Advances in Anomaly Detection

Tom Dietterich
Alan Fern
Weng-Keen Wong

Sharmodeep Battacharyya
Debashis Mondal

Andrew Emmott
Shubhomoy Das
Risheek Garrepalli
Zoe Juozapaitis
Si Liu
Md. Amran Siddiqui
Tadesse Zemicheal

Oregon State University

Outline

- Defining the Anomaly Detection Problem
- Benchmarking Current Algorithms for Unsupervised Anomaly Detection
- PAC Theory of Rare Pattern Anomaly Detection
- Incorporating Analyst Feedback
- Applications
  - Weather network anomaly detection
  - Open Category detection

Defining Anomaly Detection

- Data $x_1, \ldots, x_N$, each $x_i \in \Re^d$
- Mixture of "nominal" points and "anomaly" points
- Anomaly points are generated by a different process than the nominal points
- Anomaly detector: $A(x) =$ anomaly score
- Goals:
  - Find all of the anomalies in the training data
  - Determine whether a new query point x_q is an anomaly

Three Settings

- Supervised
  - Training data labeled with “nominal” or “anomaly”
- Clean
  - Training data are all “nominal”, test data contaminated with “anomaly” points.
- Unsupervised
  - Training and test data consist of mixture of “nominal” and “anomaly” points

Well-Defined Anomaly Distribution Assumption

- WDAD: the anomalies are drawn from a well-defined probability distribution
  - example: repeated instances of known machine failures

- The WDAD assumption is often risky
  - adversarial situations (fraud, insider threats, cyber security)
  - diverse set of potential causes (novel device failure modes)
  - user’s notion of “anomaly” changes with time (e.g., anomaly == “interesting point”)

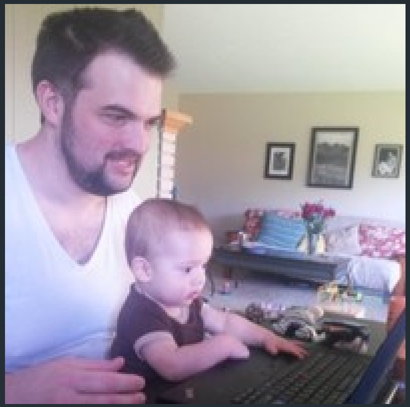
Strategies for Unsupervised Anomaly Detection

- Let α be the fraction of training points that are anomalies
- Case 1: α is large (e.g., > 5%)
  - Fit a 2-component mixture model
    - Requires WDAD assumption
    - Mixture components must be identifiable
    - Mixture components cannot have large overlap in high density regions
- Case 2: α is small (e.g., 1%, 0.1%, 0.01%, 0.001%)
  - Anomaly detection via Outlier detection
    - Does not require WDAD assumption
    - Will fail if anomalies are not outliers (e.g., overlap with nominal density; tightly clustered anomaly density)
    - Will fail if nominal distribution has heavy tails

Outline

- Defining the Anomaly Detection Problem
- Benchmarking Current Algorithms for Unsupervised Anomaly Detection
- PAC Theory of Rare Pattern Anomaly Detection
- Incorporating Analyst Feedback
- Applications
  - Weather network anomaly detection
  - Open Category Classification

Benchmarking Study
[Andrew Emmott]



- Most AD papers only evaluate on a few datasets
- Often proprietary or very easy (e.g., KDD 1999)
- Research community needs a large and growing collection of public anomaly benchmarks

[Emmott, Das, Dietterich, Fern, Wong, 2013; KDD ODD-2013]
[Emmott, Das, Dietterich, Fern, Wong. 2016; arXiv 1503.01158v2]

Benchmarking Methodology

- Select 19 data sets from UC Irvine repository
- Choose one or more classes to be “anomalies”; the rest are “nominals”
- Manipulate
  - Relative frequency
  - Point difficulty
  - Irrelevant features
  - Clusteredness
- 20 replicates of each configuration
- Result: 25,685 Benchmark Datasets

Metrics

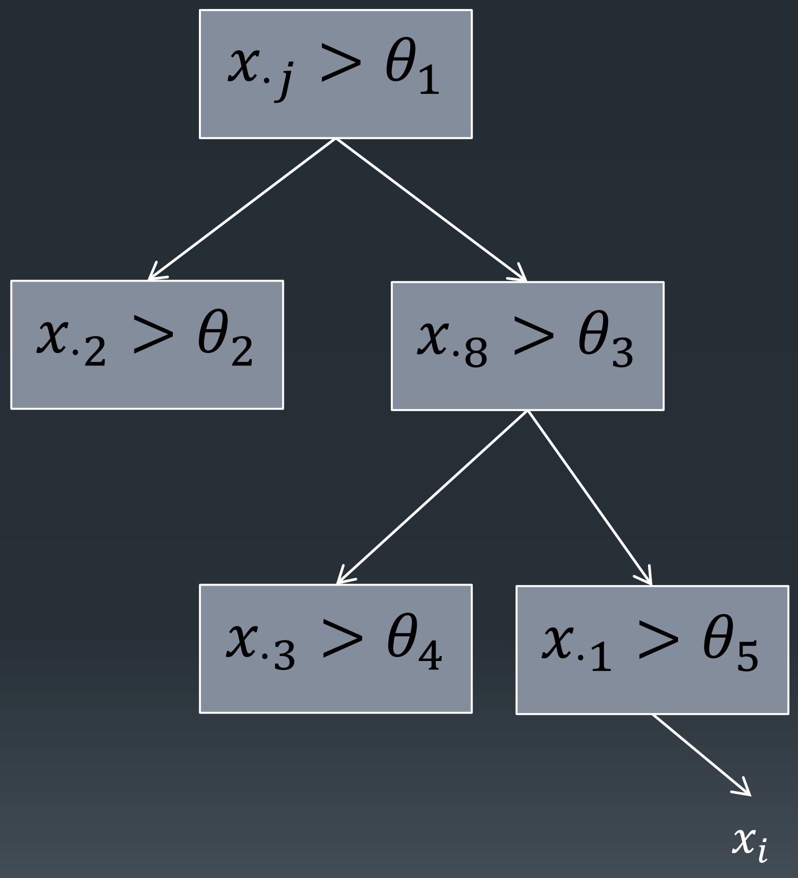
- AUC (Area Under ROC Curve)
  - ranking loss: probability that a randomly-chosen anomaly point is ranked above a randomly-chosen nominal point
  - transformed value: $\log \frac{AUC}{1-AUC}$
- AP (Average Precision)
  - area under the precision-recall curve
  - average of the precision computed at each ranked anomaly point
  - transformed value: $\log \frac{AP}{\mathbb{E}[AP]} = \log LIFT$

Algorithms

- Density-Based Approaches
  - RKDE: Robust Kernel Density Estimation (Kim & Scott, 2008)
  - EGMM: Ensemble Gaussian Mixture Models (our group)
- Quantile-Based Methods
  - OCSVM: One-class SVM (Schoelkopf, et al., 1999)
  - SVDD: Support Vector Data Description (Tax & Duin, 2004)
- Neighbor-Based Methods
  - LOF: Local Outlier Factor (Breunig, et al., 2000)
  - ABOD: kNN Angle-Based Outlier Detector (Kriegel, et al., 2008)
- Projection-Based Methods
  - IFOR: Isolation Forest (Liu, et al., 2008)
  - LODA: Lightweight Online Detector of Anomalies (Pevny, 2016)

Isolation Forest [Liu, Ting, Zhou, 2011]

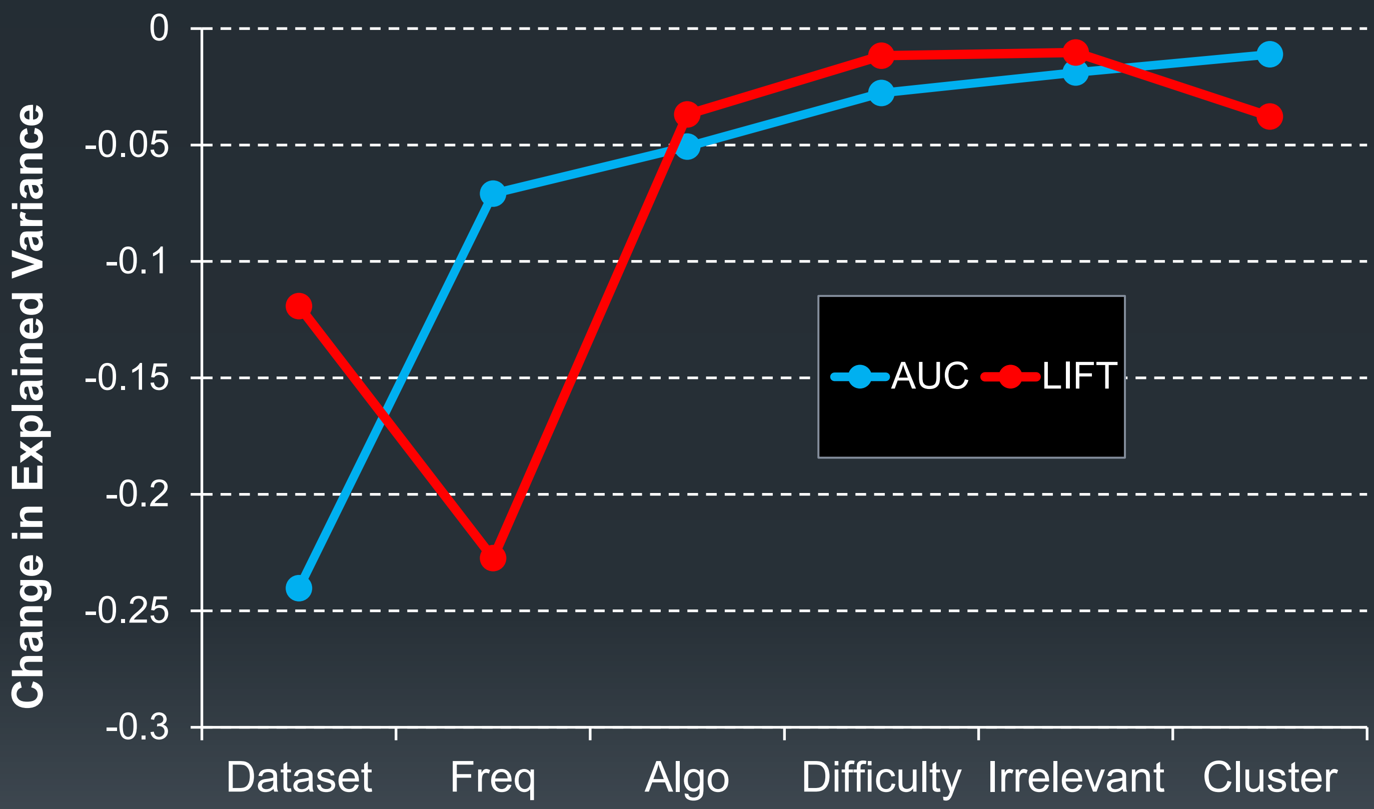
- Construct a fully random binary tree
  - choose attribute j at random
  - choose splitting threshold θ_1 uniformly from $[\min(x_{\cdot j}), \max(x_{\cdot j})]$
  - until every data point is in its own leaf
  - let $d(x_i)$ be the depth of point x_i
- repeat 100 times
  - let $\bar{d}(x_i)$ be the average depth of x_i
  - $score(x_i) = 2^{-\left(\frac{\bar{d}(x_i)}{r(x_i)}\right)}$
    - $r(x_i)$ is the expected depth



Analysis

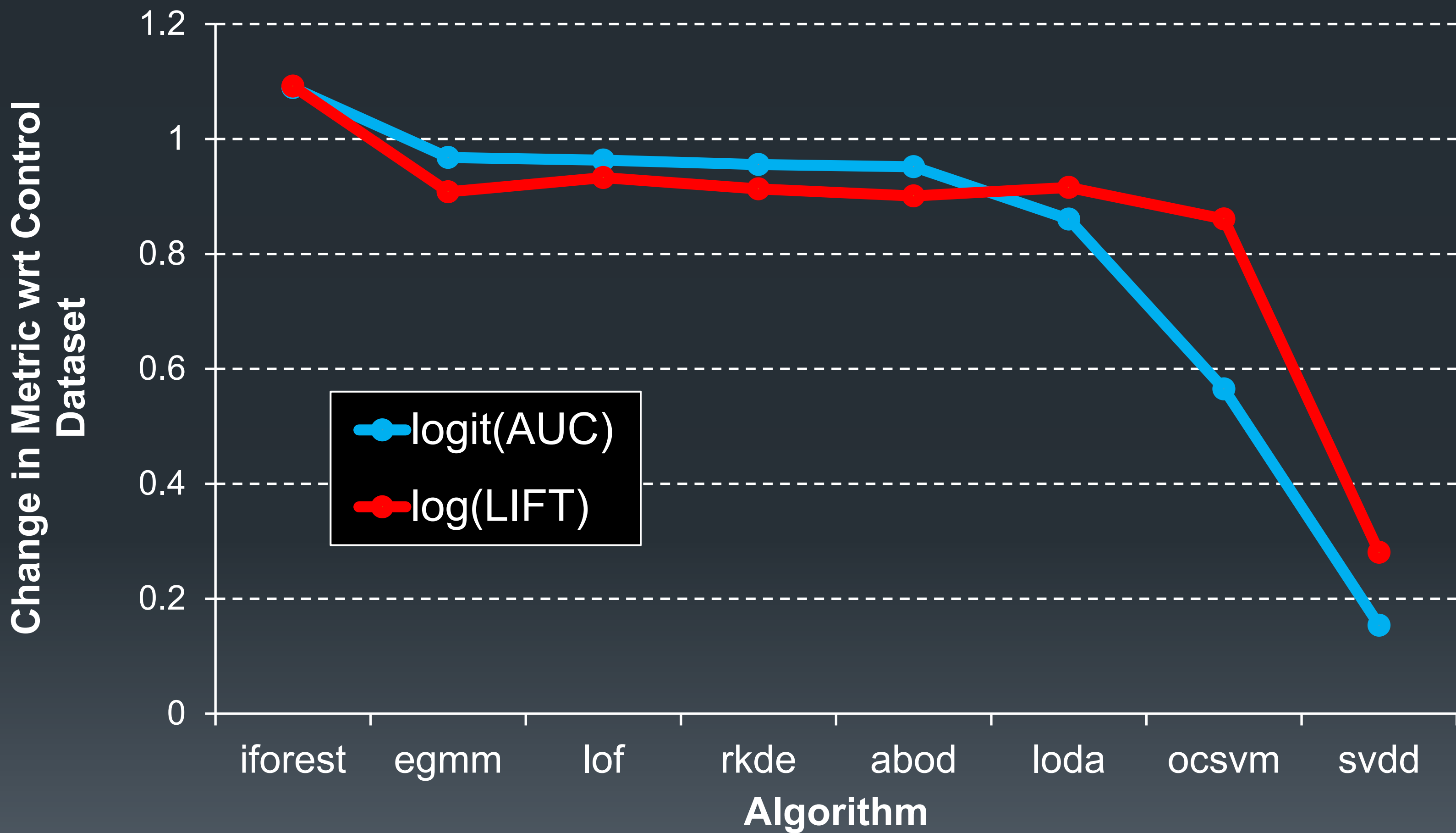
- Linear ANOVA
  - $metric \sim rf + pd + cl + ir + mset + algo$
    - rf: relative frequency
    - pd: point difficulty
    - cl: normalized clusteredness
    - ir: irrelevant features
    - mset: “Mother” set
    - algo: anomaly detection algorithm
- Validate the effect of each factor
- Assess the *algo* effect while controlling for all other factors

What Matters the Most?



- Problem and Relative Frequency!
- Choice of algorithm ranks third

Algorithm Comparison



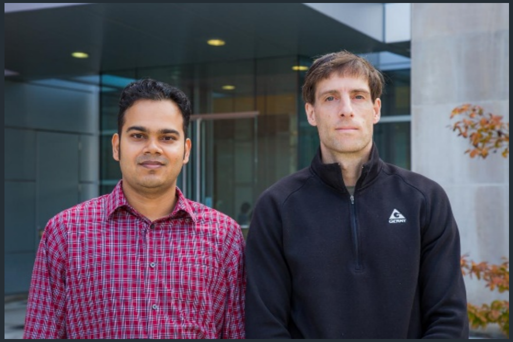
iForest Advantages

- Most robust to irrelevant features
  - for both AUC and LIFT
  - Hypothesis: effect of irrelevant features can be averaged out by computing a large forest
- Second most robust to clustered anomaly points
  - for AUC
  - Why?

Outline

- Defining the Anomaly Detection Problem
- Benchmarking Current Algorithms for Unsupervised Anomaly Detection
- PAC Theory of Rare Pattern Anomaly Detection
- Incorporating Analyst Feedback
- Applications
  - Weather network anomaly detection
  - Open Category Classification

Towards a Theory of Anomaly Detection [Siddiqui, et al.; UAI 2016]



- Existing theory on sample complexity
  - Density Estimation Methods:
    - Exponential in the dimension d
  - Quantile Methods (OCSVM and SVDD):
    - Polynomial sample complexity
- Experimentally, many anomaly detection algorithms learn very quickly (e.g., 500-2000 examples)
- New theory: Rare Pattern Anomaly Detection

Pattern Spaces

- A pattern $h: \Re^d \rightarrow \{0,1\}$ is an indicator function for a measurable region in the input space
  - Examples:
    - Half planes
    - Axis-parallel hyper-rectangles in $[-1,1]^d$
- A pattern space $\mathcal{H}$ is a set of patterns (countable or uncountable)

Rare and Common Patterns

- Let μ be a fixed measure over $\Re^d$
  - Typical choices:
    - uniform over $[-1, +1]^d$
    - standard Gaussian over $\Re^d$
- $\mu(h)$ is the measure of the pattern defined by h
- Let p be the "nominal" probability density defined on $\Re^d$ (or on some subset)
- $p(h)$ is the probability of pattern h
- A pattern h is τ-rare if

$$f(h) = \frac{p(h)}{\mu(h)} \leq \tau$$

- Otherwise it is τ-common

Rare and Common Points

- A point x is τ-rare if there exists a τ-rare h such that $h(x) = 1$
- Otherwise a point is τ-common

- Goal: An anomaly detection algorithm should output all τ-rare points and not output any τ-common points

PAC-RPAD

- Algorithm $\mathcal{A}$ is PAC-RPAD for
  - pattern space $\mathcal{H}$,
  - measure μ,
  - parameters τ, ϵ, δ

  if for any probability density p and any τ, $\mathcal{A}$ draws a sample from p and with probability $1-\delta$ detects all τ-rare points and rejects all $(\tau+\epsilon)$-commons in the sample

- ϵ allows the algorithm some margin for error
- If a point is between τ-rare and $(\tau+\epsilon)$-common, the algorithm can treat it arbitrarily
- Running time: polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, and $\frac{1}{\tau}$, and some measure of the complexity of $\mathcal{H}$

RarePatternDetect

- Draw a sample of size $N(\epsilon, \delta)$ from p
- Let $\hat{p}(h)$ be the fraction of sample points that satisfy h
- Let $\hat{f}(h) = \frac{\hat{p}(h)}{\mu(h)}$ be the estimated rareness of h
- A query point x_q is declared to be an anomaly if there exists a pattern $h \in \mathcal{H}$ such that $h(x_q) = 1$ and $\hat{f}(h) \leq \tau$.

Results

- Theorem 1: For any finite pattern space $\mathcal{H}$, RarePatternDetect is PAC-RPAD with sample complexity

$$N(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)\right)$$

- Theorem 2: For any pattern space $\mathcal{H}$ with finite VC dimension $\mathcal{V}_{\mathcal{H}}$, RarePatternDetect is PAC-RPAD with sample complexity

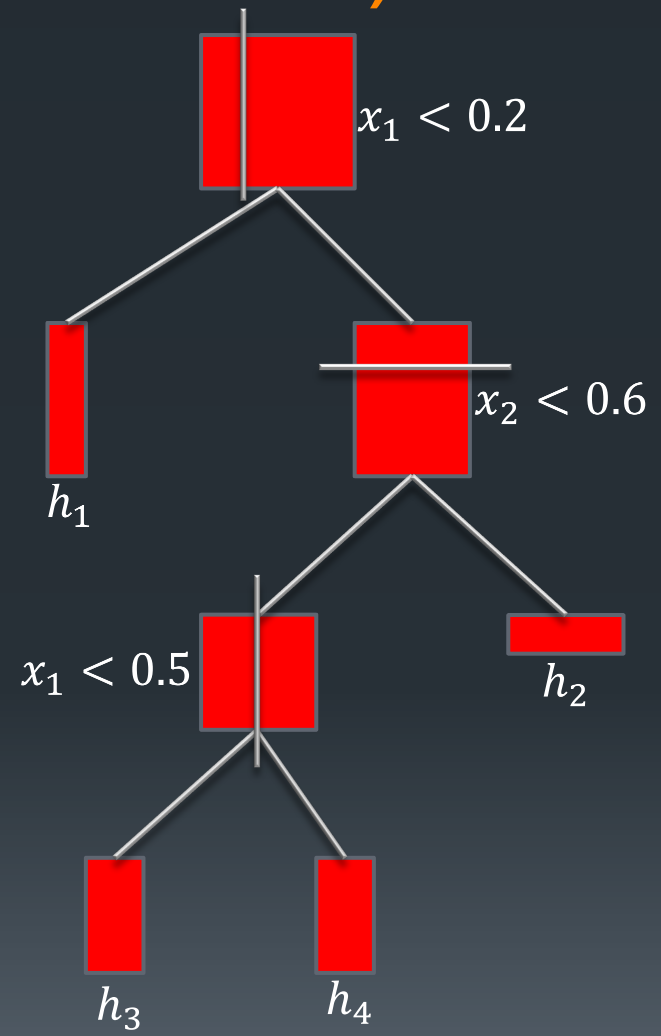
$$N(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2}\left(\mathcal{V}_{\mathcal{H}} \log\frac{1}{\epsilon^2} + \log\frac{1}{\delta}\right)\right)$$

Examples of PAC-RPAD $\mathcal{H}$

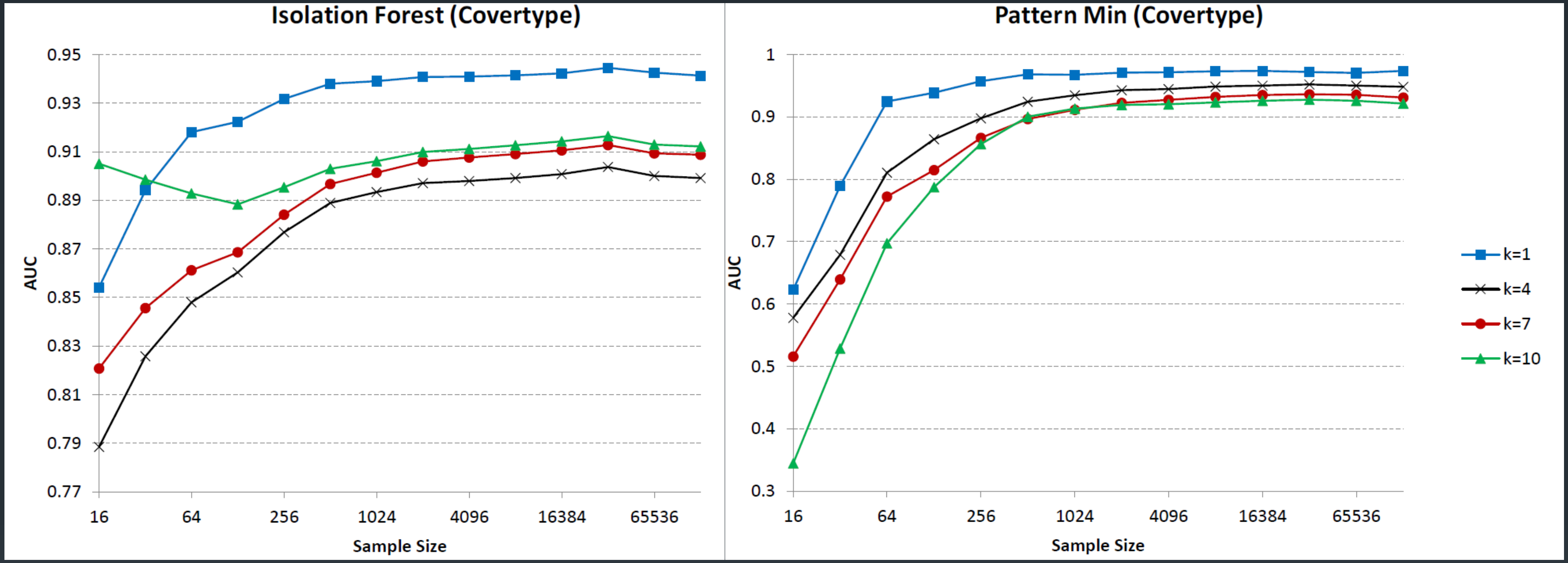
- Half spaces
- Axis-aligned hyper-rectangles (related to iForest leaves)
- Stripes (equivalent to LODA's histogram bins)
- Ellipsoids
- Ellipsoidal shells (difference of two ellipsoidal level sets)

Isolation RPAD (aka Pattern Min)

- Grow an isolation forest
  - Each tree is only grown to depth k
  - Each leaf defines a pattern h
  - μ is the volume (Lebesgue measure)
  - Compute $\hat{f}(h)$ for each leaf
- Details
  - Grow the tree using one sample
  - Estimate $\hat{f}$ using a second sample
  - Score query point(s)

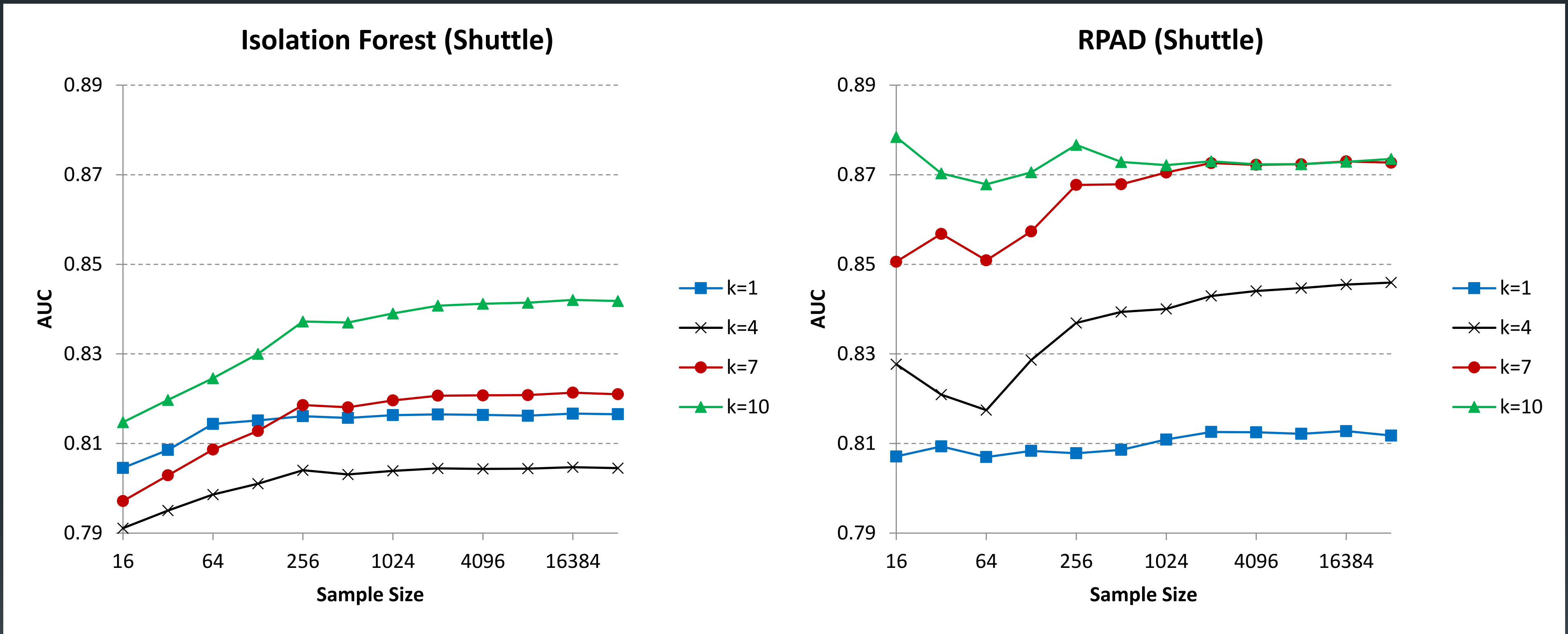


Results: Covertype



- PatternMin learns more slowly, but eventually beats IFOREST

Results: Shuttle



- PatternMin consistently beats iForest for $k > 1$

RPAD Conclusions

- The PAC-RPAD theory seems to capture the qualitative behavior of algorithms such as IFOREST
- It is easy to design practical RPAD algorithms
- Theory needs further work to handle sample-dependent pattern spaces $\mathcal{H}$

Outline

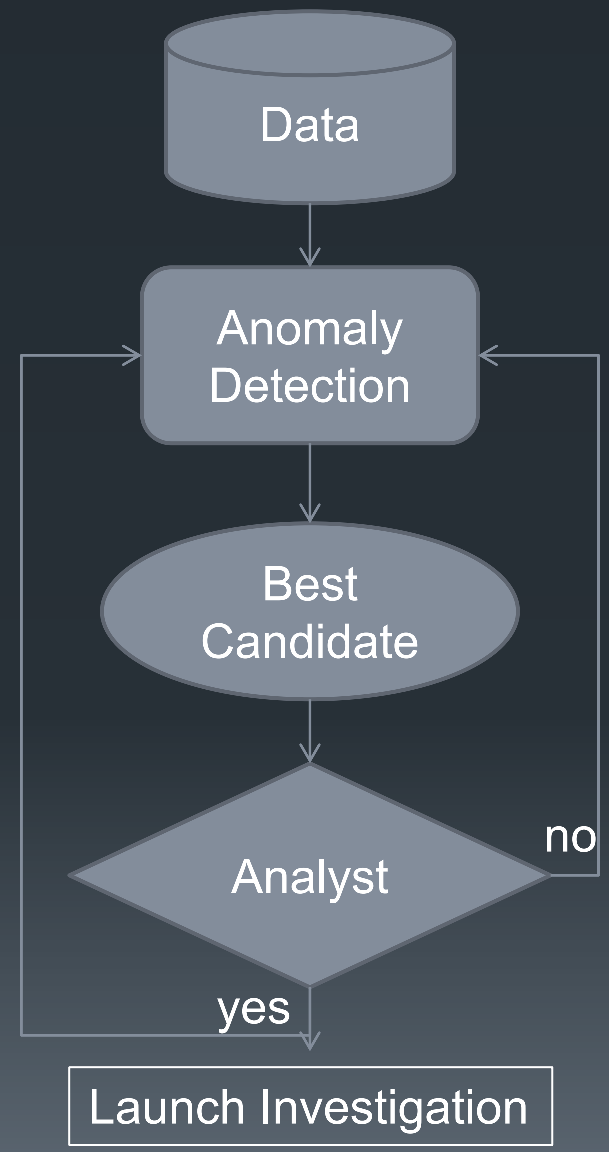
- Defining the Anomaly Detection Problem
- Benchmarking Current Algorithms for Unsupervised Anomaly Detection
- PAC Theory of Rare Pattern Anomaly Detection
- Incorporating Analyst Feedback
- Applications
  - Weather network anomaly detection
  - Open Category Classification

Outline

- Defining the Anomaly Detection Problem
- Benchmarking Current Algorithms for Unsupervised Anomaly Detection
- PAC Theory of Rare Pattern Anomaly Detection
- Incorporating Analyst Feedback
- Applications
  - Weather network anomaly detection
  - Open Category Classification

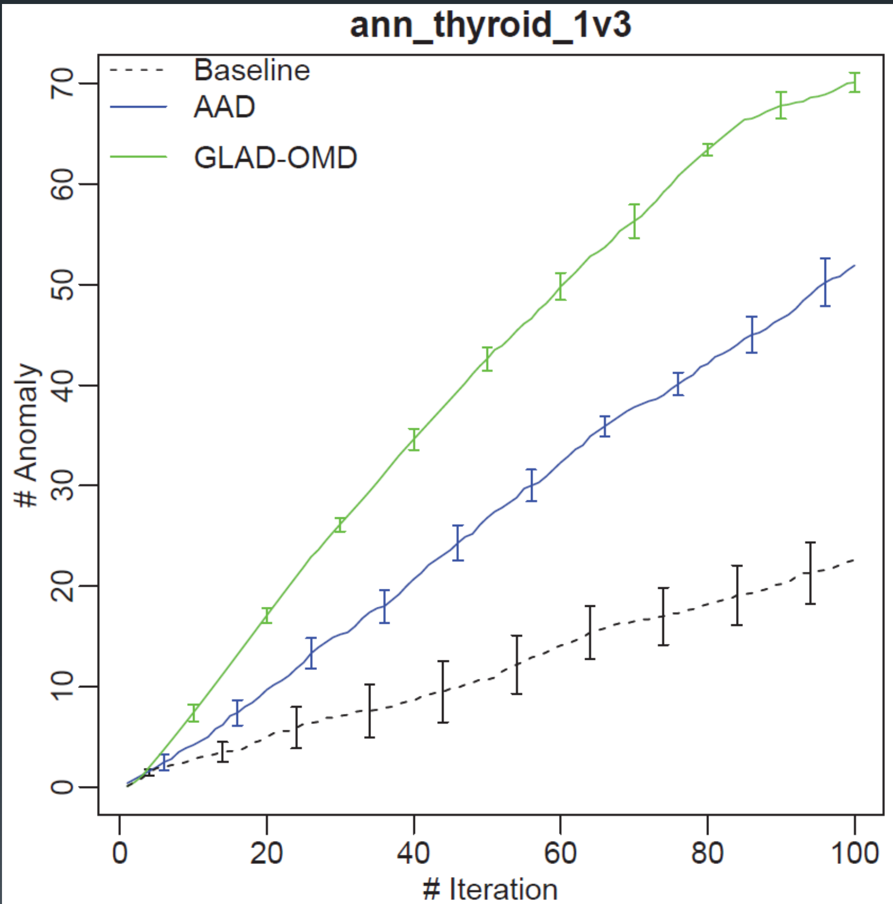
Incorporating Analyst Feedback

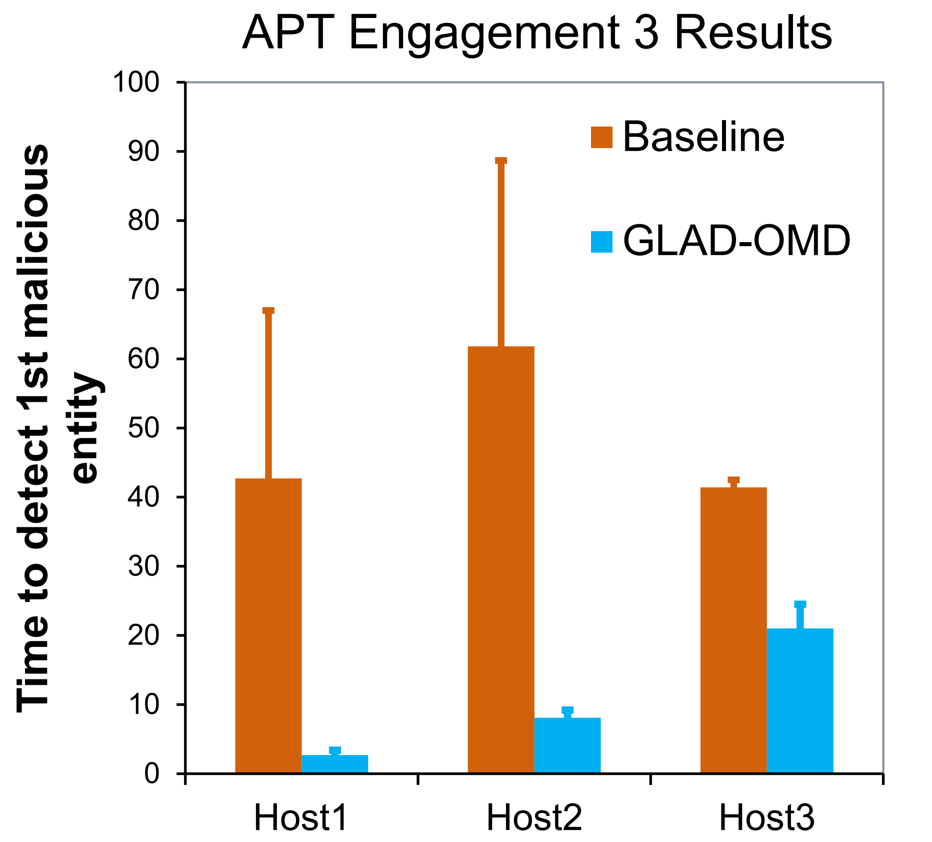
- Show top-ranked (unlabeled) candidate to the Analyst
- Analyst labels candidate
- Label is used to update the anomaly detector



[Das, et al, ICDM 2016]
[Siddiqui, et al., KDD 2018]

Analyst Feedback Yields Huge Improvements in Anomaly Discovery





Method

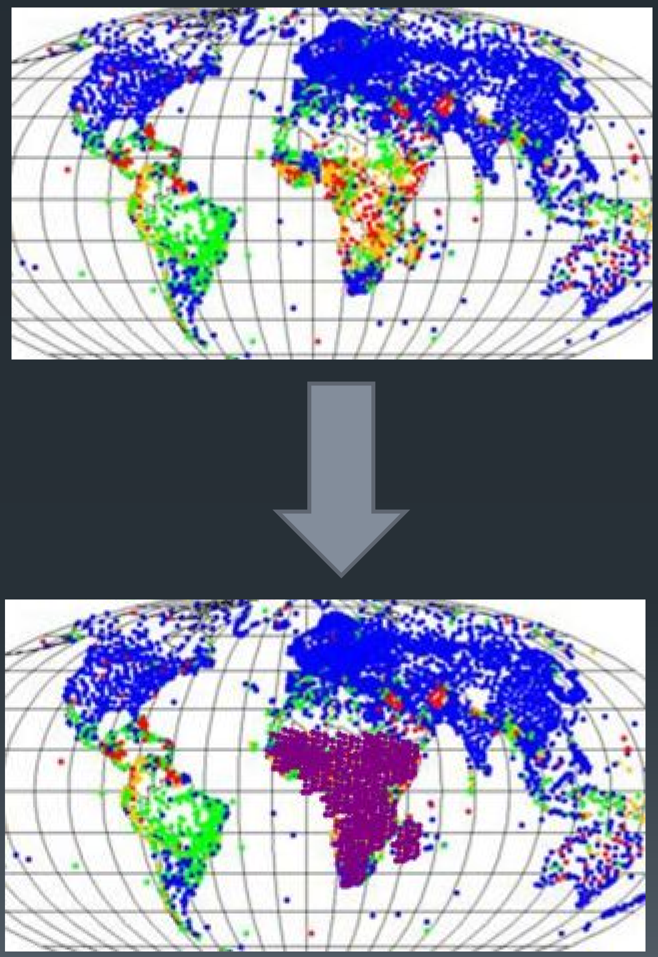
- Transform the Isolation Forest into a gigantic linear model
  - Each node in each tree becomes a Boolean feature that is 1 if x_q visits that node
  - Initial weight of each feature is 1.0, so that the weighted sum == sum of isolation depths in the forest
- Apply online convex optimization algorithms to learn from analyst feedback
  - Online Mirror Descent adjusts the weights to reduce the score of anomalies and increase the score of nominals

Outline

- Defining the Anomaly Detection Problem
- Benchmarking Current Algorithms for Unsupervised Anomaly Detection
- PAC Theory of Rare Pattern Anomaly Detection
- Incorporating Analyst Feedback
- Applications
  - Weather network anomaly detection
  - Open Category Classification

TAHMO: Trans-Africa Hydro-Meteorological Observatory

- Africa is very poorly sensed
  - Only a few dozen weather stations reliably report data to WMO (blue points in map)
  - Poor sensing ➔No crop insurance ➔Low agricultural productivity
  - Goal: Make Africa the best-sensed continent & improve agriculture
- Project TAHMO (tahmo.org)
  - TU-DELFT & Oregon State University
  - Design low-cost weather station
  - Deploy 20,000 such stations across Africa
  - Create data products (e.g., drought assessments, inundation estimates)
- Automated Data Quality Control
  - Detect broken sensors as anomalies



SENSOR-DX Architecture: Design and Training

- Define a set of views of the TAHMO data
  - A view involves 1 or more sensors from 1 or more stations over 1 or more time points
  - Each view defines a set of view tuples v
- Fit an anomaly detector to the view tuples
- Introduce a state variable s for each sensor at each station and time point
- Fit probabilistic models $P(A(v)|parents(v))$
- Hypothesis: It is easier to model the anomaly score distribution than it is to model the sensor readings

Example View

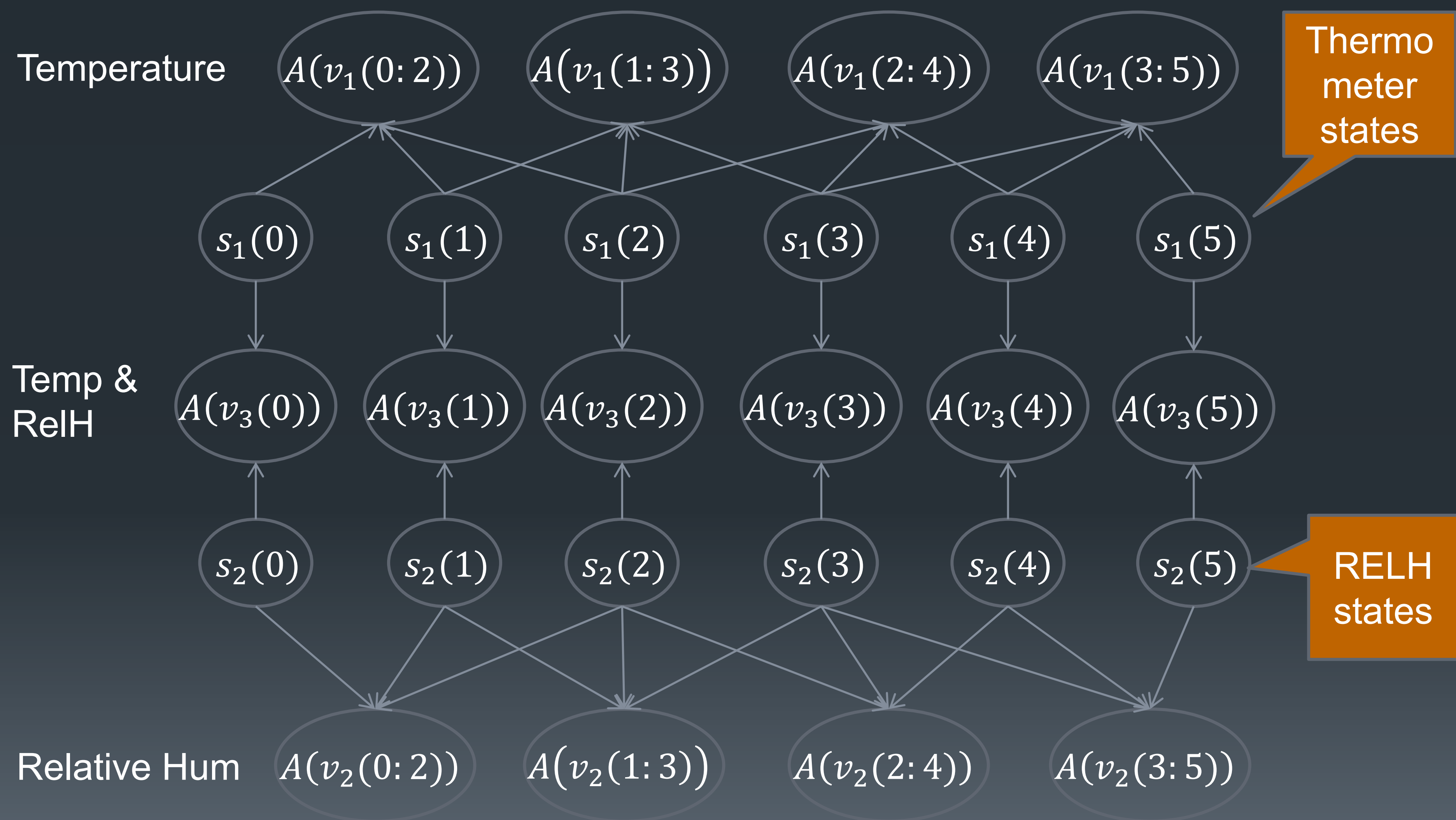
- Temperature T at times $t-2, t-1, t$

0	1	2	3	4
15	15.5	16.2	17	16.5

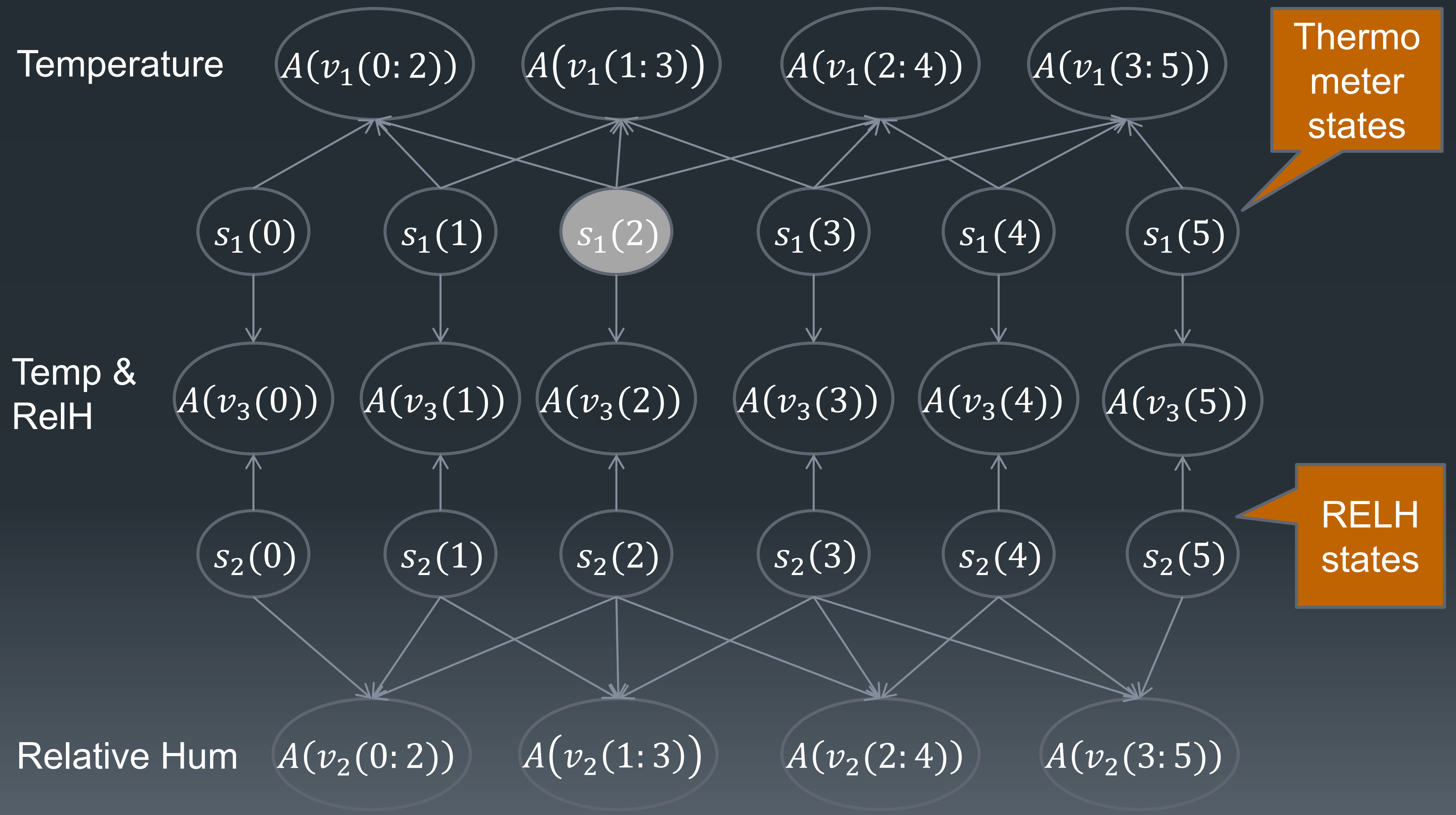
- View tuples

	$\boldsymbol{t-2}$	$\boldsymbol{t-1}$	$\boldsymbol{t}$
$v_1(0:2)$	15	15.5	16.2
$v_1(1:3)$	15.5	16.2	17
$v_1(2:4)$	16.2	17	16.5

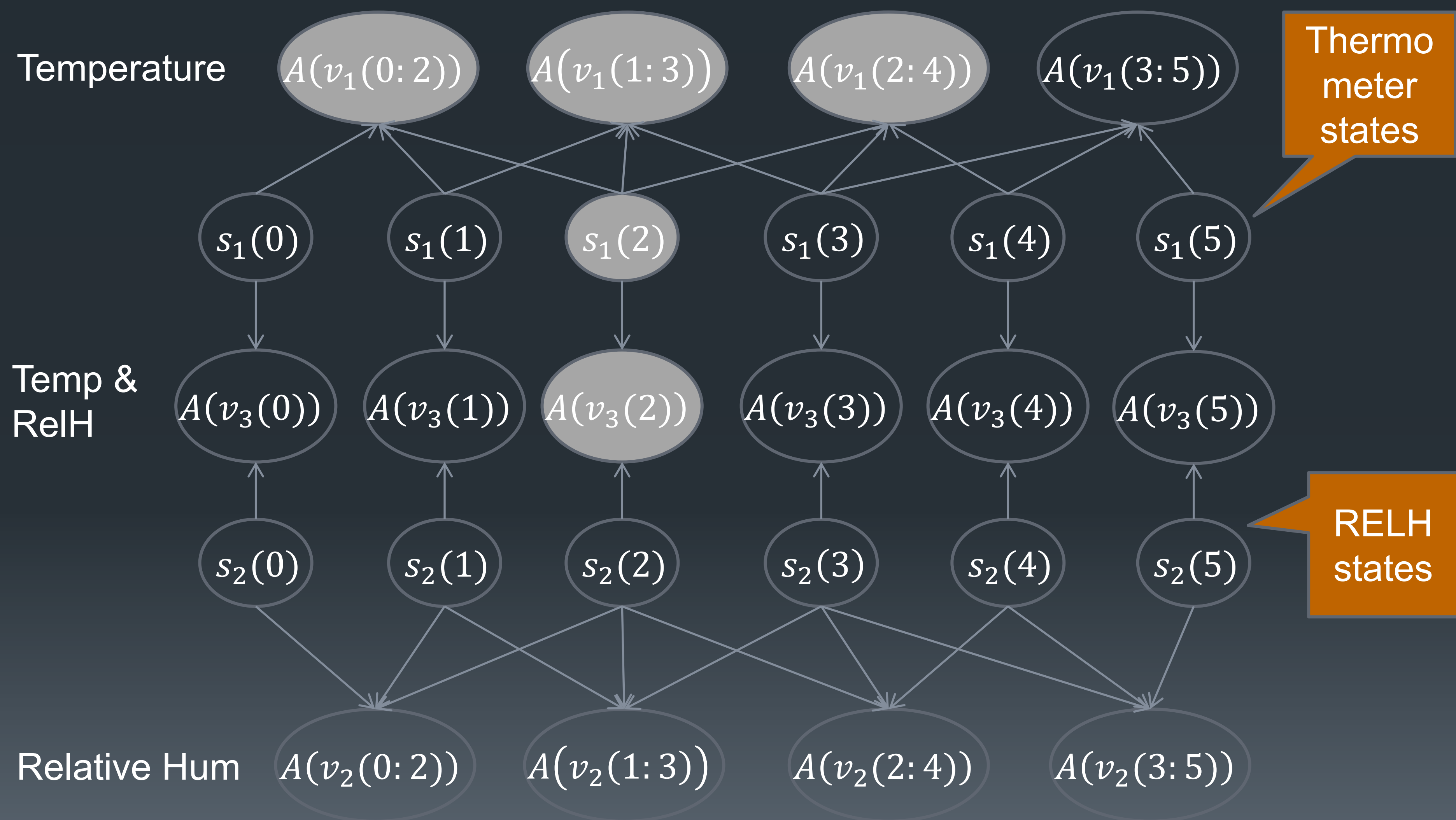
Diagnostic Model



Diagnostic Model



Diagnostic Model



Run Time Quality Control

- Assemble incoming data into view tuples
- Compute anomaly score for each view tuple
- Perform probabilistic inference to determine which sensor states best explain the observed anomaly scores:

$$\arg\max_{S} P(S|A(V))$$

Experimental Evaluation

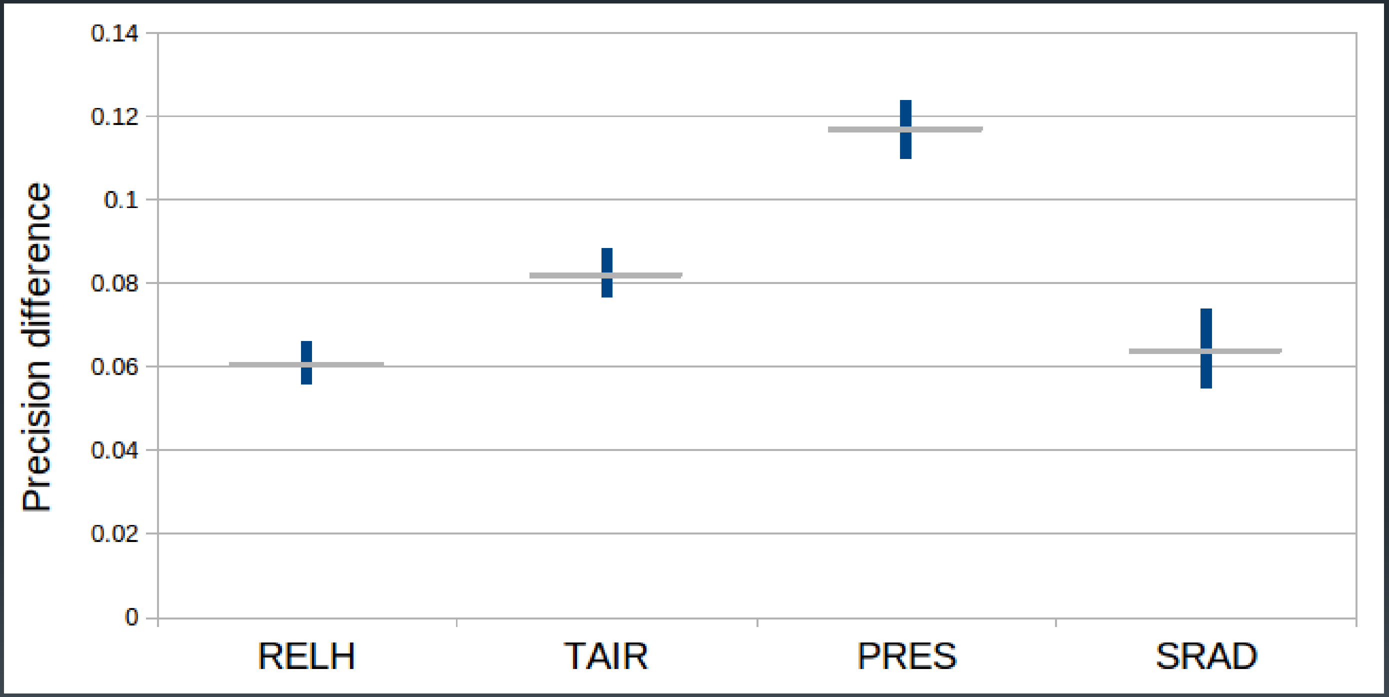
[Tadesse Zemicheal]

- Data: Oklahoma Mesonet
  - 1 year for training, 1 year for testing
  - 5 minute reporting interval; 20-day blocks
  - Hourly sensor state variable
  - Sensors:
    - Temperature (TAIR), relative humidity (RELH), atmospheric pressure (PRES), and Solar Radiation (SRAD)
  - Stations:
    - OKCE, OKCN, OKCW, NRMN
- Synthetic faults
  - spike noise, flatline, offset
- Isolation Forest
- Baseline:
  - Single sensor view
- SENSOR-DX:
  - Four views
- Metrics:
  - Precision and recall

View type	State/period	Total #views
Single sensor view	1	16
Same sensor two station view	2	24
Two sensor single station view	2	24
Single sensor three hour view	3	14
Total views per block		80

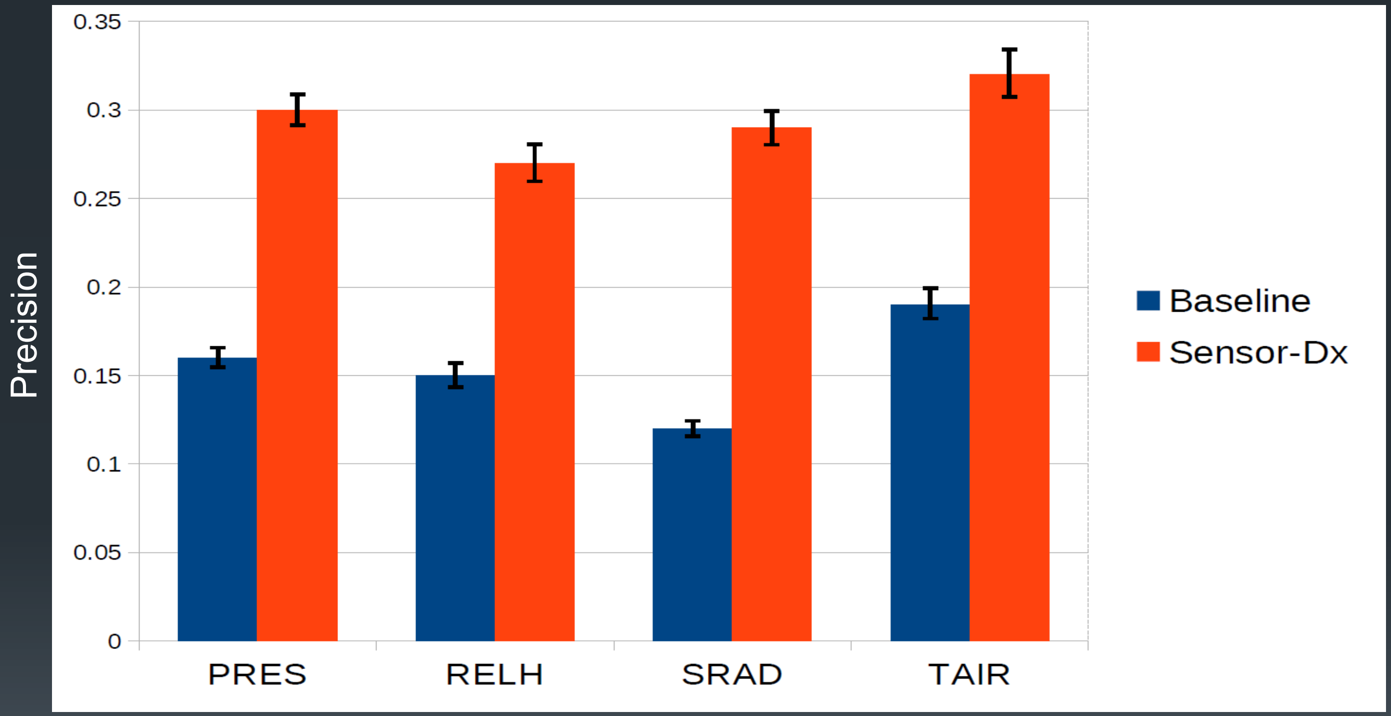
Result: SENSOR-DX improves precision

Difference in precision of multi-view method versus single-view baseline



95% two-sided paired differences bootstrap confidence intervals

Precision at Matching Recall Level



95% confidence intervals

Sensor-DX improves precision, but the false alarm rate will still be quite high

Status

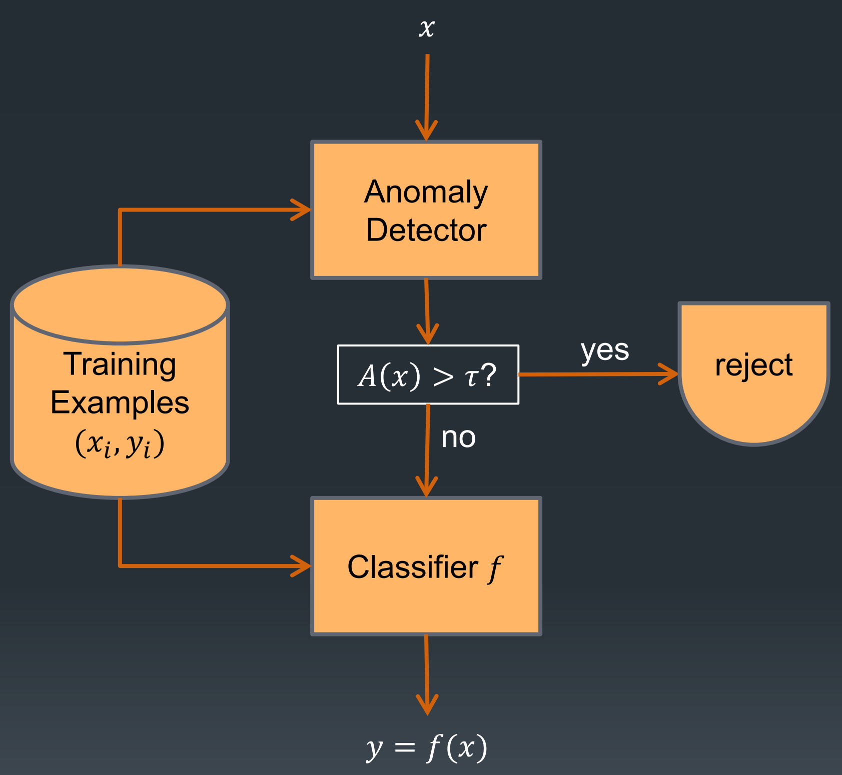
- Deployment on TAHMO network is in progress
- Integrated with
  - Network Manager dashboard
  - Trouble ticket system

Open Category Classification

[Liu, Garrepalli, Fern, Dietterich, ICML 2018]

- Training data for classes {1,…,K}
- Test data may contain queries corresponding to additional classes
- Can we detect them?

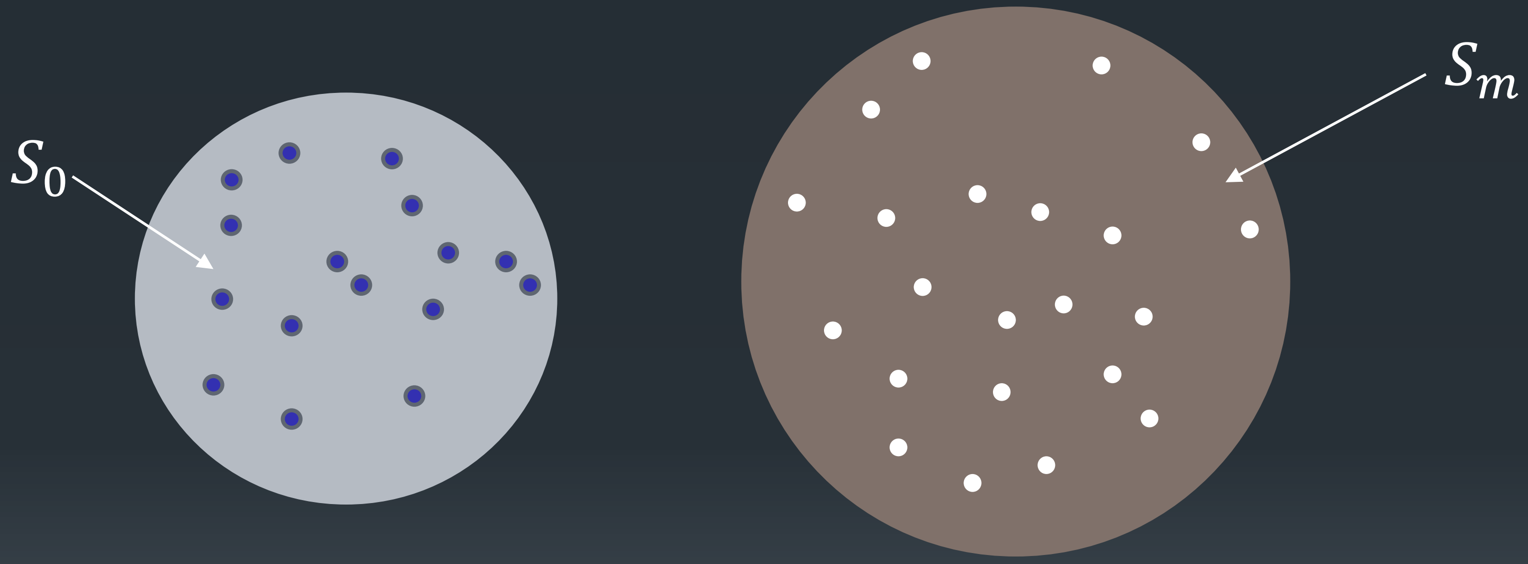
Prediction with Anomaly Detection



Training Data

P_0
Nominal Distribution

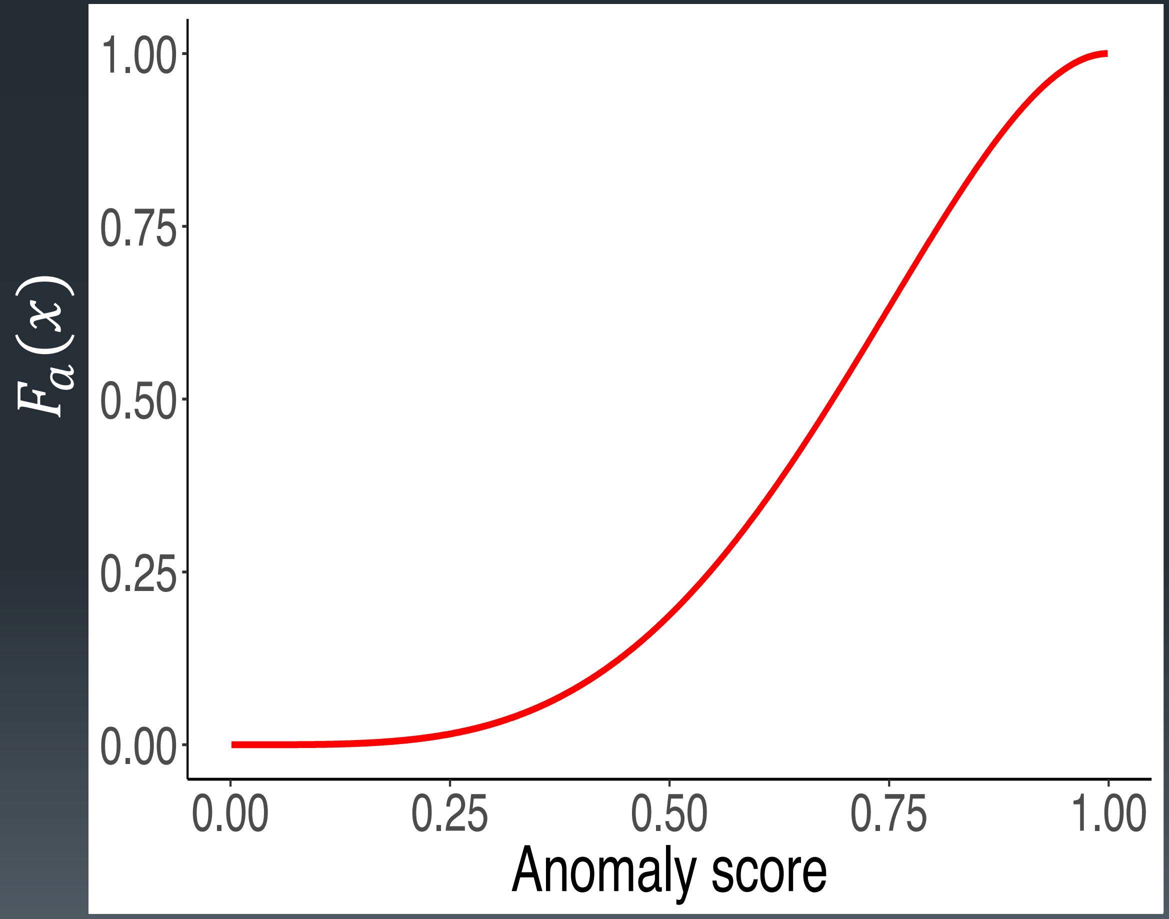
P_m
Mixture Distribution



Proportion of Aliens = α

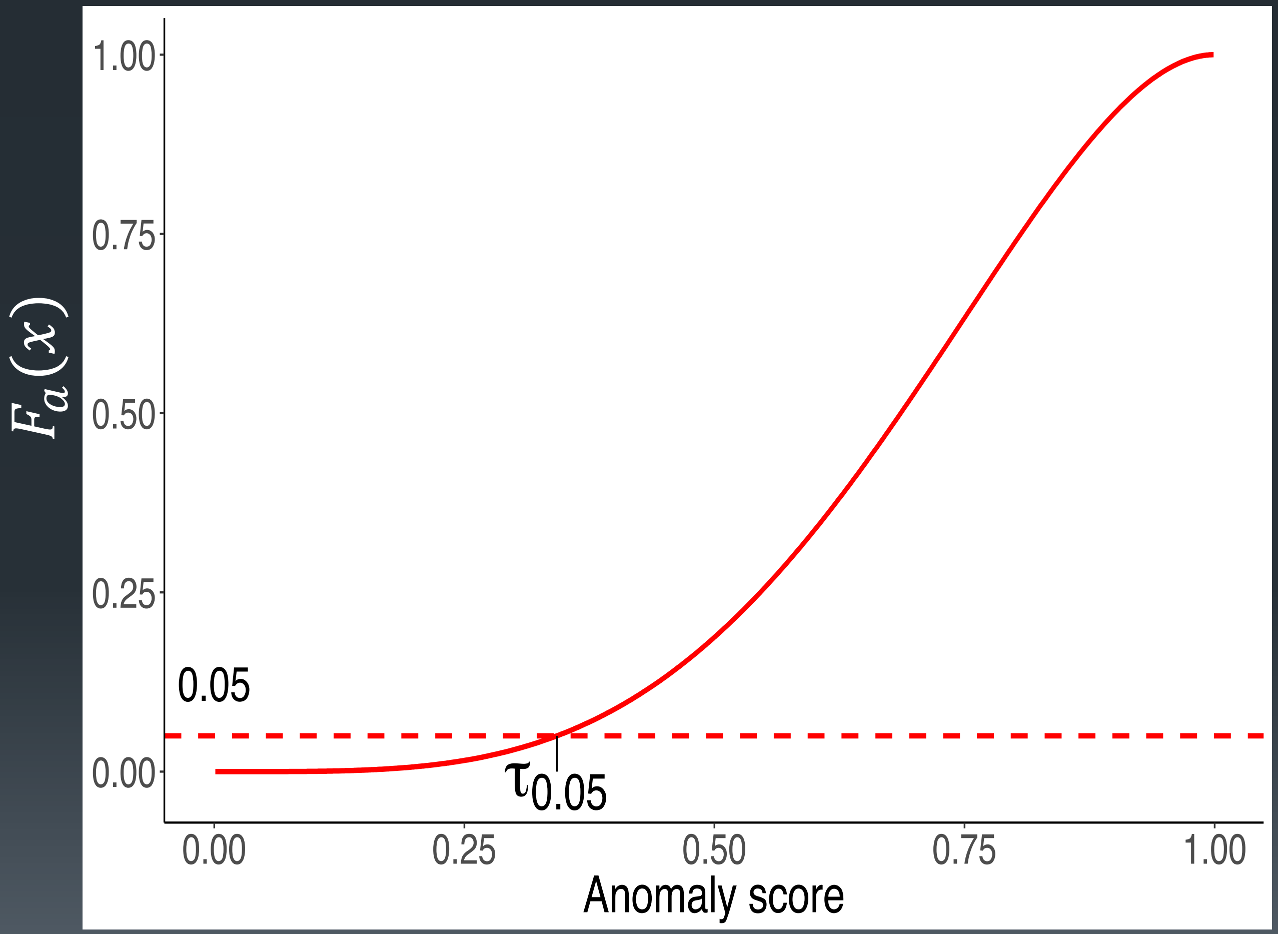
$$P_m = (1 - \alpha)P_0 + \alpha P_a$$

CDF of Alien Anomaly Scores: F_a



Want to have recall $= 1 - q$

Choosing τ for target quantile q



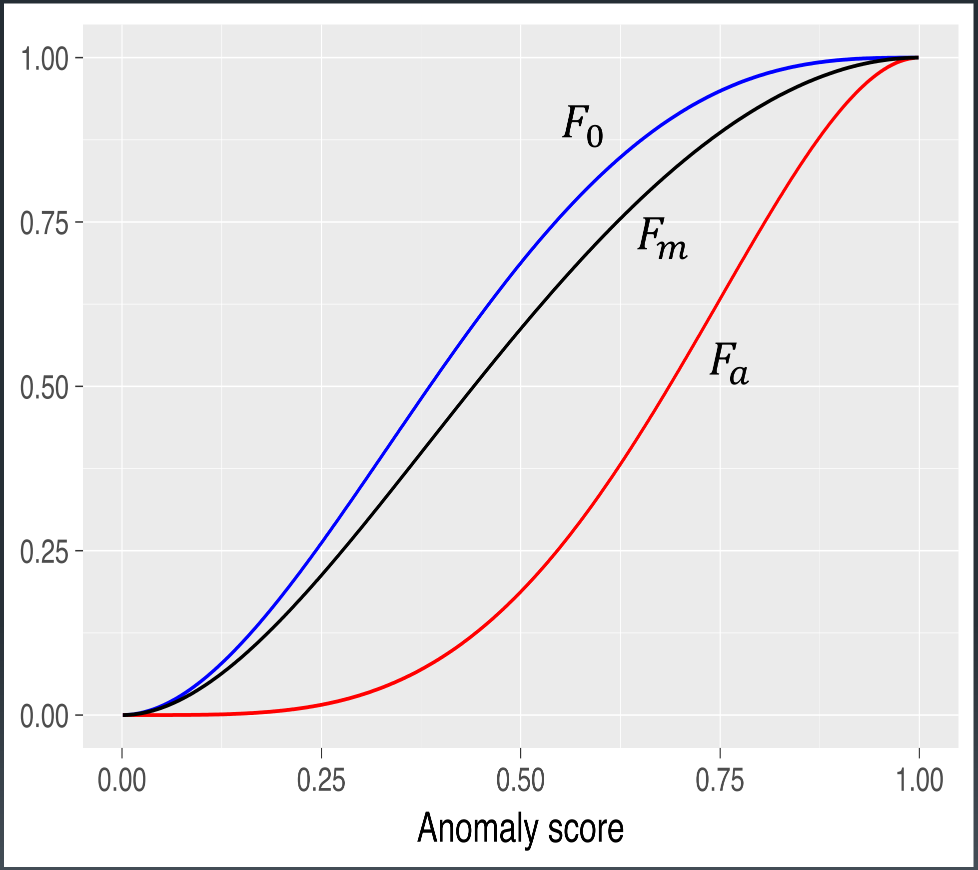
$$q = 0.05$$

$$P_m = (1-\alpha)P_0 + \alpha P_a$$

implies that

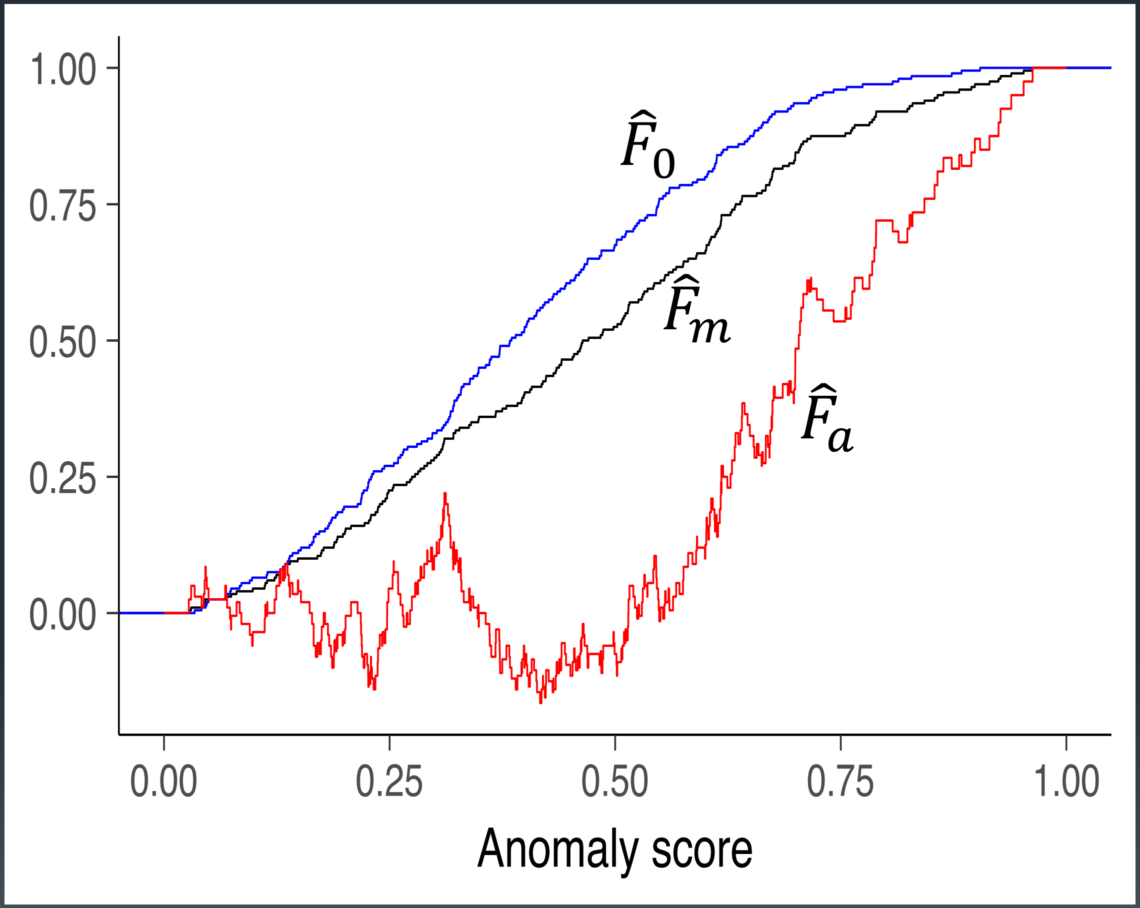
$$F_m(x) = (1-\alpha)F_0(x) + \alpha F_a(x)$$

CDFs of Nominal, Mixture, and Alien Anomaly Scores



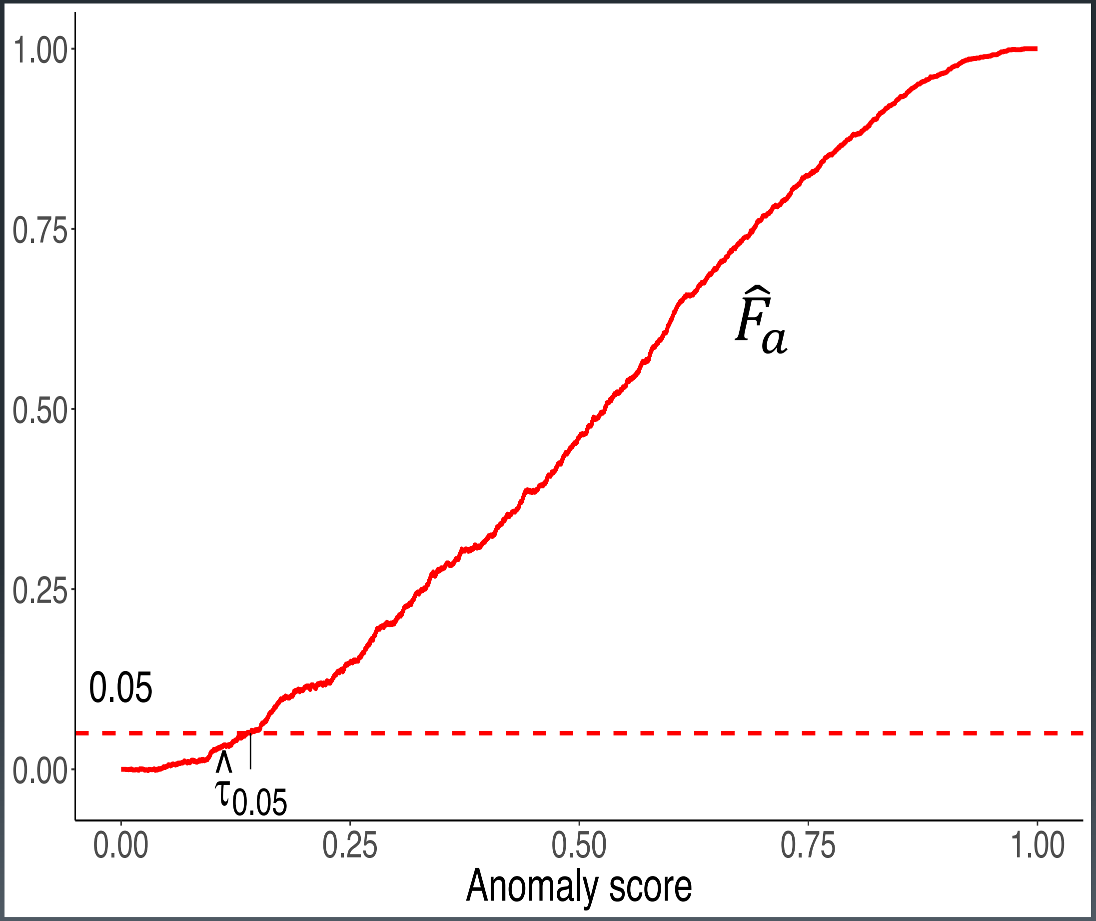
$$F_a(x) = \frac{F_m(x) - (1-\alpha)F_0(x)}{\alpha}$$

What We Have Are Empirical CDFs



$$\widehat{F}_a(x) = \frac{\widehat{F}_m(x) - (1-\alpha)\widehat{F}_0(x)}{\alpha}$$

We Use the Empirical Estimate $\hat{\tau}_{0.05}$



EstimateTau(S_0, S_m, q, α)

1: Anomaly scores of S_0: $x_1, x_2, \cdots, x_k$

2: Anomaly scores of S_m: $y_1, y_2, \cdots, y_m$

3: Compute empirical CDFs $\hat{F}_0$ and $\hat{F}_m$.

4: Calculate $\hat{F}_a$ using

$$\hat{F}_a(x) = \frac{\hat{F}_m(x) - (1-\alpha)\hat{F}_0(x)}{\alpha}.$$

5: Output detection threshold

$$\hat{\tau}_q = \max_{u \in S} \hat{F}_a(u) \leq q,$$

where $S = \{x_1, x_2, \cdots, x_k, y_1, y_2, \cdots, y_m\}$.

Theoretical Guarantee

[Liu, Garrepalli, Fern, Dietterich, ICML 2018]

- Theorem: If

$$n > \frac{1}{2} \ln \frac{2}{1 - \sqrt{1 - \delta}} \left(\frac{1}{\epsilon}\right)^2 \left(\frac{2 - \alpha}{\alpha}\right)^2$$

then with probability $1 - \delta$ the alien detection rate will be at least $1 - (q + \epsilon)$

Proof based on Massart (1990) concentration bound for empirical CDFs

Summary

- Outlier Detection can perform unsupervised or clean anomaly detection when the relative frequency of anomalies, α is small
- Algorithm Benchmarking
  - The Isolation Forest is a robust, high-performing algorithm
  - The OCSVM and SVDD methods do not perform well on AUC and AP. Why not?
  - The other methods (ABOD, LODA, LOF, EGMM, RKDE) are very similar to each other
- PAC-RPAD theory may account for the rapid learning of many anomaly detection algorithms
- Expert Feedback can double or triple the efficiency of detecting anomalies
- Anomaly detection can help find broken IoT sensors
- Anomaly detection can provide guarantees for open category detection

Acknowledgements

- Partially supported by
  - DARPA Contract W911NF-11-C-0088
  - DARPA Contract FA8650-15-C-7557
  - DARPA Contract FA8750-19-C-0092
  - US NSF Grants 1514550 & 1521687
  - FLI program FLI-RFP-AI1, grant number 2015-145014
  - Gift from Huawei, Inc.