Analysis in Motion Initiative

NIAC DAY@PNNL

Presented by: Mark Greaves
AIM is developing new methods for semi-automated knowledge discovery from high-volume data streams. AIM will reduce the time to discovery by performing hypothesis generation and testing in parallel with a stream.
AIM Streaming Context

- Data is forgotten
  - Each model’s cache is small relative to the data volume
- Single-pass
  - No access to the data stream beyond the sample
- Cooperative user
  - Important problem knowledge isn’t in training data
AIM Initiative Goals

- AIM will develop new techniques for building multiple classifier systems in a streaming context
  - Employ a user-directed fusion function to augment training data, so that the user can iteratively tweak and re-weight models on the fly
  - Include diverse dynamic model types (symbolic, PGMs, terminological)
  - Use high-level user feedback to steer the data production system

- FY 2014 research focus
  - Get smart
    - Build a broad resource of known streaming algorithms and techniques
  - Fail fast
    - Can we perform scalable symbolic deduction on streams?
    - Can statistical models evolve new structures to track the stream?
    - Can we gain useful information from implicit user behavior?
  - Get ready
    - Construct AIM’s integration and testing infrastructure
AIM Research Challenges

▶ **Algorithms**
- Stream sampling and cache maintenance/eviction
- Anytime online algorithms for feature extraction and analytics
- Algorithm ensembles and multiple classifier systems
- Model evolution
- Continuous time-sensitive hypothesis generation, testing, and filtering
- Non-relational and noisy data formats
- Scaling to high data rates

▶ **Human-machine feedback**
- UX to usefully perform model steering and training
- UX for data exploration and hypothesis testing in a streaming context
- Hypothesis depiction
Additional Expertise Needs

- All of the previously identified research challenges

- Streaming data wrangling
  - Data ingestion and cleaning
  - Normalization and semantic processing
  - Feature extraction over streams

- Cloud-based stream processing architectures

- Novel stream analytics algorithms and approaches

- Processing distributed streams