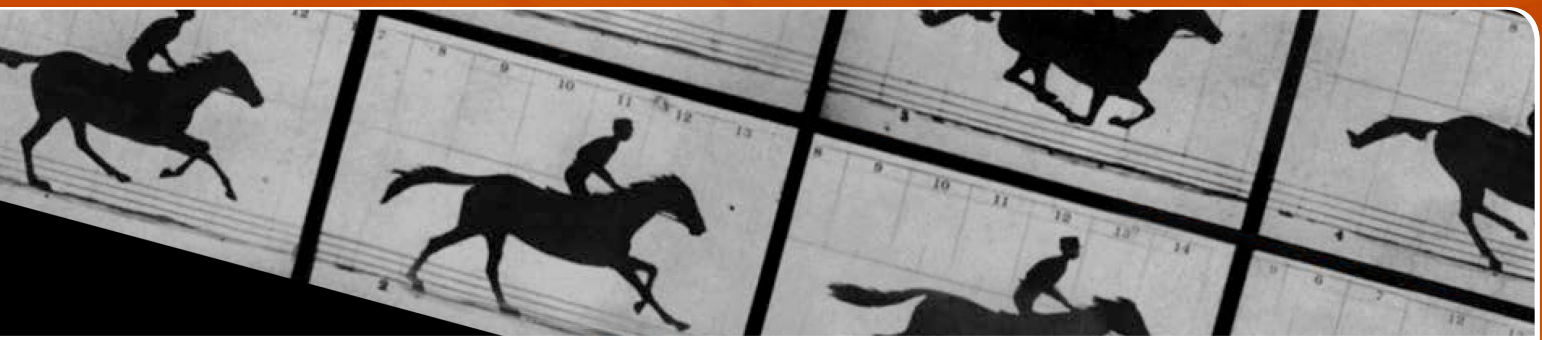
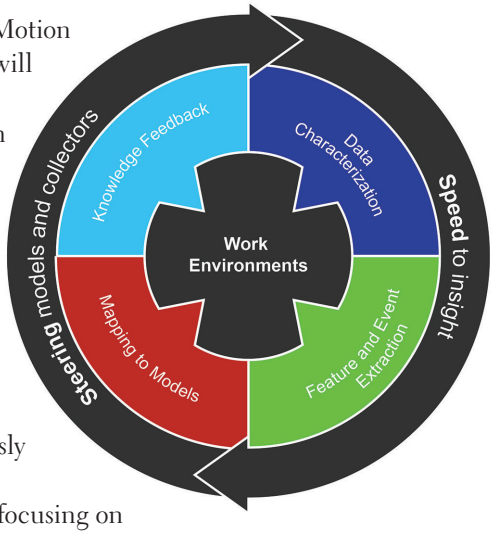
Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*



Analysis in Motion Initiative

The ability to *make sense of data* at larger volumes and faster speeds is foundational to all of Pacific Northwest National Laboratory's science, energy, and security missions. To accelerate scientific discovery, enable timely threat discovery, and make critical decisions, it is necessary to identify and interpret phenomena as they emerge and to adapt our analysis and data collection methods as those phenomena evolve. However, the dominant analysis paradigm in many fields remains post hoc evaluation of results, often relying heavily on manual effort for critical knowledge construction activities like hypothesis generation and testing. This strongly human-centered approach does not always offer the scalability required for timely decision making in big data environments that are increasingly marked by continuous data streams, and result in delaying or preventing necessary decisions and actions.



The Analysis in Motion (AIM) Initiative will develop a new analysis paradigm to provide continuous, automated synthesis of new knowledge and dynamic control of measurement systems contemporaneously with observed phenomena. By focusing on new semi-automated methods for hypothesis construction from streaming data, and by rebalancing effort between humans and machines, AIM will result in an improved ability to rapidly put new observations in the context of evolving domain knowledge. To achieve this goal, AIM will focus its research activities on four areas, the combination of which accelerates a complete analysis cycle that connects data collection, interpretation, and action.

1. **Streaming data characterization** methods that can identify and tag features of importance in high rate, large volume data streams.
2. **Hypothesis generation and testing** methods based on continuously evolving models that can relate features in data to candidate explanations to assist humans and machines in interpreting results and identifying possible future states.
3. **Human-machine feedback** based on capturing human background knowledge through new interaction paradigms, to evaluate candidate hypotheses rapidly and steer models and data collectors in response to evolving knowledge.
4. An integration and execution **environment for streaming analysis** that will provide the infrastructure for testing and demonstrating AIM's advances.

The initiative will initially demonstrate its evolving capabilities in two core strategic areas for the Laboratory: chemical imaging of dynamic systems and disrupting illicit nuclear trafficking. AIM capabilities will help accelerate scientific discovery in chemical imaging through near real time analysis, interpretation, and steering of imaging experiments. Similarly, it will assist in the discovery of hidden trafficking networks by continuously assimilating observations from diverse data sources.

FY 2014 AIM RESEARCH INVESTMENTS

Online Predictive Analytics on Streaming Data

PI: Bobbie-Jo Webb-Robertson

Dealing efficiently with streaming data to identify events, threats, or anomalies is a major challenge in many real-world applications. In particular, a common approach that includes building a static model via machine learning and deploying it for classification of streaming data often quickly results in model failure because of changes in the underlying dynamics of the system that have not been captured or foreseen in generating the model. This has led to a field of incremental machine learning, which is a class of algorithms that were developed to reduce the time, memory, and storage required to train models from fixed massive training sets. These algorithms, however, still remain largely untested for streaming data and are focused primarily on classification and little on anomaly detection on-the-fly. The goal of this research is to develop a computational framework that builds upon incremental machine learning algorithms that perform classification and anomaly detection simultaneously on streaming data. Furthermore, the developed methods will be designed to work with incomplete data, identify hidden variables to help steer future data collection, and quantify the change between one or more states of the model. We believe that combining these improvements in modeling in combination with high-throughput computing can have a major effect on response time to events in streaming data for which the dynamics are continually changing.

Streaming Hypothesis Reasoning

PI: Jesse Weaver

AIM will advance streaming reasoning techniques to overcome a limitation in contemporary inference that performs analysis only over data in a fixed cache or a moving window. This research will lead to methods that continuously shed light on proposed hypotheses as new knowledge arrives from streams of propositions, with a particular emphasis on the effect that removing the expectation of completeness has on the soundness and performance of deductive algorithms. The work will address challenges in sampling rates, cache maintenance, deductive reasoning, ranking of conclusions, and handling uncertainty in a logically coherent way. Using a fixed-size cache, the challenge is to identify and preserve "interesting" assertions in a stream. Deductive reasoning will continuously be performed over the cache so as to draw conclusions (relevant to proposed hypotheses) as quickly as possible. The use of a cache differentiates our work from state-of-the-art works in deductive stream reasoning in that the cache enables us to "remember" propositions that are no longer in the stream window. Such a capability is potentially transformative to data-intensive science as well as investigative processes that draw from heterogeneous data sources. The primary challenge in the research is in determining which propositions should be cached at any point in time in order to maximize the benefit of deductive reasoning to the user. Subjective logic will enable us to overcome the inability of traditional deductive inference to account for uncertainty.

Science of Interaction

PI: Alex Endert

Sensemaking is the cognitive process of gaining understanding of data by testing one's own assertions and domain expertise against the data. Current approaches to incorporating the user's domain expertise into visual and mathematical analytic systems rely on user interfaces that require humans to explicitly steer the underlying computation—for instance through sliders, menus, and knobs—that correspond directly to parameters of the specific algorithm or technique underlying a visualization. This approach for user interaction (called direct manipulation) has been the predominant method for interacting with visual displays during the sensemaking process. However, as data rates and analytic complexity increase, human users can no longer be expected to manually adjust model parameters. Rather, a new interaction method is needed in which user knowledge is captured continuously and implicitly as users work, with little or no additional burden on them. This knowledge is then fed back into evolving models and hypothesis generation algorithms. AIM's research in human-machine feedback will advance a user interaction paradigm called semantic interaction that demonstrates how interaction with an information space can be used as evidence for the underlying cognitive discourse users have with information. As a result, capturing that tacit knowledge associated with user interaction in a mathematically structured form can lead to a more fluid cognitive process of discovery, leading to faster and more meaningful insights. This research will enable users of AIM capabilities to transition between explicit, formal assertions or hypotheses that algorithmic techniques can compute on, and the informal, fluid exploration that users require for discovery.

For more information, please contact:

Bill Pike
Pacific Northwest National Laboratory
william.pike@pnnl.gov | (509) 375-2689