Shanan E. Peters, Dept. Geoscience, University of Wisconsin-Madison

1) Computing innovations that have had the most impact in your field.

My area of geoscience involves sedimentary geology and paleobiology, disciplines which depend on sample-based data. Gathering these type of data requires the direct mediation of humans, who must travel to rock outcrops or drilling facilities, collect, and then process physical samples, some small fraction of which then enter into an automated workflow involving measurement-producing instruments (e.g., mass spectrometers, electron backscatter diffraction. Specific research questions motivate the collection of most sample-based data, but ultimately much of our understanding of fundamental, long-term changes in the Earth system hinge on synthesizing large numbers of observations/measurements so that new synthetic results can be made. For these reasons, I feel the following innovations are of particular note:

1. Advances in web-based GIS technology. One of the most fundament common "joins" between data in all geoscience research domains is space and time. Driven largely by the explosion of location-aware mobile devices, GIS capabilities from open-source initiatives, such as PostGIS and open standards such as GeoJSON, all operating in combination with a rapidly growing and rich set of geographic data services (e.g., Mapbox, Leaflet, OpenStreenMap) has transformed our ability to establish the critical "base map" for geoscience data. This might at first seem mundane, but the advances in the past several years have been marked and this opens up a wide range of possibilities in the sample-based sciences, both quantitative and qualitative.
2. Machine reading and learning approaches. For several hundred years sample-based geoscientists (and scientists generally) have made their data and knowledge available via publications. Synthesizing paleontological, geological, geochemical, and biological data from the literature is critical to many important questions in geoscience, ranging from the causes and consequences of global climate change to the oxygenation of the Earth's atmosphere and oceans. Compiling these data by hand is in many cases prohibitive time consuming and results in a static data resource that is difficult or impossible to assess, improve, or augment with new data. We desperately need a machine reading system that is trained with the language of geoscience, supplied with existing knowledge bases, and capable of locating and extracting structure data from the unstructured published literature. Advances in the past several years have put us down this path, and it is ripe for further exploration and development. Unfortunately, this also requires coordination with publishers, who maintain a rigid grip on the body of scientific work that is being produced today and in the past.

2) discuss science challenges that you think would benefit from innovations in intelligent and information systems research.

I think this is well described by our NSF EarthCube proposal. Below is an excerpt from the Project Summary.

Overview: We are in an era when access to information and data is often less of a problem than our ability to efficiently process and use it. In some cases, these problems are caused by massive, monolithic datasets that are difficult to store, transfer, and/or analyze. In other cases, the first-order problem is discovering and then aggregating relevant data that are widely disseminated in many different locations and formats, such as in the tables, text, and figures of published papers, government agency reports, spreadsheets, and websites. Geoscience currently lacks a cyberinfrastructure that can efficiently, cheaply, and with high precision and accuracy find, extract, and organize many different types of data that are critical to advancing science and leveraging current and past investments in data acquisition. Instead, there are dozens of isolated, sometimes redundant, geoscience data mining efforts that use humans as the primary mechanism for finding data and then keystroking them into structured databases. This mode of operation is not only costly and slow, but it is also an inefficient use of human resources and scientific expertise. Here we propose to develop a geoscience-oriented trained computing system that can serve as a cross-disciplinary tool for rapidly finding, extracting, and organizing geoscience data. Three geoscience domain-specific test cases are proposed, which we believe span much of the complexity that will be encountered when our system is extended to the broader geoscience community. Our longer-term vision is to establish an EarthCube trained computing system that can aid in finding, extracting, and aggregating data, as well as in processing, summarizing, and synthesizing them in a way that helps geoscientists to tackle new problems and better understand and model Earth systems.

Intellectual Merit: There are many scientific problems that can only be addressed when large, synthetic datasets are compiled and made available. Examples include the long-term history of biological extinction and diversification, which is best measured when the global stratigraphic ranges of fossil taxa are tabulated from field-observed occurrences of fossils, and long-term carbon cycling, which is best constrained when many stable carbon isotopic measurements are derived from carbonate rock and organic matter samples and then combined with global mass flux estimates for the burial of inorganic, organic, and authigenic carbon in sediments. A trained computing system capable of facilitating the creation of aggregate synthetic datasets, and augmenting existing datasets with new types of information, could greatly accelerate the pace of scientific discovery and enable geoscientists to more rapidly ask fundamentally new types of questions. Our work seeks to develop this type of transformative capability for EarthCube. In order for us to provide a tangible test of our infrastructure and its capabilities, we propose to develop our building block in the context of three geoscience case studies: paleobiology, sedimentary geology, and structural geology. Each test will generate or augment geoscience databases in ways that can immediately facilitate new science. Thus, our activities will not only address challenges in computing and machine learning as part of EarthCube building block development, but they will also test the ability of our system to enable geoscientists to make new progress in their research.