific suggestions for short-term follow-on activities.

6.1 Transformative Effect of Intelligent Systems and Geosciences Collaborations

- **Findings**: There is an existing foundation of significant research contributions from intelligent systems and geosciences collaborations. Some contributions lie at the intersection of those disciplines, such as innovative robotic devices (intelligent systems)

to explore the ocean floor (geosciences), novel approaches to learn regularities (intelligent systems) in climate data (geosciences), and new frameworks for data integration (intelligent systems) of geospatial datasets (geosciences). Through those interactions, new fundamental contributions are enabled in each discipline. Motivated by challenging geoscience problems and datasets, intelligent systems researchers can investigate new fundamental techniques to tackle new kinds of complex practical problems. Equipped with an innovative IS technique, geoscientists can apply it to other aspects of their work and make transformative advances in their science. These new techniques and advances have significant transformative effect since they tackle problems that could not be solved without a cross-disciplinary collaboration and disseminate concepts and techniques across fields. They also have broader impact, since they can change how a research community approaches problems.

- **Recommendations**: Increasing intelligent systems and geosciences collaborations would expand the research contributions and broaden their scope, with significant transformative impact in the research agendas of both areas. Although initial grants could be obtained through the NSF EAGER programs, but those are limited short-term investments that would simply establish initial collaborations. What is needed is multi-year funding programs that are formulated with multi-disciplinary research in mind.

6.2 Sustaining and Broadening Intelligent Systems and Geosciences Interactions

- **Findings**: The interactions between intelligent systems and geosciences are very beneficial but still limited. Although the workshop participants were in agreement about the benefits of such interactions, they also recognized that the opportunities for intelligent systems and geosciences discussions are not many. The NSF EarthCube program and the CISE Expeditions programs have facilitated some important connections in specific areas. There are some workshops focused on intelligent systems for climate and environmental science at AAAI and IJCAI among others, but the interactions have been limited to the machine learning community. Other areas of geosciences have not had similar events. On the geosciences side the AGU and other meetings hold sessions on informatics, and although there are some sessions in semantics and metadata they tend to focus on data management and computation issues rather than novel research opportunities. The IS and GEO communities do not have many chances to interact and explore collaborations at present. This limits the potential for new scientific contributions at the intersection of the two areas and back to the more fundamental research in each discipline.
- **Recommendations**: Fostering more comprehensive and deeper intelligent systems and geosciences interactions will increase and augment the range of research topics pursued at the intersection between the disciplines. The existence of funding programs as suggested above will attract researchers from both communities to participate in such

events. Additionally, specific sessions on intelligent systems and geosciences should be held at conferences and other events, and special working meetings and retreats on target topics of interest should be planned. Examples include:

- Creating Special Interest Groups or Research Coordination Networks as part of EarthCube. Interest groups are the basic structures for the conversations about EarthCube that take place on the EarthCube web platform. Research Coordination Networks are funded projects to support meetings and community activities.
- Establishing working groups under the auspices of the USGS Powell Center. This would fund community activities and workshops to start to address some of the suggested directions in this report.

6.3 Growing an Intelligent Systems and Geosciences Research Community

- **Findings**: Researchers shy away for many reasons from cross-disciplinary research areas, and synergistic research in intelligent systems and geosciences falls in this category. There is a learning curve to embark on such collaborations, which can hamper the growth of this nascent community. In addition, researchers need to be in a position to make those investments and reap the results. A key challenge for researchers interested in this area is sustaining multi-year collaborations that enable the initial learning stages and support the gradual understanding and generation of ideas and ultimately their realization and application. Moreover, once a collaboration is fruitful there is usually a ripple effect of ideas and potential follow on projects that would be productive from the beginning. These long-term collaborations are often hard to support and maintain.
- **Recommendations**: Sustained multi-year funding programs would attract a substantial amount of scientists to pursue synergistic intelligent systems and geosciences research. In addition, reducing the growth of the community can be encouraged by reducing this learning curve when possible. Developing community research resources would be very effective, including:
    - *A community-led synthesis of generalizable problems, priorities, and expected impact.* An important investment of effort is required for an investigator to identify a concrete problem, abstract the aspects that make it applicable to other situations, and understand the impact that solving that problem may have in other fields. Community activities to identify, abstract, and characterize the impact of key challenges at the intersection of intelligent systems and geosciences would help attract researchers to tackle them. These could be accomplished by standing committees with periodic visioning retreats and resulting public reports.
    - *A community-led repository of datasets and challenge problems*. This would follow the example of the UC Irvine Machine Learning Repository [Lichman 2013], which has collected benchmark datasets from other disciplines that are

widely used by the machine learning community. A repository for climate related data sets has also been started for climate [CI 2015], but it is still in its infancy. In addition to data, organizing community challenge problems and events help attract interest and participation. Good examples are the machine learning challenges organized by ChaLearn, a tax-exempt organization founded for the purpose. Robotics is another area where shared challenges have driven the research focus and shown significant improvements over time, such as the RoboCup robot soccer tournament steered by a research symposium and a shared platform [RoboCup 2015]. Other areas of research would benefit from these shared datasets and challenges, including knowledge representation, information integration, and intelligent interfaces and visualization.

6.4 Facilitating IS-GEO Communication and Education

- **Findings**: Understanding the problems and existing approaches is the basis for new ideas and collaborations, but this is very challenging due to the diversity of terminologies and conceptualizations in each discipline.
- **Recommendations**: Organizing joint intelligent systems and geosciences efforts to develop curated collections of materials for learning across disciplines, from simple glossaries to educational modules to class syllabi and curricula, which could be interlinked and cross-referenced with the literature. These collections would have oversight from a rotating editorial board composed of both intelligent systems and geosciences researchers. The process of identifying and reviewing the materials in the collections is likely to lead to new collaborations among contributors, providing incentives for researchers to participate in addition to serving the community and increasing their name recognition. These collections could include:
    - *A cross-indexed glossary* that defines the key terms and concepts used in both the intelligent systems and geosciences communities. Researchers on both areas would need to ensure that the descriptions are accessible to the other side.
    - *A self-study repository* for intelligent systems and geosciences researchers to quickly learn about relevant topics in the other discipline. This would consist of high-quality, self-contained, accessible, and well-organized modules for independent learning. Materials need not be developed from scratch, but would build on the plethora of existing tutorials, papers, and recordings that are already available. Possible vehicles for this work could be the NSF Research Traineeship Program (NRT) and ESRI Educational grants.
    - *An educational catalog* of existing classes, programs, and curricula. Some universities have geoinformatics programs, but they tend to focus on data management and computational infrastructure rather than intelligent systems

or geosciences research topics. Other universities have bioinformatics programs that could be used as a basis to jumpstart geoinformatics programs. Most universities do not yet offer even a single class in this important interdisciplinary area. This educational catalog would include information on existing courses and programs, teaching materials, and instructor interviews to share best practices. This effort could build on the past work by the Interdisciplinary Teaching about Earth for a Sustainable Future (InTeGrate) project [InTeGrate 2015]. Funding support could be as part of the NSF Science of Learning Centers (SLC) and the NSF Idea Labs.

6.5 Short-Term Follow-Up on Recommendations

Workshop participants identified a number of specific opportunities to follow up on the above recommendations:

- **Engagement with EarthCube**. The participants propose to take actions to engage with EarthCube, initially through a Special Interest Group. Interest groups are the basic structures for the conversations about EarthCube that take place on the EarthCube Web platform. Such a setting allows virtual communities of practice within EarthCube to form around common interests. EarthCube also funds coordination meetings and other community activities such as Research Coordination Networks (RCN).
- **Engagement with USGS Powell Center**. Launched by USGS in 2009, The Powell center of the US Geological Survey in Fort Collins, CO aims to develop an analysis and synthesis-centered strategy outlining the major natural-science issues facing the Nation in the next decade. MOU was signed between USGS and NSF Geosciences Directorate in 2012. The Powell center provides an excellent framework to jumpstart and establish new working groups, by providing facilities and travel support for groups to meet face-to-face for a week at a time [USGS 2015]. While it is too late to propose such a working group this year (deadline in late April), the participants strongly suggest the community members to take advantage of this opportunity in 2016. Marcia McNiff (USGS) gave a presentation at the workshop that is available in the workshop web site.
- **Engagement with NSF Research Traineeship Program (NRT)**. The NRT program encourages the development of bold, new, potentially transformative, and scalable models for STEM graduate training. NRT grants could focus on developing graduate training programs to ensure that geoscience graduate students in research-based master's and doctoral degree programs develop the skills, knowledge, and competencies needed to pursue a range of STEM careers.
- **Engagement with NSF Science of Learning Centers (SLC)**. The SLC program encourages the development of large-scale, long-term centers for the long-term advancement of Science of Learning research. SLC grants could supports cross-

disciplinary intelligent systems and geosciences research to build a common groundwork of intelligent conceptualization, experimentation and explanation towards a deeper understanding of learning of geoscience.

- **Engagement with NSF Idea Labs**. An "Ideas Lab" is a new merit review strategy coined by NSF to address grand challenges in STEM research and education. NSF has launched a new program for "Improving Undergraduate STEM Education" (IUSE) through its Division of Undergraduate Education (EHR/DUE). Its "IUSE Phase I Ideas Labs" solicits proposals focusing on discipline-specific workforce development needs, where geoscience is one of the three identified disciplines. IUSE Phase I Ideas Labs grants could support the work on geoscience training initiatives addressed in this report.
- **Engagement with Other Research Initiatives.** Many of the initial synergistic research and activities reported here came to life under the NSF ITR program and developed further into other programs such as the NSF and AFOSR DDDAS programs, and NSF EXPEDITIONS and INSPIRE programs. These kinds of programs must be pursued by this nascent community as mechanisms for funded collaborations in the near term.

Workshop participants also suggested organizing specific events at scientific meetings. Specific activities are already ongoing at the International Conference on Computational Science (DDDAS workshop), Dynamic Data-Driven Environmental Systems Science Conference, the DDDAS session with American Controls Conference, and the Climate Informatics Workshop series. Additional opportunities are presented by specific meetings in geosciences including the American Geophysical Union (AGU) Fall Meeting (held annually in December), the Geological Society of America's Annual Meeting (held annually in November), the American Physical Society Focus Group on Climate (yearly meeting in March), the SIAM Geosciences (yearly meeting in June), and the International Environmental Modeling and Software Systems (iEMSs) conference (biannual event held in June).

In the intelligent systems area, conferences such as the International Joint Conference in AI (IJCAI) and the annual Conference of the Association for the Advancement of Artificial Intelligence (AAAI) have held sustainability tracks that could be expanded to other areas of geosciences.

Other target conferences include the International Conference on Machine Learning (ICML), the ACM Conference on Knowledge Capture (K-CAP), the ACM Intelligent User Interfaces (IUI) conference, the IEEE Visual Analytics Science and Technology (VAST), the IEEE Information Visualization (InfoVis), and the IEEE Scientific Visualization (SciVis). There is a void of high quality conferences in geoinformatics in general, which presents an opportunity for synergies in this space centered on intelligent systems for geosciences.

7. Conclusions

This workshop investigated how new research in intelligent systems could enable groundbreaking geosciences research. In recent years, intelligent systems have demonstrated significant transformative impact in the commercial sector. These techniques have been applied to geosciences with some success, but they are inadequate to meet the challenges presented by geosciences research. First, using data alone is insufficient to create models of the complex phenomena under study. Second, geoscientists need to reach across disciplines to synthesize disparate data and models, which requires extensive qualification and context. Third, scientists need powerful partnerships with computers in order to explore complex hypotheses and understand how new findings relate to the existing body of knowledge. Therefore, in order to tackle complex geosciences phenomena new approaches are needed.

A new generation of geoscience-aware intelligent systems needs to emerge with a much deeper understanding of the physical laws that provide context and structure to the data. This report presents research opportunities in information and intelligent systems inspired by geosciences challenges. Crucial capabilities are needed that require major research in knowledge representation, robust sensors, information integration, machine learning, and interactive analytics.

Enabling these advances requires that intelligent systems and geosciences researchers work together to formulate knowledge-rich frameworks, algorithms, and user interfaces. Recognizing that these interactions are not likely to occur without significant facilitation, workshop participants made four major recommendations to facilitate this collaborative research. First, long-term sustained funding programs are needed to foster collaborations across these disciplines. Second, research coordination networks and annual events would enable sustained communication across these fields that do not typically cross paths. Third, repositories of challenge problems and datasets with crisp challenge statements would lower the barriers to getting involved. Fourth, a curated repository of learning materials to educate researchers and students alike is needed to reduce the steep learning curve involved in understanding advanced topics in the other discipline.

Major investments in this area have the potential to have transformative impact in artificial intelligence research, and at the same time have transformative impact in geosciences as well as in other science disciplines and beyond into commercial world.

Acknowledgments

This work was sponsored by the Directorate for Computer and Information Science and Engineering and the Directorate for Geosciences of the US National Science Foundation under grant number IIS-1533930. The authors would like to thank CISE and GEO Program Directors for their guidance and suggestions, in particular Hector Muñoz-Avila and Eva Zanzerkia for their

guidance, and Todd Leen, Frank Olken, Sylvia Spengler, Amy Walton, and Maria Zemankova for suggestions and feedback.

References

[ChaLearn 2015] Challenges in Machine Learning. http://www.chalearn.org/. Last accessed July 21, 2015.

[CI 2015] Climate Informatics. http://www.climateinformatics.org/?q=data-sets/. Last accessed July 21, 2015.

[Che 2015] Zhengping Che, David C. Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. *Deep Computational Phenotyping*. Proceedings of the 21st ACM Conference on Knowledge Discovery and Data Mining (SIGKDD), 2015.

[Gil et al 2014] Gil, Y.; Chan, M.; Gomez, B.; and B. Caron (Eds). "EarthCube: Past, Present, and Future." EarthCube Project Report EC-2014-3, 2014.

[InTeGrate 2015] Interdisciplinary Teaching about Earth for a Sustainable Future. http://serc.carleton.edu/integrate/participate/index.html. Last accessed July 21, 2015.

[Kawale et al. 2013] Kawale, J., S. Liess, A. Kumar, M. Steinbach, P. Snyder, V. Kumar, A. R. Ganguly, N. F. Samatova, and F. Semazzi. "A graph-based approach to find teleconnections in climate data." Stat. Anal. Data Mining, 6, 158-179, 2013.

[Lichman 2013] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[Liess et al. 2014] Liess, S., A. Kumar, P. K. Snyder, J. Kawale, K. Steinhaeuser, F. H. M. Semazzi, A. R. Ganguly, N. F. Samatova, and V. Kumar. "Different modes of variability over the Tasman Sea: Implications for regional climate." J. Climate, 27, 8466-8486, 2014.

[NRC 2012] National Research Council, New Research Opportunities in the Earth Sciences at the National Science Foundation, Committee on new Research Opportuniteis in the Earth Sciences; National Research Council, ISBN 978-0-309-21924-2, National Academies Press, Washington, DC, p. 216.

[NRC, 2012a] National Research Council, Challenges and Opportunities in the Hydrologic Sciences, Committee on Challenges and Opportunities in the Hydrologic Sciences, Water Science and Technology Board, Division on Earth and Life Studies, ISBN: 978-0-309-22283-9, National Academies Press, Washington, DC, p. 188.

[NRC 2013] National Research Council, Solar and Space Physics: A Science for a Technological Society, Committee on a Decadal Strategy for Solar and Space Physics (Heliophysics); Space Studies Board; Aeronautics and Space Engineering Board; Division of Earth and Physical Sciences; National Research Council, ISBN 978-0-309-16428-3, National Academies Press, Washington, DC, p. 466.

[NRC 2014a] National Research Council, Review of the National Science Foundation's Division on Atmospheric and Geospace Sciences Draft Goals and Objectives Document, Committee to Review the NSF AGS Draft Science Goals and Objectives, ISBN 978-0-309-31048-2, National Academies Press, Washington, DC, p. 36.

[NRC 2014b] National Research Council, Sea Change: 2015-2025 Decadal Survey of Ocean Sciences, Committee on Guidance for NSF on National Ocean Science Research Priorities: Decadal Survey of Ocean Sciences, Ocean Studies Board; Division on Earth and Life Studies, ISBN 978-0-309-36688-5, National Academies Press, Washington, DC, p. 98.

[NSF 2014] "Dynamic Earth: GEO Imperatives and Frontiers 2015-2020." National Science Foundation, eds. NSF Advisory Committee for Geosciences.

[Peters et al. 2014] Peters SE, Zhang C, Livny M, Ré C (2014) "A Machine Reading System for Assembling Synthetic Paleontological Databases." PLoS ONE 9(12): e113523. doi:10.1371/journal.pone.0113523.

[Pundsack et al. 2013] Pundsack J, Bell R, Broderson, D, Fox GC, Dozier J, Helly J, Li W, Morin P, Parsons M, Roberts A, Tweedie C, Yang C (2013) "Report on Workshop on Cyberinfrastructure for Polar Sciences." St. Paul, Minnesota. University of Minnesota Polar Geospatial Center, 17pp.

[RoboCup 2015] RoboCup. http://www.robocup.org/. Last accessed July 21, 2015.

[USGS 2015] USGS Powell Center Proposals. https://powellcenter.usgs.gov/proposals. Last accessed July 21, 2015.