2015 NSF Workshop on Intelligent Systems for Geosciences
Participant Statement: Kirk Borne (GMU)

1) Summarize 3-4 big ideas in your field in the last few years that you think may have an impact in geosciences research

(a) **Fast Machine Learning on Streaming Data**: Because of the enormous volume and rate at which data are now being collected in many real-time environments, it has become imperative for organizations and project teams to mine, analyze, and respond to massive streaming data in fast dynamic workflows. A previous NSF program in DDDAS (Dynamic Data-Driven Application Systems) focused on fast machine learning for decision support in streaming data applications. Such a program needs to be revived. The world of science (as well as business, government, healthcare, cybersecurity, etc.) needs such fast machine learning techniques and algorithms more than ever. These are essential in order to exploit the maximum discovery potential and to enable the best responses in time-critical geoscience application areas in which vast amounts of sensor data are being collected.

(b) **Predictive Modeling of Spatiotemporal Data**: Time series data lend themselves to predictive modeling. Historic data can be used as training data for predictive models. This area (Predictive Analytics) has become practically synonymous with "Big Data" in the commercial world, as organizations have discovered the great power of data to inform and provide insights into the behavior of all types of systems (behavioral, social, machines, markets, cyber, political, health, etc.). These techniques and commercial tools are receiving enormous attention and R&D investments. Scientists can and should take advantage of these advances. With the added geospatial dimension, more accurate predictive modeling of critical geoscience phenomena is improved and imperative.

(c) **Prescriptive and Cognitive Analytics**: Truly intelligent systems will have more advanced analytics capabilities beyond predictive, such as prescriptive and cognitive. These intelligent systems concepts can be deployed in operational and discovery environments in geoscience. Prescriptive analytics refers to the optimization of a particular outcome (e.g., disaster response, risk mitigation, crop yield, sensor deployment, etc.) based upon insights learned from existing data and sensors. The optimization component of "Prescriptive Analytics" refers to maximizing a benefit function (or minimizing a cost function) whose behavior and dependencies have been learned through mining of sensor data in any system that is being monitored. Cognitive analytics refers to the ability to know the "right questions" to ask (or "right problems" to address) in an application where a comprehensive set of data (both content and context) has been assembled from numerous rich sources: the "360 view" of a geoscience problem or domain from many sensors and sources delivers a deep view into that problem, thus giving scientists improved questioning, inquiry, and exploratory guidance.

2) Highlight 2-3 important research trends in your area that can be relevant to the workshop goals

(a) **Intelligent (Self-Learning) Data Systems**: Machine learning is the core algorithmic engine of artificial intelligence. This should not be applied only to robots, but also to data systems. Machine learning can be used to create metadata (using human-trained algorithms) automatically for large data sets. Such intelligent systems are self-learning, thus enabling scientific data systems to "learn" both context and content metadata as data are ingested. This automatic generation of scientific metadata will greatly enhance the discovery and access of relevant data granules within massive geoscience data repositories. Scientists will be able to focus more time and attention on discovery from data, and less on data discovery.

(b) **Discovery in Semantic Linked Data**: Someone (not Shakespeare) has said "All the world is a graph!" Well, the truth is that "knowledge" consists of the full collection of linked facts within a domain (i.e., a knowledge graph). Consequently, the more connections and links that we can find and instantiate within our geoscience data collections, then the greater will be the scientific knowledge discovery potential of those data repositories. Intelligent systems for geoscience should therefore include attention to ontologies, semantics, linked data, and graph-based capabilities (graph databases, graph analytics, and graph search tools). The exploding interest in social network analysis, and the subsequent generation of powerful tools for exploration of networks, can be applied to any linked-data collection. Since geoscience data are naturally linked (temporally, geospatially, functionally, and scientifically), then it is a natural evolution in geoscience data systems to incorporate this "intelligent" knowledge-based (semantic linked-data) approach to the data science experience of geoscientists.