Communications of the ACM, January 2019, Vol. 62 No. 1, Pages 76-84. DOI: 10.1145/3192335

Intelligent Systems for Geosciences: An Essential Research Agenda

Yolanda Gil, University of Southern California
Suzanne A Pierce, The University of Texas Austin
Hassan Babaie, Georgia State University
Arindam Banerjee, University of Minnesota
Kirk Borne, George Mason University
Gary Bust, Johns Hopkins University
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
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

ABSTRACT

This article outlines a research agenda for intelligent systems that will result in fundamental new capabilities for understanding the Earth system, based on requirements from Earth, ocean, polar, atmospheric, and geospace sciences. In order to meet the challenges presented by complex geosciences phenomena with uncertain, intermittent, sparse, multi-resolution, and multi-scale data, new approaches must be developed to incorporate existing scientific knowledge and the user's context into intelligent systems. This will result in a new generation of knowledge-rich intelligent systems that will require significant research in artificial intelligence: 1) In knowledge representation, capturing scientific knowledge in the form of physical, geological, chemical, biological, and ecological processes will push the limits of the state of the art; 2) In sensing and robotics, that scientific knowledge should be used to prioritize and collect data; 3) In information integration, all those processes need to form a "system of systems" where all data and knowledge are interconnected; 4) In machine learning, algorithms need to be enriched with models of the relevant physical, geological, chemical, biological, and ecological processes; and 5) In user interfaces, interaction modalities must be guided by a knowledge-rich user model that provides context for interactions with scientists. This research agenda in knowledge-rich intelligent systems will be essential to unlock much needed integrated discoveries to chart our planet's history and future.

1 INTRODUCTION

Many aspects of geosciences pose novel problems for intelligent systems research. Geoscience data is challenging 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 requires breakthroughs that would significantly transform intelligent systems, while greatly benefitting the geosciences in turn. Although there have been significant and beneficial interactions between the intelligent systems and geosciences communities [1;2], the potential for synergistic research in intelligent systems for geosciences is largely untapped. A recently launched Research Coordination Network on Intelligent Systems for Geosciences followed a workshop at the National Science Foundation on this topic [3]. This expanding network builds on the momentum of the NSF EarthCube initiative for geosciences, and is driven by practical problems in Earth, ocean, atmospheric, polar, and geospace sciences [4]. Based on discussions and activities within this network, this article presents a research agenda for intelligent systems inspired by geosciences challenges.

Geosciences research aims to understand the Earth as a system of complex highly interactive natural processes and their interactions with human activities. Current approaches have fundamental shortcomings given the complexity of geosciences data. First, using data alone is insufficient to create models of the very complex phenomena under study so prior theories need to be taken into account. Second, data collection can be most effective if steered using knowledge about existing models to focus on data that will make a difference. Third, to combine disparate data and models across disciplines requires capturing and reasoning about extensive qualifications and context to enable their integration. These are all illustrations of the need for knowledge-rich intelligent systems that incorporate significant amounts of geosciences knowledge.

The article begins with an overview of research challenges in geosciences. It then presents a research agenda and vision for intelligent system to address those challenges. It concludes with an overview of ongoing activities in the newly formed research network of intelligent systems for geosciences that is fostering a community to pursue this interdisciplinary research agenda.

2 GEOSCIENCE CHALLENGES REQUIRING INNOVATIONS IN INTELLIGENT SYSTEMS

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-term effects of human activities on the planet [4;5;6;7;8;9]. 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.

KEY INSIGHTS

Advances in artificial intelligence are needed 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 scale.

Intelligent systems need to incorporate extensive knowledge about the physical, geological, chemical, biological, ecological, and anthropomorphic factors that affect the Earth system while leveraging recent advances in data-driven research.

A new generation of knowledge-rich intelligent systems will enable more robust sensor platforms, more effective information integration, more capable machine learning algorithms, and intelligent interactive environments that have the potential to significantly transform geosciences research practices and expand the nature of the problems under study.

Different disciplines in geosciences are facing these challenges from different motivations and perspectives:

- **Forecasting rates of sea level change in polar ice shelves:** Polar scientists, along with atmospheric and ocean scientists, face an urgent need to understand sea level rise around the globe. 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. 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. New research on intelligent sensors would support selective data collection, on-board data analysis, 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.
- **Unlock deep Earth time:** Earth Scientists focus on understanding the dynamics of the Earth, including the interior of the Earth or *deep Earth* (such as tectonics, seismology, magnetic or gravity fields, and volcanic activity) and the near-surface Earth (such as the hydrologic cycle, the carbon cycle, the food production cycle, and the energy cycle). While collecting data from the field is done by individuals in select locations, the problems under

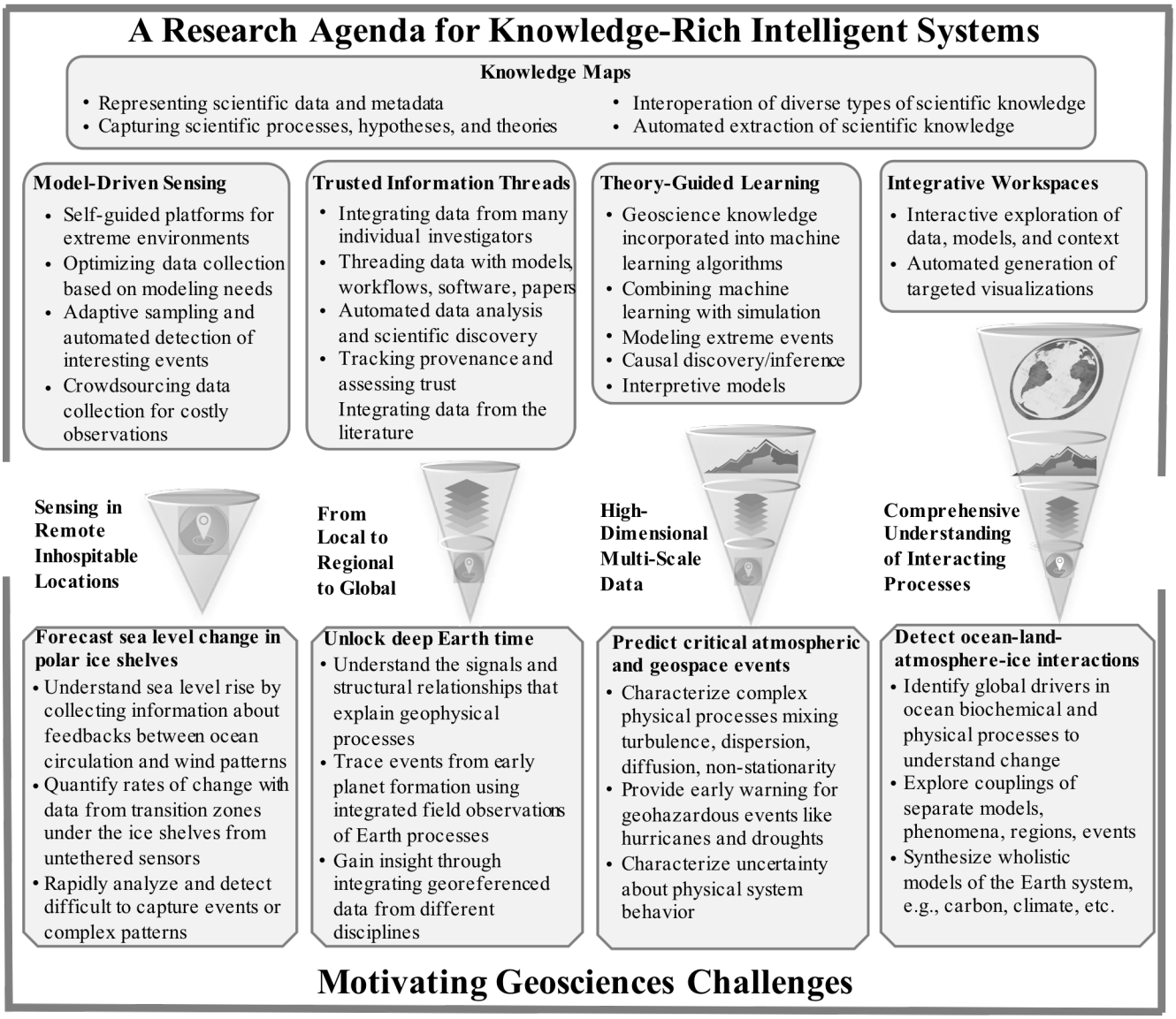

Figure 1. New research in artificial intelligence (top) will result in a new generation of knowledge-rich intelligent systems that could address the significant challenges faced by geosciences (bottom). Knowledge-rich intelligent systems will exploit knowledge maps containing models and pre-existing knowledge in order to drive sensor data collection, create trusted information threads, power theory-guided learning, and enable integrative analytics.

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. Earth systems are integrated, but current geoscience data and models are not. To unravel significant questions about topics, such as Deep Earth Time, geoscientists need intelligent systems to efficiently integrate data from disparate locations, data types, and collection efforts within a wide area.

- **Predict critical atmosphere and geospace events**: 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. Although the data collected is very large, it is miniscule given the complexity of the phenomena under study. Therefore, the data available must be augmented with knowledge about physical laws underlying the phenomena in order to generate effective models.

- **Detect ocean-land-atmosphere-ice interactions**: Our ability to understand the Earth system is heavily dependent on our ability to integrate geoscience models across time, space, and discipline. This requires sophisticated approaches that support composition and discover structure, diagnose and compensate for compound model errors and uncertainties, and generate rich visualizations of multi-dimensional information that take into account a scientist's context.

Figure 1 illustrates intelligent systems research directions inspired by these geoscience challenges, organized at various scales. Studying the Earth as a system requires fundamentally new capabilities 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.

3 A ROADMAP FOR INTELLIGENT SYSTEMS RESEARCH WITH BENEFITS TO GEOSCIENCES

Earth systems phenomena are characterized by non-linear, multi-resolution, multi-scale, heterogeneous, and highly dynamic processes. Geosciences research is also challenged by extreme events and long-term shifts in Earth systems. The data available is intermittent, has significant sources of uncertainty, and is very sparse given the complexity and rich phenomena under study. Therefore, the small sample size of the datasets must be supplemented with the scientific principles underlying geosciences processes in order to guide knowledge discovery. 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. We need approaches that leverage the advances in data-driven research with methods that exploit the domain knowledge and scientific principles that govern the phenomena under study. These geoscience-aware systems will need to incorporate extensive knowledge about phenomena that combine physical, geological, chemical, biological, ecological, and anthropomorphic factors.

This body of research will lead to a new generation of *knowledge-rich intelligent systems* that contain rich knowledge and context in addition to data, enabling fundamentally new forms of reasoning, autonomy, learning, and interaction. The research challenges for creating knowledge-rich intelligent systems center on five major areas:

"Studying the Earth as a system requires fundamentally new capabilities 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."

1. *Knowledge Representation and Capture*: Capturing scientific knowledge about processes, models, and hypotheses.
2. *Sensing and Robotics*: Prioritizing data collection based on the scientific knowledge available.
3. *Information Integration*: Representing data and models as a "system of systems" where all knowledge is interconnected.
4. *Machine Learning*: Enriching algorithms with knowledge and models of the relevant underlying processes.
5. *Interfaces and Interactive Systems*: Exploring and understanding user context using interconnected knowledge.

We describe these five areas in turn. For each area, we introduce major research directions followed by an overarching vision for that area.

3.1 Knowledge Representation and Capture

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

3.1.1 Research Directions

Representing Scientific Data and 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, creating links among datasets to further interoperability [10]. This leads to Web-embedded semantic networks and knowledge graphs that provide vast amounts of open interconnected knowledge about geosciences. Semantics, ontological representations, scientifically accurate concept mappings across domains, knowledge graphs, and the application of Linked Open Data are all areas of active

research to facilitate search and integration of data without a great deal of manual effort [11].

Capturing Scientific Processes, Hypotheses, and Theories. To complement the ontologies and data representations just discussed, a great challenge is representing the ever-evolving, uncertain, complex, and dynamic scientific knowledge and information. Important challenges will arise in representing dynamic processes, uncertainty, theories and models, hypotheses and claims, and many other aspects of a constantly growing scientific knowledge base. 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. In addition, scientists will need to understand the representations and 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. 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 smaller to larger scales, 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. 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 should 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 identify and import these kinds of data into structured knowledge bases.

3.1.2 Research Vision: Knowledge Maps

We envision rich knowledge graphs that will contain explicit interconnected representations of scientific knowledge linked to time and space to form multi-dimensional *knowledge maps*. 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, but they contain simple facts that lack the depth and grounding needed for scientific research. Knowledge maps will have deeper spatio-temporal representations of processes, hypotheses, and theories and will be grounded in the physical world, interconnecting the myriad models of geoscience systems.

3.2 Robotics and Sensing

Knowledge-informed sensing and data collection has great potential to do more cost-effective data gathering across the geosciences.

3.2.1 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 for 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 tradeoff analysis of the cost of data collection versus the utility of the data to be collected.

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 sensing through adaptive sampling, resulting in richer data sets at lower costs. Interpreting sensor data onboard allows autonomous vehicles to make decisions guided by real-time variations in data, or to react to unexpected deviations from the current physical model.

Crowdsourcing Data Collection for Costly Observations. 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 area of research is to improve methods of evaluating crowdsourced data collection empirically, and to gain an understanding of the biases involved in the collection process.

3.2.2 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. These devices will use that knowledge to optimize their performance and improve their effectiveness. This will result in new *model-driven sensors* that will have more autonomy and exploratory capabilities.

3.3 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.

3.3.1 Research Directions

Integrating Data from Distributed Repositories. The geosciences have phenomenal data integration challenges. Most of

the hard geoscience problems require that scientists work across sub-disciplinary boundaries and share very large amounts of data. Another facet of this issue is that the data spans a wide variety of modalities and greatly varying temporal and spatial scales. Distributed data discovery tools, metadata translators, and more descriptive standards are emerging in this context. Open issues include cross-domain concept mapping, entity resolution and scientifically valid data linking, and effective tools for finding, integrating, and reusing data.

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

Automated Data Analysis and Scientific Discovery. Capturing complex integrative data analysis processes as workflows facilitates reuse, scalable execution, and reproducibility. The pace of research could be significantly accelerated with intelligent workflow systems that automatically select data from separate repositories and carry out integrated analyses of data from different experiments. Through workflows that integrate large amounts of diverse data and interdisciplinary models, intelligent systems will lead to new discoveries.

Tracking Provenance and Assessing 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 and scientific discovery processes.

Integrating Data from the Published Literature. Important historical data in geosciences is often only available in the published literature, requiring significant effort to integrate with new data. Text mining and natural language processing tools can already extract scientific evidence from articles [4]. Important research challenges in this area include improving the quality of existing 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 and integrating newly extracted information with existing data repositories.

3.3.2 Research Vision: Trusted Information Threads

The proposed research will result in a scientifically accurate, useful, and trusted knowledge-rich landscape of data, models, and information that will include integrated broad-scale byproducts derived from raw measurements. These products will be described to explain the derivations and assumptions to increase understanding and trust of other scientists. These *trusted information threads* will be easily navigated, queried, and visualized.

3.4 Machine Learning

In order to address the challenges of analyzing sparse geosciences data given the complexity of the phenomena under study, new machine learning approaches that incorporate scientific knowledge will be needed so that inferences will be obtained better than from data alone.

3.4.1 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 [12]. 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 changes in the frequency and spatial distribution of extremely high temperature or extremely low precipitation in response to increase in greenhouse gas emissions. However,

existing climate simulation models are often unable to reproduce realistic extreme values and therefore the results are not reliable. Although data science models offer an alternative approach, the heavy-tail property of the extreme values and its spatio-temporal nature 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 datasets with ground-truth labels. In geosciences there are no gold standard datasets for many problems, and in those cases it is unclear how to demonstrate the value of machine learning models. One possible approach involves making predictions, collecting observations, and then adjusting the models to account for differences between prediction and observations. Holding data mining competitions using such data would be a very effective attractor for the machine learning community. Another alternative could be 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 geoscience 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? While it may be very hard to prove causal connections, it is possible to generate new (likely) hypotheses for causal connections that can be tested by a domain expert 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. 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 Machine Learning Methods Motivated by Geosciences Problems. A wide range of advanced machine learning methods could be effectively applied to geoscience problems. Moreover, geosciences problems drive researchers to develop entirely new machine learning algorithms. For example, attempts to build a machine learning model to predict forest fires in the tropics using multi spectral data from earth observing satellites led to a novel methodology for building predictive models for rare phenomena [3] that can be applied in any setting where it is not possible to get high quality labeled data even for a small set of samples, but poor quality labels (perhaps in the form of heuristics) are available for all samples. Machine learning methods have already shown great potential in a few specific geoscience applications, 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 with 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. Transfer active learning is still in early stages and many opportunities exist for novel machine learning research.

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 that can be understood by scientists and related to existing geosciences theories and models.

3.4.2 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 approaches for *theory-guided learning* will need to be developed, where knowledge about underlying geosciences processes will guide the machine learning algorithms in modeling complex phenomena.

3.5 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.

3.5.1 Research Directions

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, suggest datasets and analytical methods, and generate

perceptually effective visualizations. A major research challenge is to design recommender systems that appropriately take into account the complex science context of a geoscientist's investigation.

Embedding Visualizations Throughout the Science Process. Pervasive use of visualizations and direct manipulation interfaces throughout the science process would need to link data to hypotheses and allow scientists to experience models from completely new perspectives. These visualization-based interactive systems require research on the design and validation of novel visual representations that effectively integrate diverse data in 2D, 3D, multi-dimensional, multi-scale, and multi-spectral views, as well as how to link models to the relevant data used to derive them.

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. Another challenge is the design of visualizations that fit a 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 and complexity of these different kinds of information that needs to be represented.

Interfaces for Spatio-Temporal Information. The vast majority of geosciences research products 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.

3.5.2 Research Vision: Integrative Workspaces

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 user interfaces focused on *integrative workspaces*, where visualizations and manipulations will be embedded throughout the analytic process. These new intelligent user interfaces and interaction modalities will support the exploration not only of data but of the relevant models and knowledge that provide context to the data. Research activities will flow seamlessly from one user interface to another, each appropriate to the task at hand and rich in user context.

3 CONCLUSIONS

This article presented research opportunities in knowledge-rich intelligent systems inspired by geosciences challenges. Crucial capabilities are needed that require major research in knowledge representation, selective sensing, 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, a new Research Coordination Network on Intelligent Systems for Geosciences has been created to enable sustained communication across these fields that do not typically cross paths. This network focuses on three major goals. First, the organization of joint workshops and other forums will foster synergistic discussions and collaborative projects. Second, repositories of challenge problems and datasets with crisp problem statements will lower the barriers to getting involved. Third, a curated repository of learning materials to educate researchers and students alike will reduce the steep learning curve involved in understanding advanced topics in the other discipline. Additionally, members of the Research Coordination Network are engaging other synergistic efforts, programs, and communities, such as artificial intelligence for sustainability, climate informatics, science gateways, and the US NSF Big Data Hubs.

A strong research community in this area has the potential to have transformative impact in artificial intelligence research with significant concomitant advances in geosciences as well as in other science disciplines, accelerating discoveries and innovating how science is done.

ACKNOWLEDGMENTS

This work was sponsored in part by the Directorate for Computer and Information Science and Engineering (CISE) and the Directorate for Geosciences (GEO) of the US National Science Foundation under awards IIS-1533930 and ICER-1632211. We thank NSF 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. We also thank all the participants in the Research Coordination Network on Intelligent Systems for Geosciences for creating the intellectual space for productive discussions across these disciplines.

REFERENCES

[1] "RAPT: Rare Class Prediction in Absence of True Labels." Mithal, V., Nayak, G., Khandelwal, A., Kumar, V., Oza, N. C.; Nemani, R. IEEE Transactions on Knowledge and Data Engineering, 2017. DOI: 10.1109/TKDE.2017.2739739.

[2] "A Machine Reading System for Assembling Synthetic Paleontological Databases." Peters SE, Zhang C, Livny M, Ré C. PLoS ONE 9(12), 2014.

[3] "Final Report of the 2015 NSF Workshop on Information and Intelligent Systems for Geosciences." Gil, Y. and S. Pierce (Eds). National Science Foundation Workshop Report, October 2015. Available from the NSF IIS collection at the ACM Digital Library: at http://dl.acm.org/collection.cfm?id=C13 and from http://is-geo.org/

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

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

[6] "Challenges and Opportunities in the Hydrologic Sciences." National Research Council, 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, 2012.

[7] "Solar and Space Physics: A Science for a Technological Society." National Research Council, 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. ISBN 978-0-309-16428-3, National Academies Press, Washington, DC, p. 466, 2013.

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

[9] "Sea Change: 2015-2025 Decadal Survey of Ocean Sciences." National Research Council, 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, 2014.

[10] "Linked Data." Berners-Lee, T. Design Issues, available from https://www.w3.org/DesignIssues/LinkedData.html, last retrieved 11 November 2017.

[11] "The Semantic Web in Earth and Space Science. Current Status and Future Directions." T. Narock and P. Fox. Studies in the Semantic Web, IOS Press, 2015.

[12] "Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data." Karpatne, A., Atluri, G., Faghmous, J., Steinbach, M., Banerjee, A., Ganguly, S., Shekhar, S., Samatova, N., and V. Kumar. IEEE Transactions on Knowledge and Data Engineering, 29(10), pp.2318-2331, 2017.