Analysis and Decision Making in Big Data Environments

Kerstin Kleese van Dam (BNL), Mark Greaves (PNNL)
HTTP://AIM.PNNL.GOV
kleese@bnl.gov

BROOKHAVEN NATIONAL LABORATORY
a passion for discovery

U.S. DEPARTMENT OF ENERGY | Office of Science

TEM: A Widely Used Research Tool

- In Transmission Electron Microscopy (TEM) a beam of electrons is transmitted through an ultra-thin specimen, interacting with the specimen as it passes through.
- Can generate atomic resolution diffraction patterns, images and spectra under wide ranging environmental conditions
- ~500 aberration corrected (S)TEM worldwide with ~20-30 in National Laboratories
- In-situ observations 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 3GB/sec for state of the art

Characteristics Of TEM Data Analysis

- **Every Experiment is different**
- Today pre-dominant analysis method is the post-doc
- New TEM systems observe processes in-operando - steering experiment to successful outcomes
- An event is one image (2Kx2K, 4Kx4K), arriving at up to 400 images / sec - 3GB/sec
- Feature detection in a high noise to signal ratio data
- Several features could form a processes, **events are connected, order matters!**
- Possibly several process at any given point in time, only one can be captured in sufficient detail, need human insight to prioritize
- Ability to interpret early indicators of processes requires collaboration between scientist and computer

PNNL Analysis in Motion Initiative

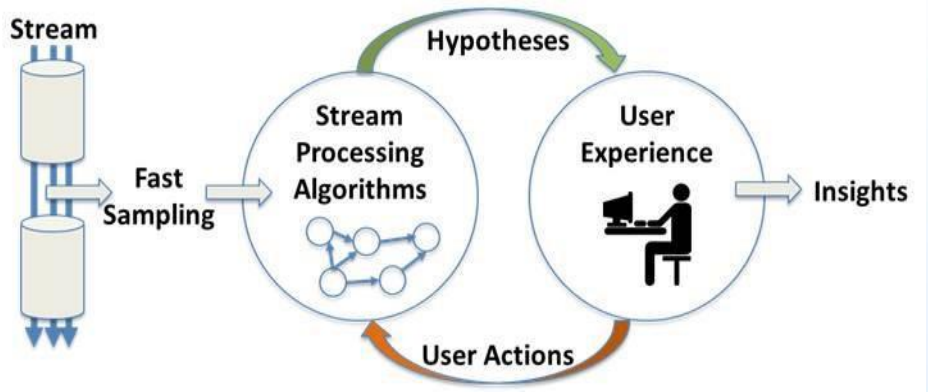
CHALLENGE

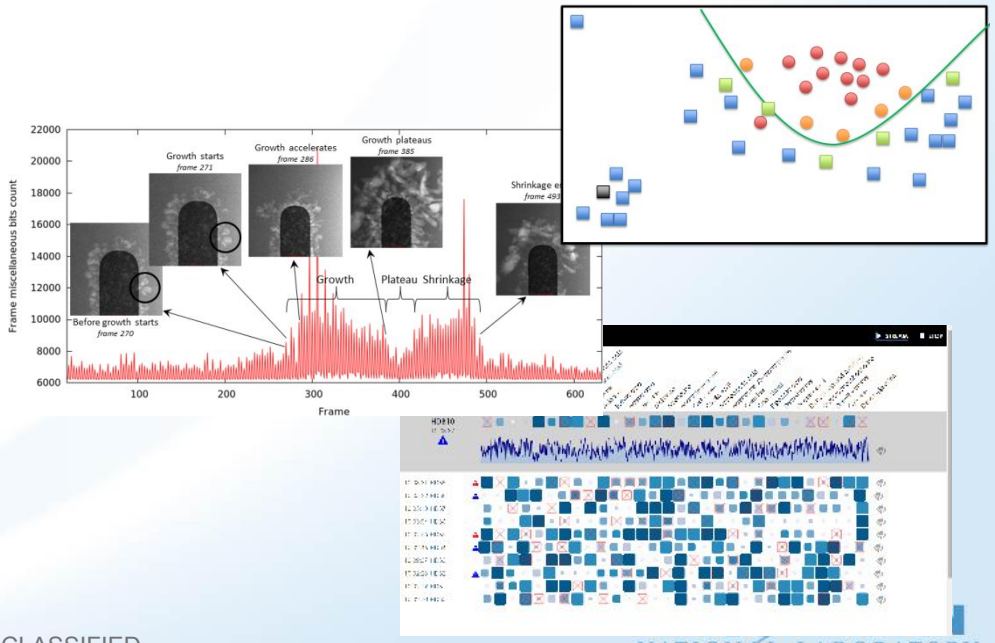
Enable users to identify, analyze, interpret and react to emerging phenomena in a high velocity, high volume streaming data environment.

Approach

Analysis in Motion is an integrated infrastructure that uses multiple classifier systems connected via a fast messaging framework to derive insights from dynamic data that is tracked in real time. It uses human insight to guide the analysis and data acquisition process.

AIM has partners in Universities and at BNL.





How Does It Work

Single-pass

- No access to the data stream beyond the sample

Data is forgotten

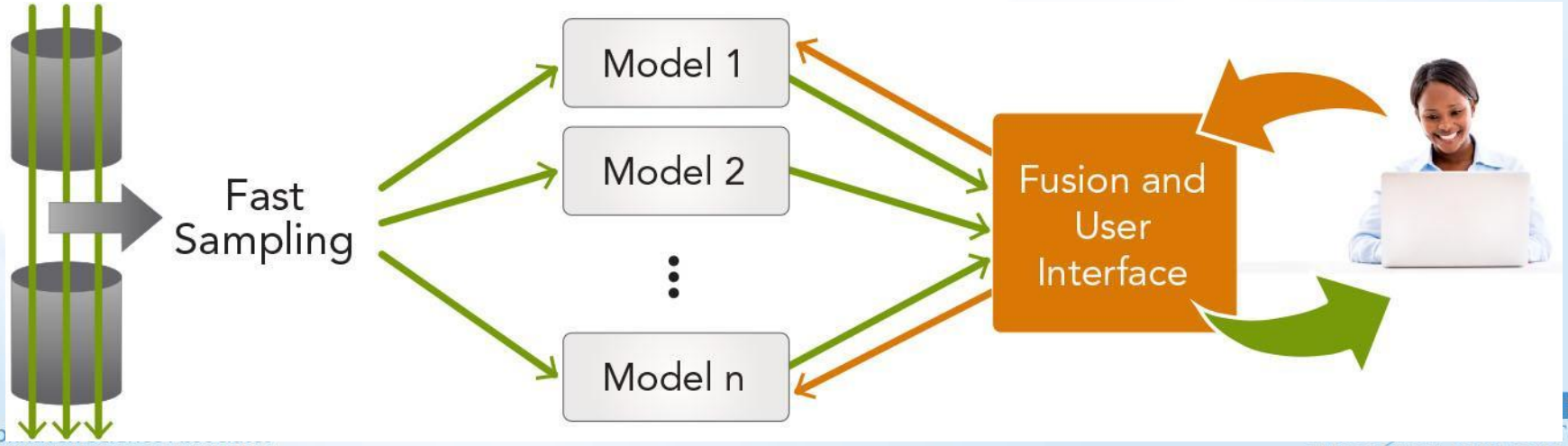
- Each model's cache is small relative to the data volume

Many algorithms options

- Ideal combination of algorithms needs to be determined at runtime

Cooperative user

- Important problem knowledge isn't in training data



What Does it Take?

- Identify emerging phenomena in high velocity streaming data **- Streaming Statistics, Data Mining, Machine Learning**
- Determine what is of interest and impact, generate candidate explanations **– Streaming Deductive Reasoning**
- Human-Computer collaboration to jointly adjust data collection, reasoning and insights **– Cognitive Depletion Detection, Hypothesis Exchange**
- Document which decisions were taken during the analysis process to explain the results **- Provenance**
- High performance event processing infrastructure **- Programming Models, Performance Models, Adaptive Workflows, Fast Networking**

First Experiences

- Initial Infrastructure Implementation - Apache Kafka
  - Reviewed many different solutions, final decision based on cost of entry and programing model support.
  - Easily achieved over 600,000 messages/sec for small messages
  - Challenge discovered - in most commercial system message ordering is not the issue, in science it is, implementation comes with initial severe performance hits
- Incremental Analysis is quite different
  - Combining models improves results, but improvement not always worth the cost
  - researchers need to adapt to new approach and create models that scale, often both skills not in one person
- Human - Computer Collaboration is Key
  - Hypothesis exchange through storyline representations and new interaction models

Not Addressed Yet

- Adaptive Workflows
- Effectively connecting different programing models in a streaming analysis process
- Analysis placement - in-situ, in-transit, remote - needs performance investigation and modeling
- How to achieve: Speed + high velocity + high volume