Analysis in Motion (AIM)

Kerstin Kleese van Dam (BNL), Mark Greaves (PNNL), Rob Jasper (PNNL), Nigel Browning (PNNL)

The ability to interactively make sense of data at large volumes and faster speeds is foundational to many national mission areas in science, energy, health, national security and industry. These domains are driven by the need to assimilate and interpret ever-increasing volumes of data to accelerate scientific discovery and make critical decisions. In these domains, the speed of analysis can be as important to the final outcome as the choice of data to be collected. Analysis in Motion is developing a new advanced analysis paradigm -- persistent / dynamic knowledge synthesis - in which we tightly integrate high velocity streaming analysis with human in the loop decision and sense making in one continuous process.

Analysis in Motion is a collection of focused research projects focused specifically on developing analytical methodologies for scenarios where human insight is critical to a successful outcome. These scenarios are characterized by at least some of the following:

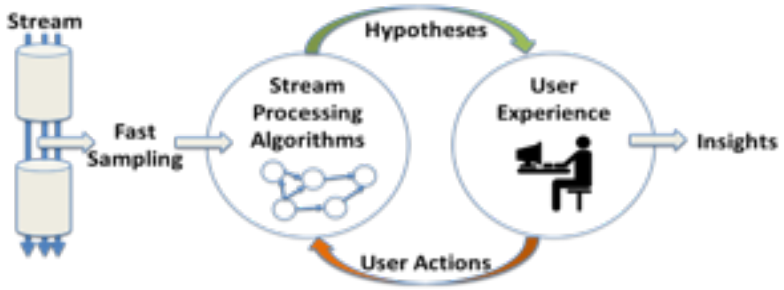
- Data arrive at such high velocity and volume that storage of the complete dataset for later analysis might not be practical.
- Critical decisions have to be made while the data are still arriving
- Events of interest are so rare, complex, or context-dependent that it is not possible to train algorithms sufficiently for reliable detection
- Human expert knowledge (e.g., specific scientific expertise or knowledge of the specific experimental context) is required for accurate interpretation of the data.

One such use case is the steering of high end electron microscopy experiments. In Transmission Electron Microscopy (TEM), a beam of electrons is transmitted through an ultra-thin specimen, interacting with the specimen as it passes through. These experiments can generate atomic resolution diffraction patterns, images and spectra under wide ranging environmental conditions. In-situ observations with these instruments, were physical, chemical or biological processes and phenomena are observed as they evolve, generate from 10GB-10's of TB (e.g. at BNL) of data per experiment (and getting larger) at rates ranging from 100 images/sec for basic instruments to 1600 images/sec for state of the art systems.

As with many experimental technologies, TEM scientists have to take critical decisions during the experiment that will affect the success or failure of their work. Higher-dose electron beams can damage the sample, while at the same time high dose beams are required to capture the quality of image required by science. It is therefore important to vary the beam intensity throughout the experiment to be able to image a crucial phenomena or process of interest, while minimizing damage to the sample. Furthermore it is often necessary to focus in on a particular area of the sample (which cannot be determined ahead of time) to capture the process or phenomena in question in sufficient resolution. Potentially several of such processes occur at the same time and so the scientist needs to decided which ones are of most scientific relevance, i.e. prove or disprove their hypothesis, or exhibit a new rarely observed detail. The experiment represents thus a dynamic optimization process of beam intensity, scanning speed and focus in the context of the scientific drivers for the experiment, the sample properties and background knowledge of existing scientific insights in the field. Timing, occurrence of phenomena and optimized steering decisions will vary with every sample, experiment, and experimenter.

The challenges of electron microscopy are representative of many other experimental instruments, where users would significantly benefit from a streaming analysis environment. As we have worked on these use cases, we have identified the following core components that a successful streaming analysis environment would need to incorporate:

- Streaming on-line Statistics, Data Mining and Machine Learning - to identify, identify and trace emerging phenomena in high velocity streaming data.
- Streaming Deductive Reasoning - to determine what is of interest and impact in large volumes of phenomena and to generate candidate explanations for phenomena of interest that support the scientists in their interpretation.
- Human-Computer collaboration supporting software components that to jointly work on data collection, analysis, reasoning and insights generation.
- Provenance - to document which decisions were taken during the analysis process to enable the explanation and verification of the results.
- Adaptive workflows - supporting the flexible adaptation of analysis and deductive reasoning algorithm ensembles at runtime in response to the evolving experimental and scientific priorities.
- Streaming Analysis Infrastructure - guaranteeing the timely and ordered delivery of streaming events to the different analysis infrastructure components, as well as providing the accompanying flexible resource discovery and scheduling.



The Analysis in Motion Initiative (http://aim.pnnl.gov) is a multi-year research initiative led by Pacific Northwest National Laboratory, with partners including Brookhaven National Laboratory, Washington State University, Rensselaer Polytechnic Institute, Ohio State, Georgia Tech, and several others. Since its beginning in 2014, the initiative has developed initial versions of the core components identified above (apart from on-demand prediction and adaptive workflows) and applied these successfully to streaming analysis in Nuclear Magnetic Resonance (NMR), Transmission Electron Microscopy and National Security.

Key challenges encountered in the initial implementation and testing phase included:

- Achieving Scalability - For the algorithms, for tightly coupled workflows across different programming languages and models, in the infrastructure while maintaining message ordering, in the hardware given the high data volumes.

- Establishing and Verifying Ground Truth, providing reproducibility in a highly adaptive environment with extensive user interaction, input and steering.