INTEGRATED VEHICLE HEALTH MANAGEMENT

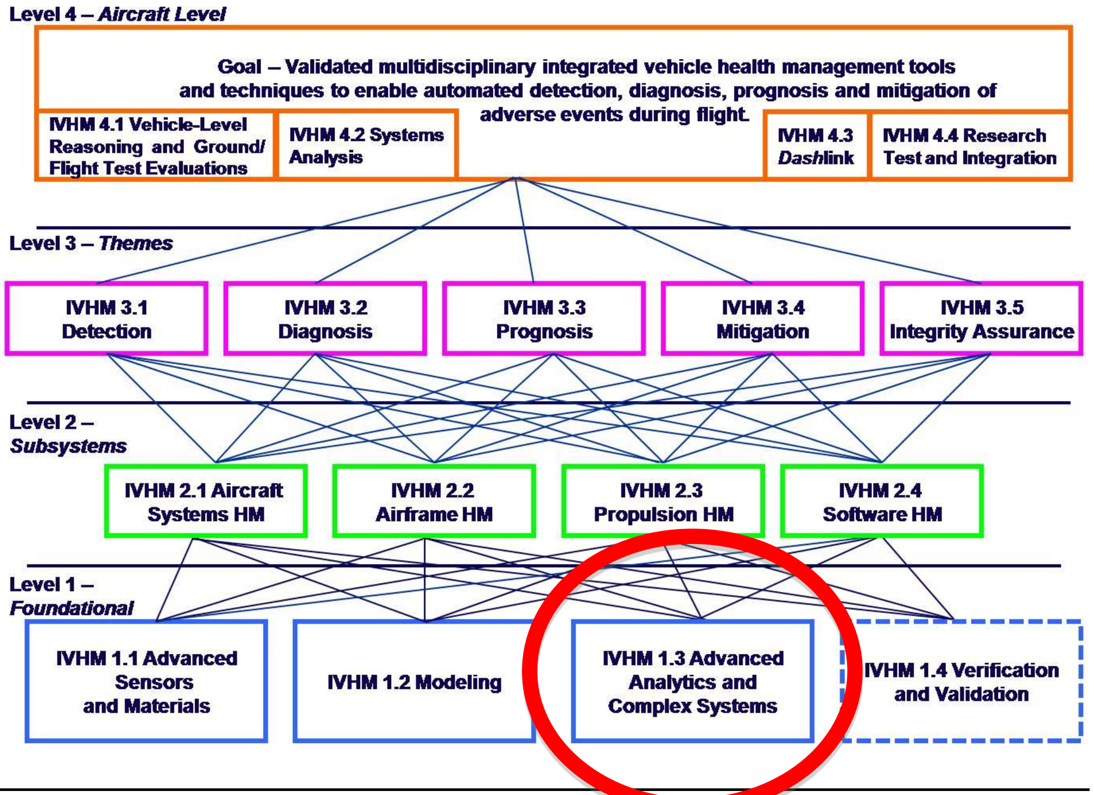
Data Mining for Fleet-wide Health Monitoring

Nikunj Oza

Aviation Safety Program Technical Conference
November 17-19, 2009
Washington D.C.

Outline

- Problem Statement
- Background
- IVHM milestones(s) being addressed
- Approach
- Results
- Conclusions
- Future Plans



Level 4 – *Aircraft Level*

Goal – Validated multidisciplinary integrated vehicle health management tools and techniques to enable automated detection, diagnosis, prognosis and mitigation of adverse events during flight.

IVHM 4.1 Vehicle-Level Reasoning and Ground/ Flight Test Evaluations

IVHM 4.2 Systems Analysis

IVHM 4.3 *Dashlink*

IVHM 4.4 Research Test and Integration

Level 3 – *Themes*

IVHM 3.1 Detection

IVHM 3.2 Diagnosis

IVHM 3.3 Prognosis

IVHM 3.4 Mitigation

IVHM 3.5 Integrity Assurance

Level 2 – *Subsystems*

IVHM 2.1 Aircraft Systems HM

IVHM 2.2 Airframe HM

IVHM 2.3 Propulsion HM

IVHM 2.4 Software HM

Level 1 – *Foundational*

IVHM 1.1 Advanced Sensors and Materials

IVHM 1.2 Modeling

IVHM 1.3 Advanced Analytics and Complex Systems

IVHM 1.4 Verification and Validation

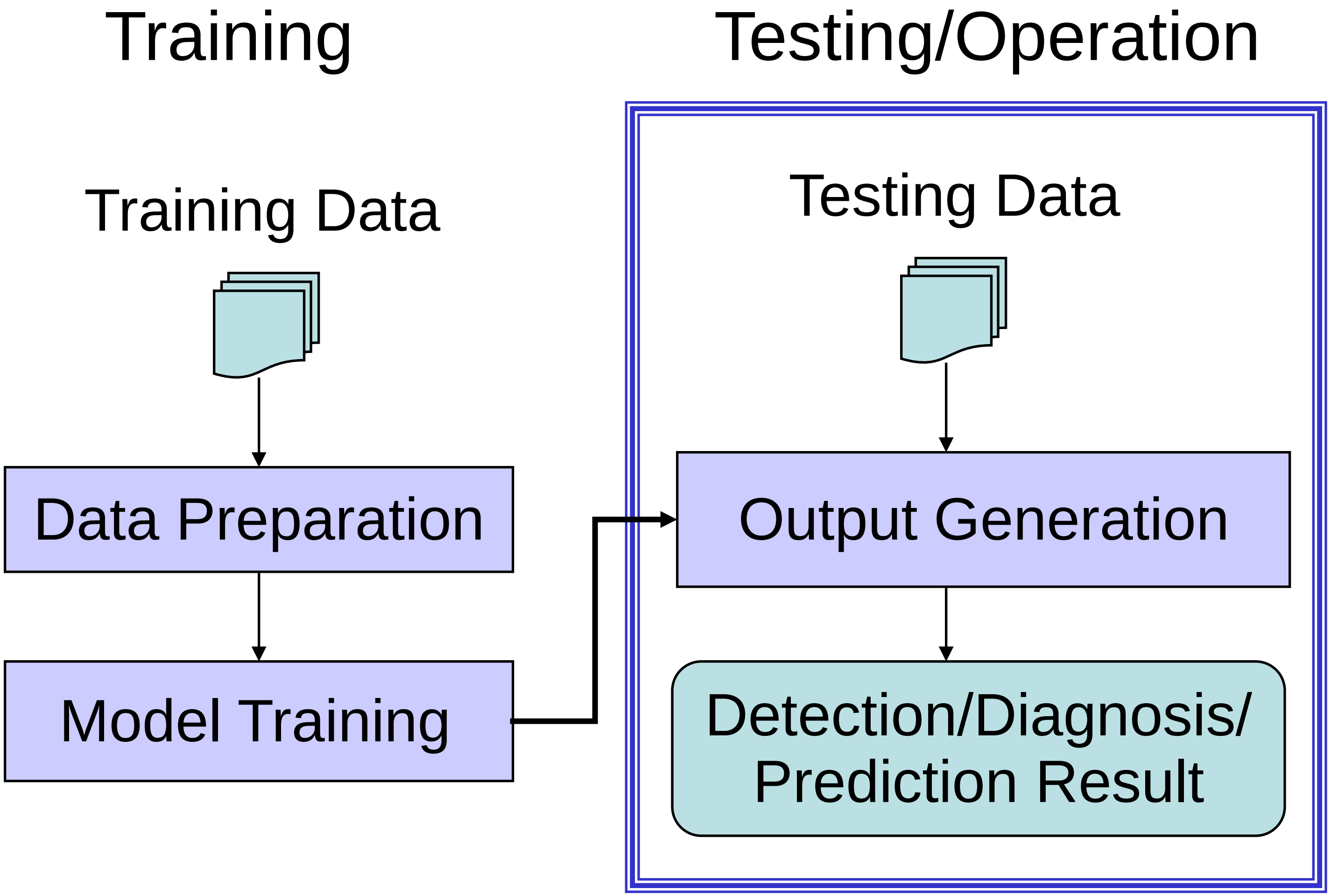
Problem Statement

Problem: Detection, diagnosis, prediction of anomalies across fleets of aircraft

Key components of solution

- Learn from data representing past operations of multiple aircraft---mostly or all normal operations.
- Heterogeneous data: discrete and continuous data representing entire aircraft and/or text describing operations or maintenance incidents
- Results/Desired Results
  - Tool that detects, diagnosis, and predicts anomalies in previously-unseen data.
  - Works across flights (e.g., within how many flights will problem become serious)
  - Works within flight (e.g., can current flight be completed)
  - Tools that allow interactive exploration (e.g., time series search)
- Emphases on accuracy, speed, handling large datasets that are sometimes distributed.

Background



Training

Testing/Operation

Training Data

Data Preparation

Model Training

Testing Data

Output Generation

Detection/Diagnosis/ Prediction Result

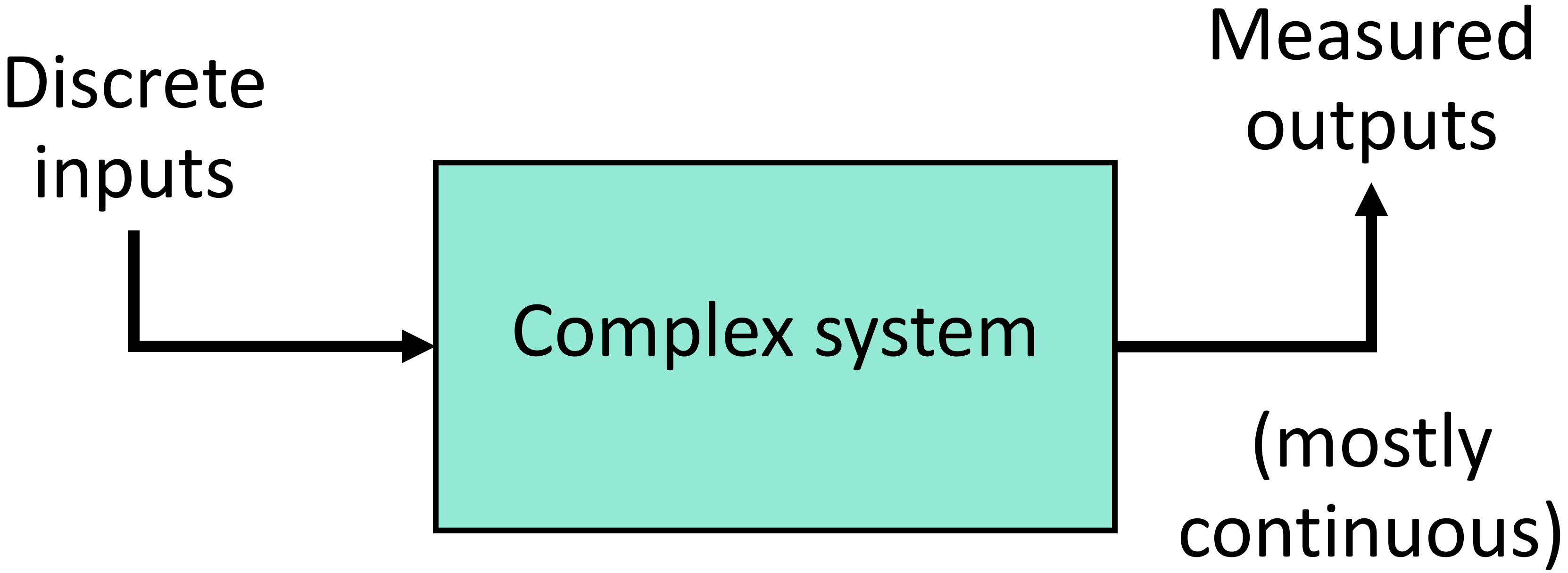
IVHM milestones(s) being worked

- Milestones 1.3.*.* are the Advanced Analytics milestones
- Particular work for this presentation
  - Milestone 1.3.1.1: Demonstrate automated anomaly detection in an offline mode on large heterogeneous datasets from multiple aircraft. Metric (i): Demonstrate ability to perform anomaly detection on a dataset containing both discrete symbols and continuous data streams and show a detection rate of at least 85%.
  - Milestone 1.3.1.3: Develop and demonstrate anomaly detection algorithms for continuous data sources in distributed data sources. Metric (iii): Demonstrate the anomaly detection algorithm's ability to produce identical results to a centralized algorithm 100% of the time on a distributed system consisting of three nodes with 10GB of data in each node.
  - Milestone 1.3.1.5: Multivariate time-series search.

Milestone 1.3.1.1

- Demonstrate automated anomaly detection in an offline mode on large heterogeneous datasets from multiple aircraft.
- Metric (i): Demonstrate ability to perform anomaly detection on a dataset containing both discrete symbols and continuous data streams and show a detection rate of at least 85%.
- Relevance to Aviation Safety
  - Data from aircraft consist of discrete and continuous sequences.
  - Need to identify anomalies involving all the variables
    - Ties between continuous and discrete variables (e.g., delays)
    - Anomalies within discrete or continuous variables
  - Anomalies often of a sequential nature
    - Looking at snapshots in time not sufficient.
    - Go beyond just consecutive pairs of time steps. Extend to gapped sequences.

Approach: Heterogeneous Anomaly Detection



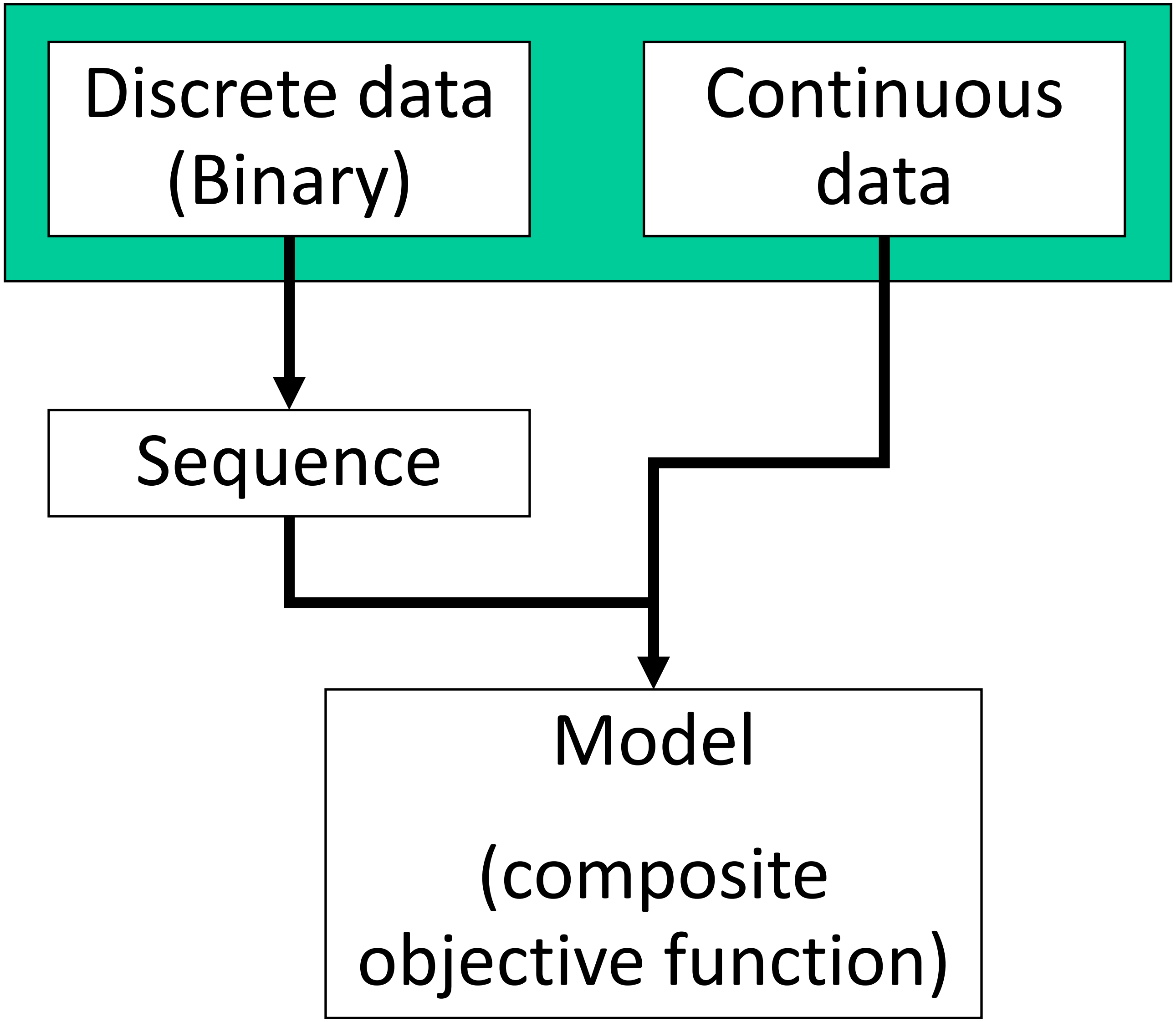
Interesting anomalies in discrete sequences

- Type I: Sequence of switching that was expected at a given stage did not occur.
- Type II: Sequence of switching that was not expected did occur.
- Type III: Sequence of switching occurred in an unexpected order.

Methodology

- To detect anomalies in heterogeneous data sets, we developed a kernel-based approach.
- Experimented using this methodology with a synthetic dataset that has anomalies of the type of interest.

Details



Discrete data (Binary)

Continuous data

Sequence

Model

(composite objective function)

Optimization problem

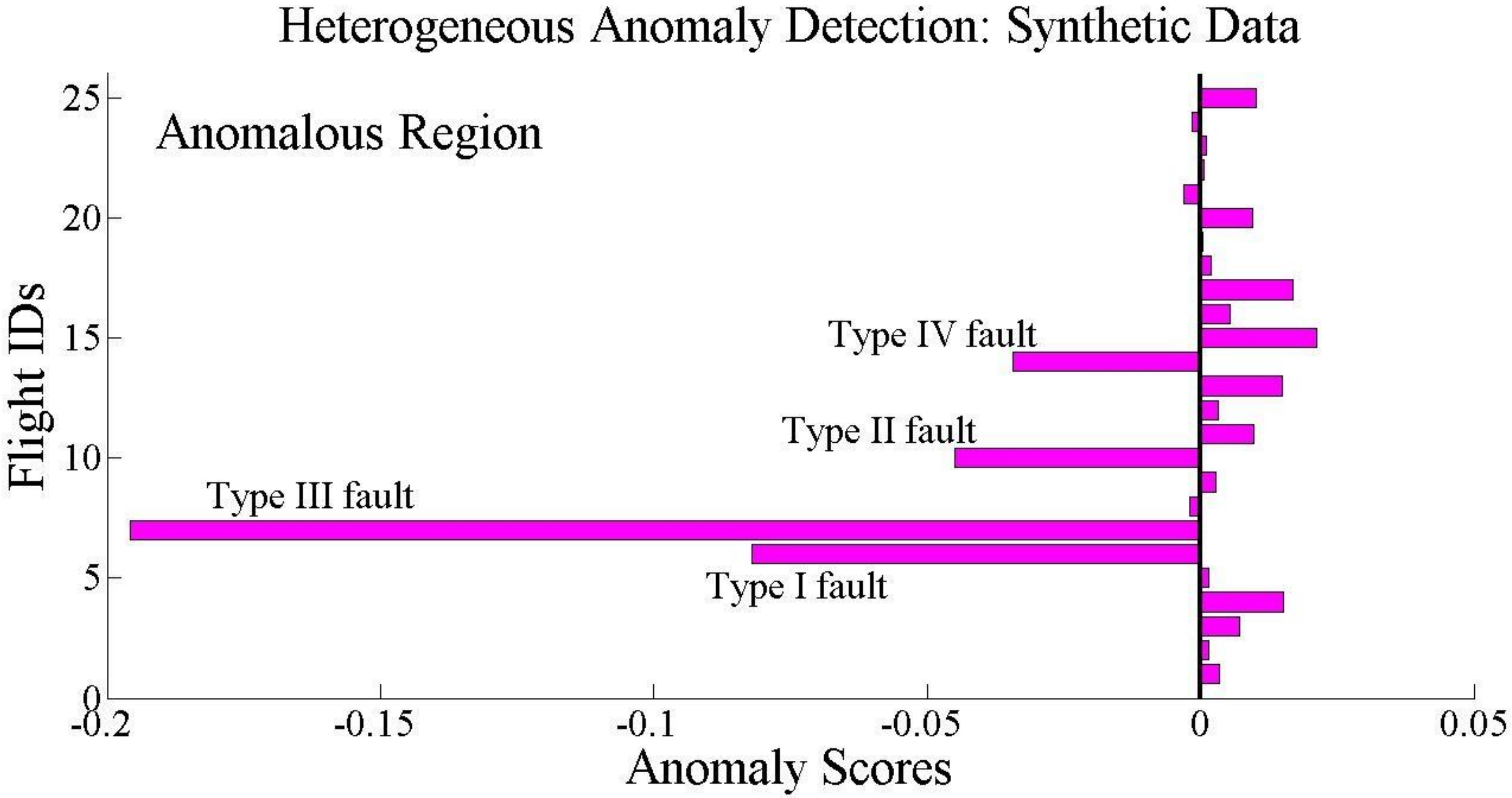
One-class Support Vector Machines (SVMs) perform anomaly detection by mapping the original data into a much higher dimensional space and then finding a small fraction of the training data (anomalies) that can be linearly separated from the remainder. The resulting model can be used to classify new examples.

$$\text{minimize} \quad Q = \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j \left(\beta \underbrace{K_d}_{\boxed{\text{Discrete kernel}}}(x_i, x_j) + (1-\beta)\underbrace{K_c}_{\boxed{\text{Continuous kernel}}}(x_i, x_j)\right)$$

$$\text{subject to} \quad 0 \le \alpha_i \le \frac{1}{\ell\nu}, \quad \nu \in [0,1], \quad \sum_i \alpha_i = 1$$

In the objective function, each entry of the discrete kernel and the continuous kernel represents the score obtained using longest common subsequence (LCS) and radial basis function (RBF) respectively.

Results



Achieved 100% detection, with a few false alarms if threshold is set to zero. However, there is a very large difference in anomaly scores between false alarms and correct detections.

Milestone 1.3.1.3

- Develop and demonstrate anomaly detection algorithms for continuous data sources in three distributed data sources.
- Metric (iii): Demonstrate the anomaly detection algorithm’s ability to produce identical results to a centralized algorithm 100% of the time on a distributed system consisting of three nodes with 10GB of data in each node.
- Relevance to Aviation Safety
  - Aviation Safety datasets are often large and distributed (stored in many places)---cannot be collected on one computer to run analysis tools.
  - Smaller datasets can also not necessarily be collected in an onboard setting (e.g., many sensors on a single aircraft).
  - MUST obtain same answer as what would be obtained when data is centralized, i.e., all information available at once. This is vital for safety applications.

Methodology

- We developed and implemented two distributed anomaly detection algorithms which are:
  - Provably correct with respect to centralized results
  - Decentralized
  - Very low communication overhead
  - Handle dynamic data
- In the absence of real distributed datasets, we experimented with synthetic dataset

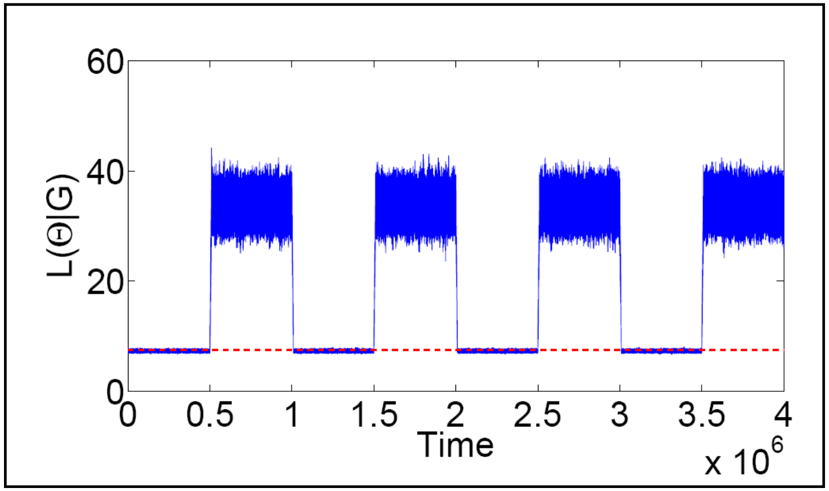
Details

- Distributed EM algorithm (Gaussian Mixture Model based)
  - Detects changes in the data at multiple nodes and adjusts the model accordingly. The anomalies correspond to points that do not fit a GMM. The detection was done based on the log-likelihood of the global data.
- Distributed distance-based
  - A point is considered anomalous if it is far away from its neighbors. Nodes exchange local outlier information with their neighbors until they all agree.

In both these cases, outliers are detected based on all nodes’ data and ***guaranteed to produce the same results*** as a corresponding centralized algorithm.

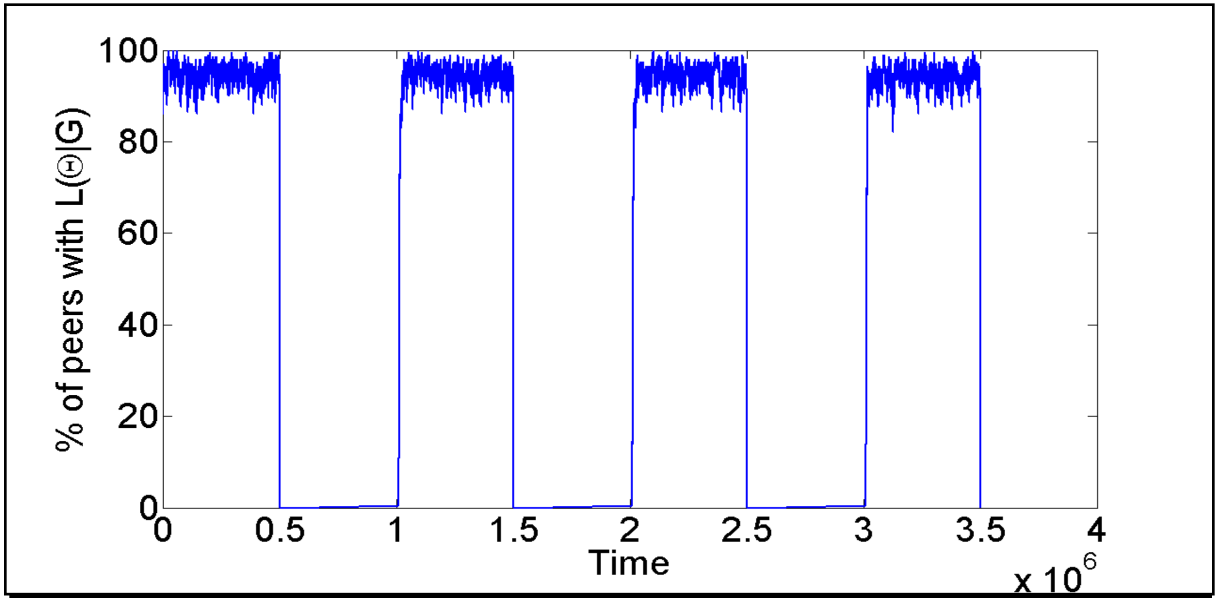
Experiments (1) EM

- Dataset
  - Synthetically generated using 5 Gaussians, 10 dimensions
  - Random choice of mean, variance of each Gaussian and mixing probability
  - Changed the mean at random intervals to simulate changes in distribution
  - Up to 100 nodes in the network, 11GB data per node
- Measurement
  - Measured the number of nodes correctly raising alert and number of messages exchanged

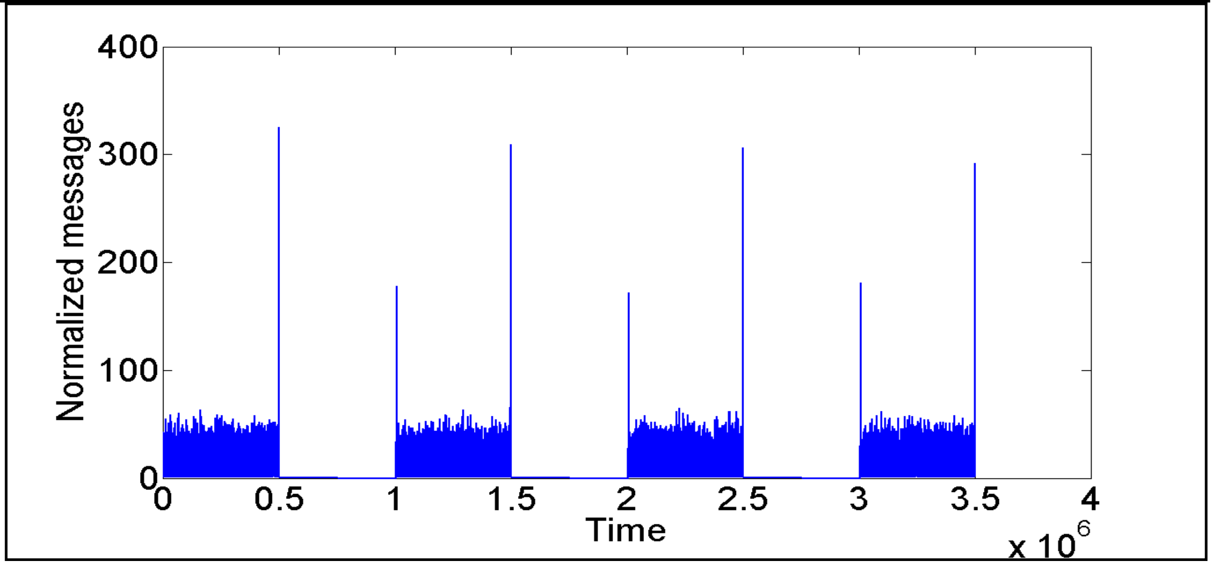


Typical dataset (red line shows threshold)

Distributed EM as detector

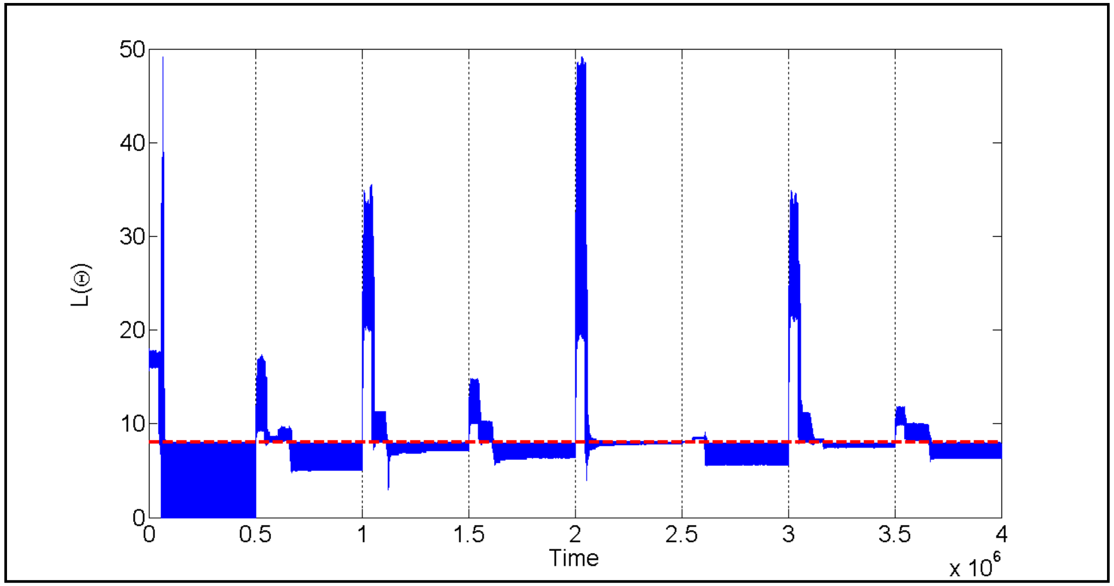


Quality

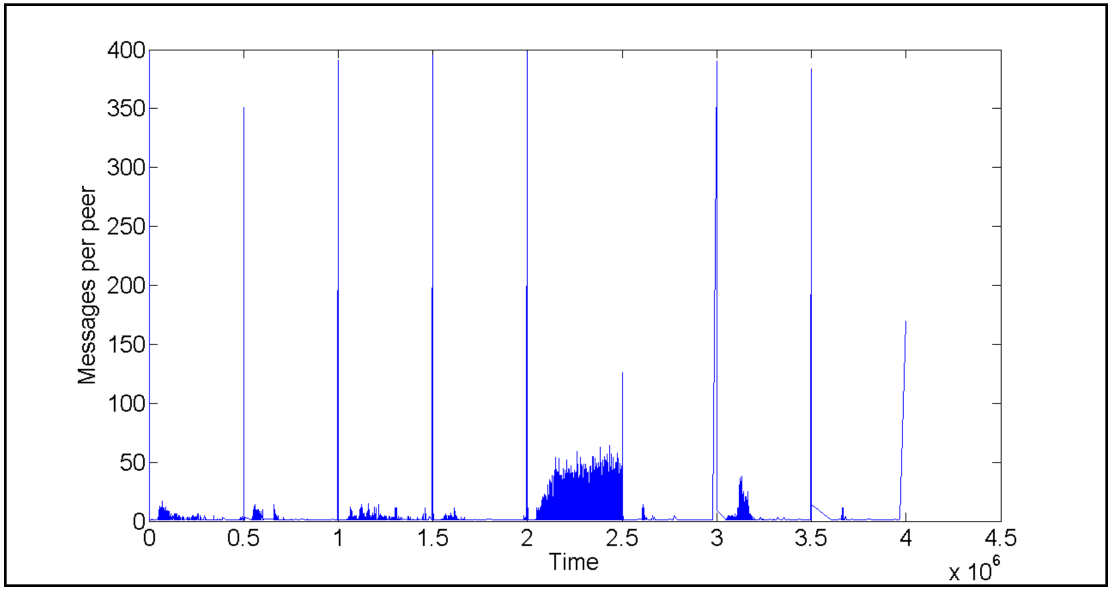


Messages

Distributed EM for building models



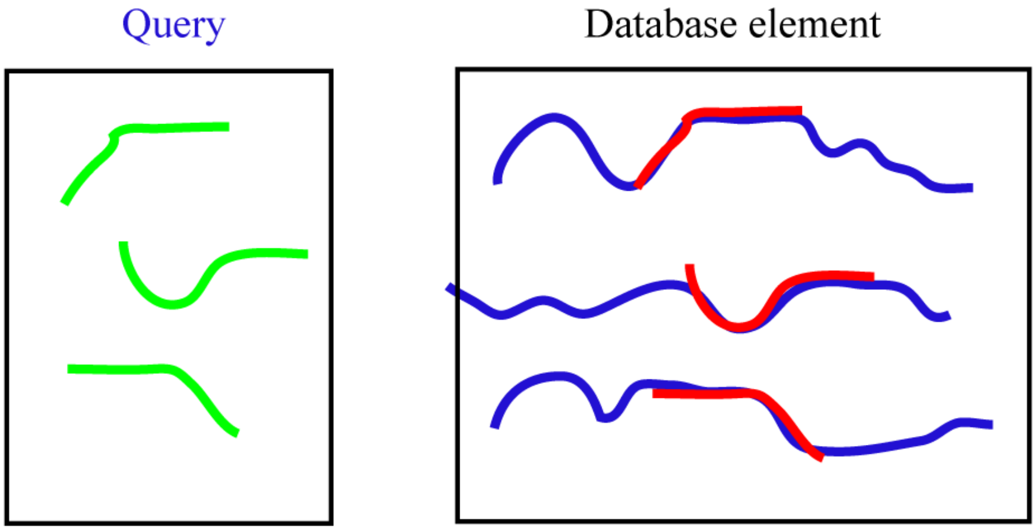
Quality



Messages

Multivariate Time Series Search

- Want a “Google” for multivariate time series.
- Given
  - Collection of multivariate time series (e.g., data from flights)
  - Multivariate query
    - Query over a relatively small set of variables (a handful)
    - Arbitrary time shifts over query variables
    - A threshold for every query variable
- Find all examples close enough to the query in the collection



Multivariate Time Series Search

- Ours is the first work that
  - Allows queries on ***any*** subset of variables
  - Allows arbitrary time shifts over those variables
  - Allows for parallel implementation
- We tested on largest datasets ever considered in time series search literature.
- See poster for more details.

Deployment to Southwest Airlines

- Experimenting with two anomaly detection algorithms (IMS and Orca) that work on continuous data.
  - Inductive Monitoring System (IMS) clusters normal data, then assesses new data by checking their distances to the closest cluster. Larger distances imply more anomalousness.
  - Orca calculates a test data point’s distance (or average distance) to kth nearest neighbor in the training set. The larger this distance, the more anomalous the point. Uses pruning rule to achieve near-linear running time.
- Preliminary results displayed in an AvSafe poster.

Conclusions

- Data Mining for Fleetwide Health Monitoring is a comprehensive area of work involving different types of data stored in different ways with different desired results.
- See posters for more details on
  - Heterogeneous anomaly detection
  - Distributed anomaly detection
  - Multivariate time series search

Next Steps

- Development of methods for anomaly diagnosis and prediction.
- Development of interactive tools with which to use these methods
- One example of such a system is the Event Cube---see the next presentation by Prof. Jiawei Han.