Finite Sample Complexity of Rare Pattern Anomaly Detection

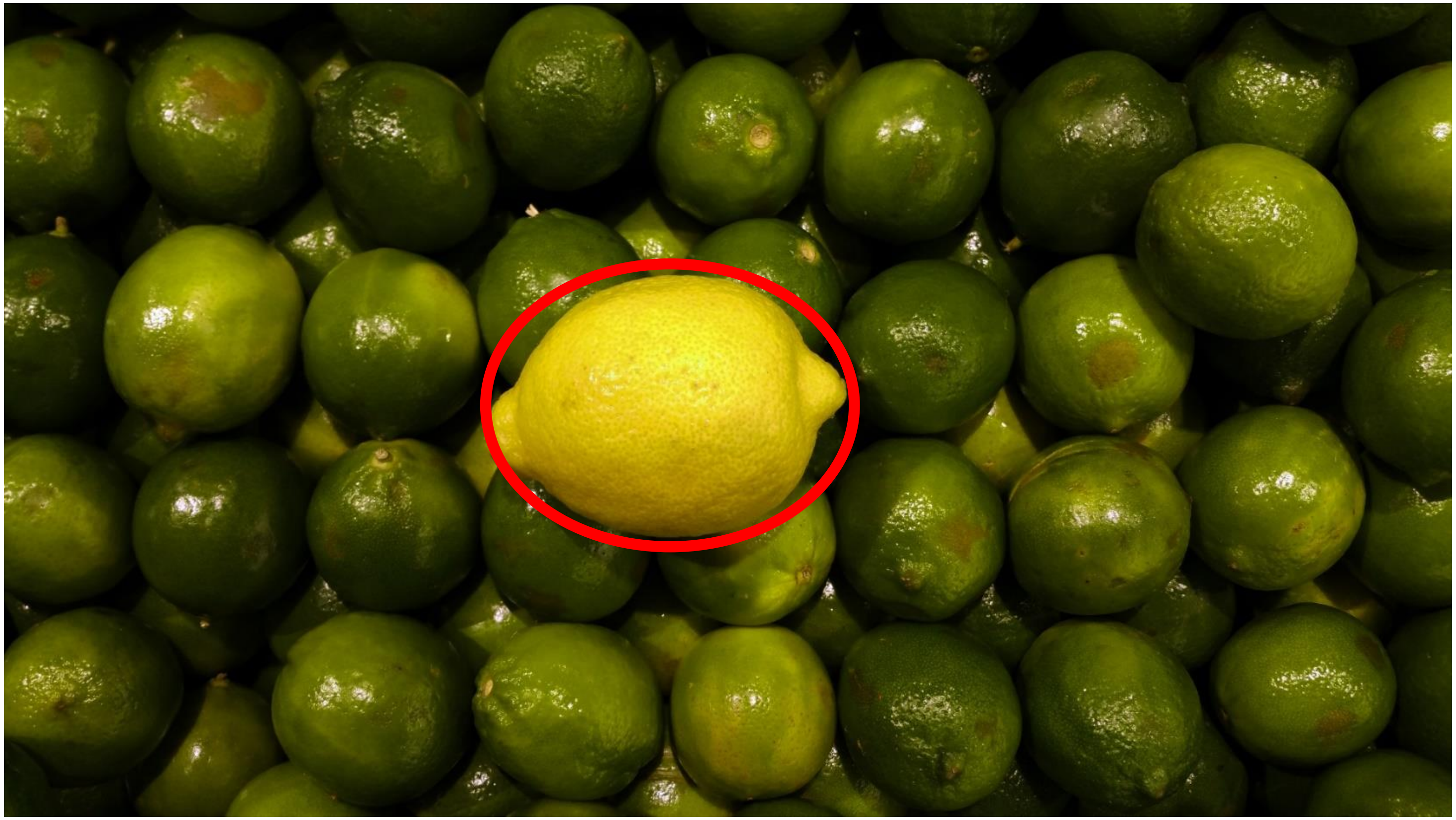
Md Amran Siddiqui, Alan Fern, Thomas G. Dietterich and Shubhomoy Das

School of EECS

Oregon State University

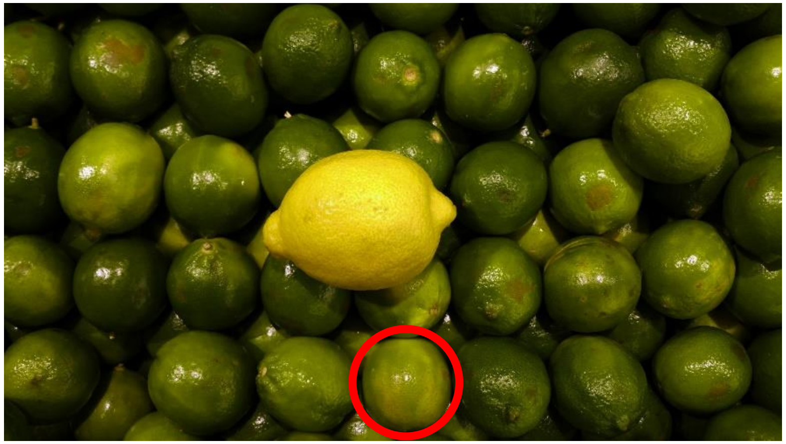
Anomaly Detection

- **Goal:** Identify rare or strange objects



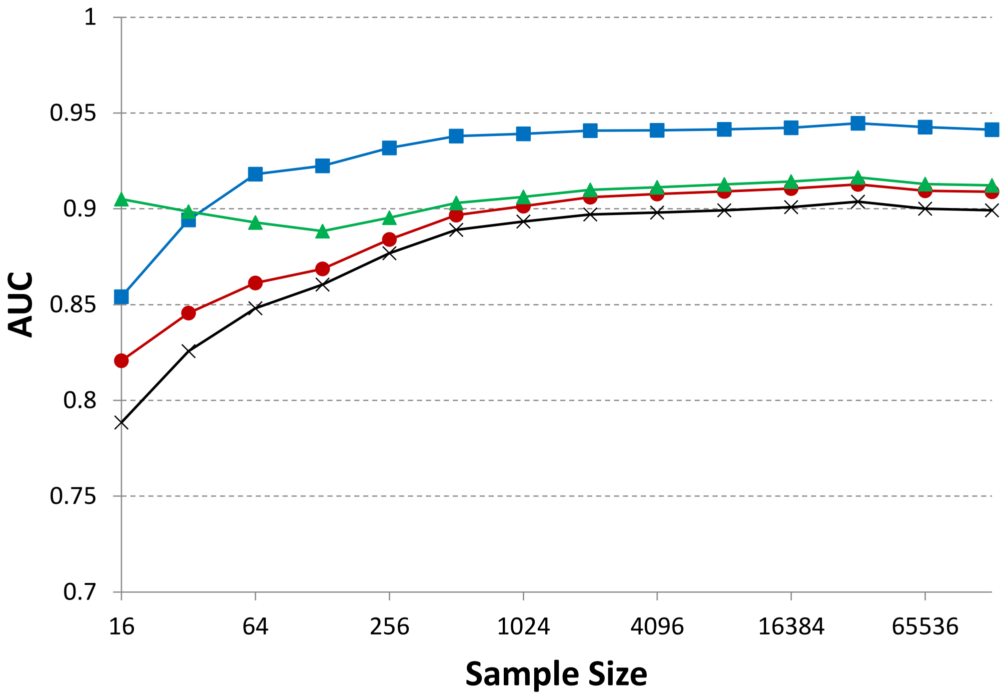
Challenges

- Every object is unusual in some ways!



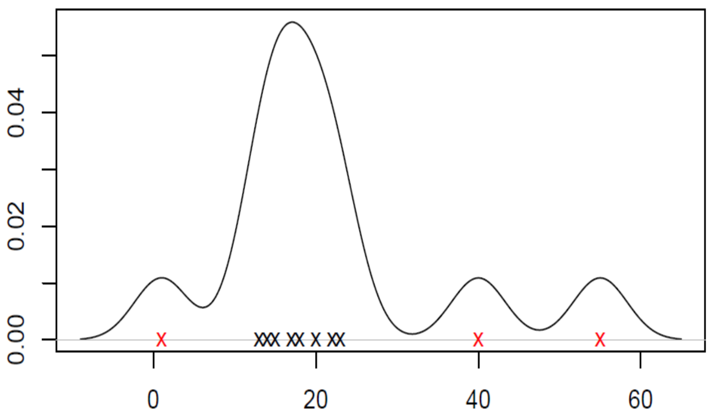
- Anomaly detection in high-dimension seems impossible ☹

State-of-the-art



- Often perform very well with a surprisingly small number of examples ☺
- Performance depends on:
    - ✓Sample Complexity
    - ✓Notion of Anomaly

Notion of Anomaly



Statistical

Semantic

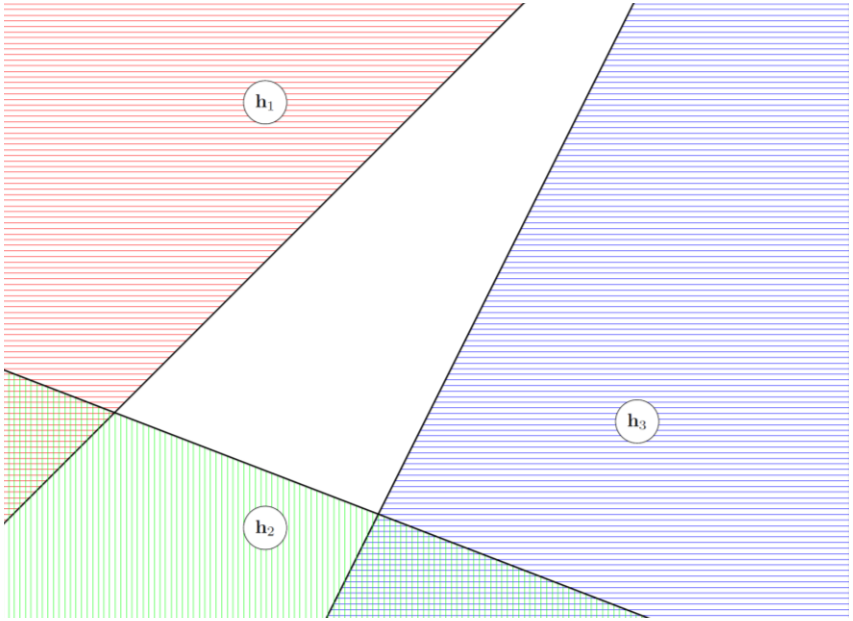
- Algorithm/representation specific
- Example: density of a point

- Application specific
- Example: A threat in security

Motivation

Many state-of-the-art algorithms [Chen et al. 2015, Liu et al. 2008, Wu et al. 2014, Tomas Pevny 2016] exhibit the following steps:

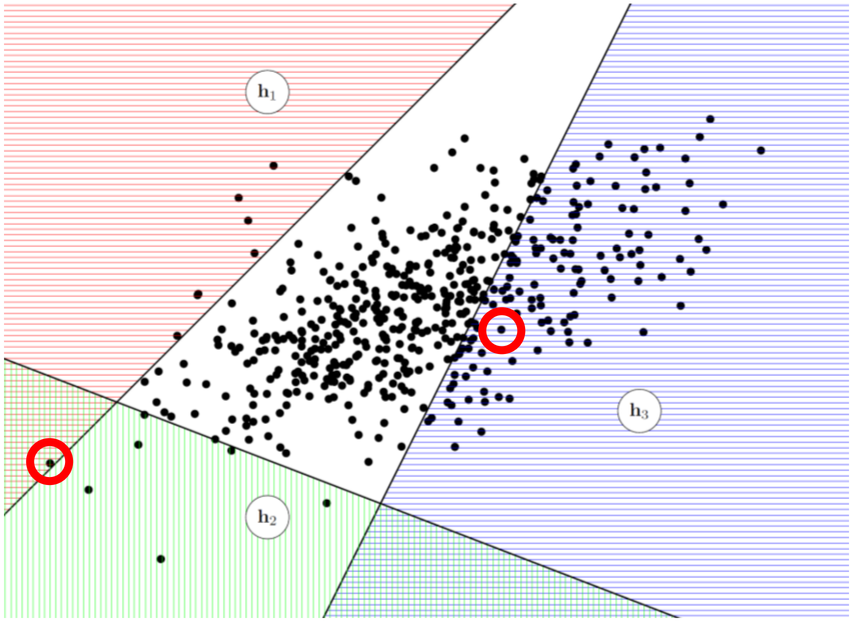
1. Choose a "pattern space" (analogous to hypothesis space)



Motivation

Many state-of-the-art algorithms [Chen et al. 2015, Liu et al. 2008, Wu et al. 2014, Tomas Pevny 2016] exhibit the following steps:

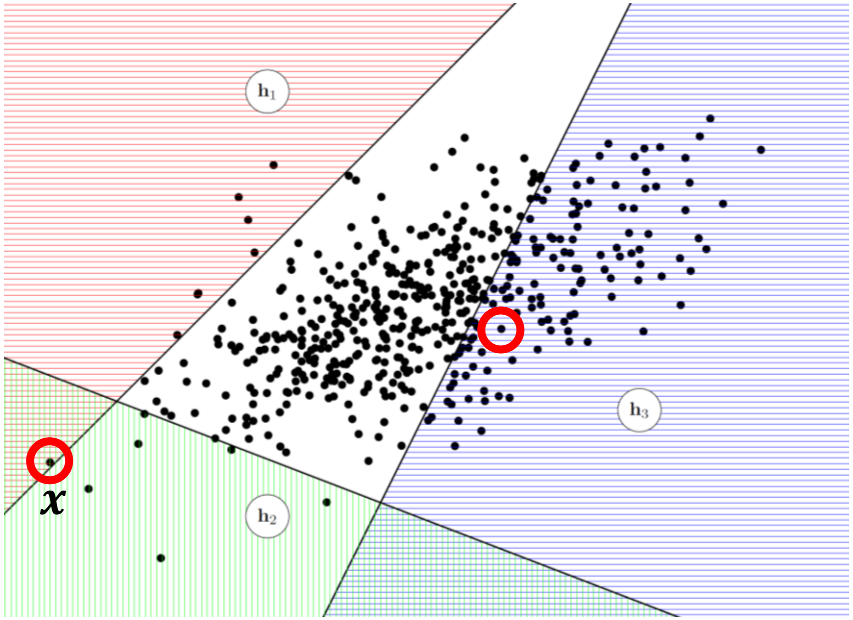
1. Choose a "pattern space" (analogous to hypothesis space)
2. Monitor the empirical frequency of the patterns
3. Compute anomaly score based on the frequencies



Rare Pattern Anomaly Detection (RPAD)

Rare Pattern Anomaly Detection (RPAD)

$\mathcal{H}$	Pattern space, $\{h_1, h_2, h_3\}$
$\mathcal{H}[x]$	Set of patterns that contain x, $\{h_1, h_2\}$
$f(h)$	Frequency of a pattern h, $f(h_1) < f(h_3)$
τ	Detection threshold



A point x is	***τ-outlier*** : If $\mathcal{H}[x]$ contains an h with $f(h) \le \tau$
	τ-common : *Otherwise*

Learning Protocol

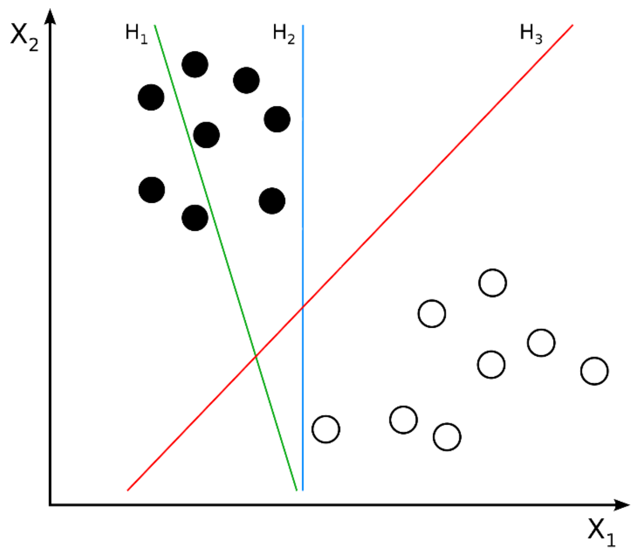
- Assumption: Input is generated from a distribution $\boldsymbol{\mathcal{P}}$ i.e. $\boldsymbol{x} \sim \boldsymbol{\mathcal{P}}$
- Let, $\boldsymbol{\mathcal{A}}$ be an anomaly detection algorithm
- $\boldsymbol{\mathcal{A}}$ can draw a training set $\boldsymbol{\mathcal{D}}$ of any size $\boldsymbol{\mathcal{N}}$ from $\boldsymbol{\mathcal{P}}$
- Given a new point x : $\boldsymbol{\mathcal{A}}$ has to either "detect" or "reject"
- Ideally, $\boldsymbol{\mathcal{A}}$ is "correct":

  if $\boldsymbol{\mathcal{A}}$ "detects" all τ-outliers and

  "rejects" all τ-commons

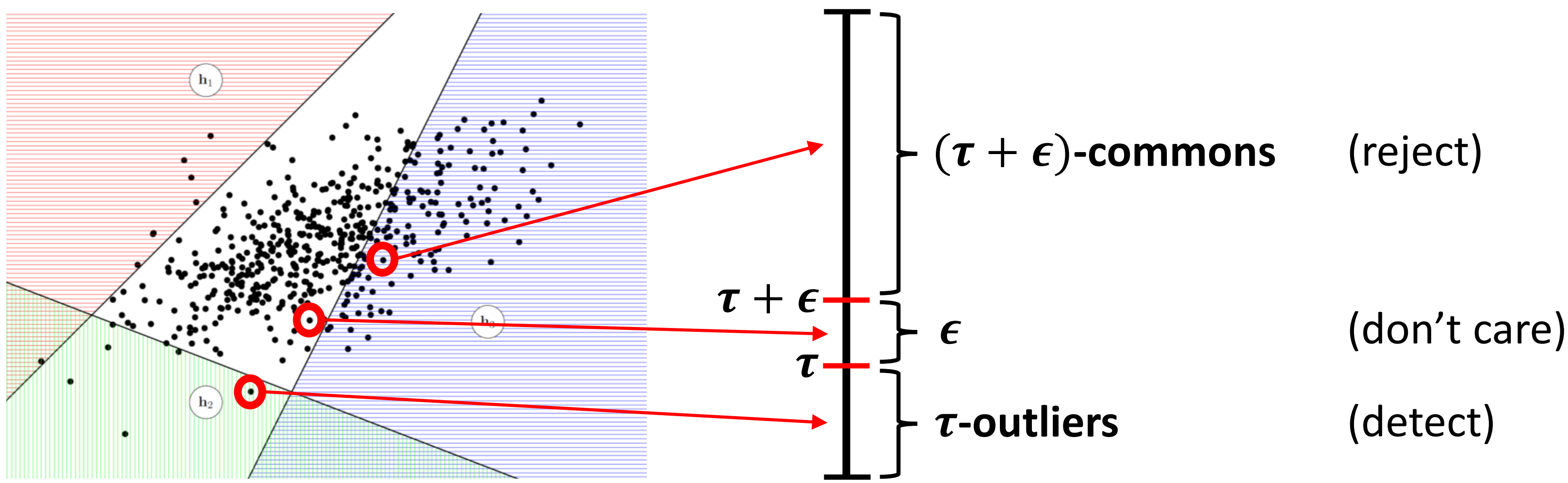
Supervised PAC Learning Framework

- Consider a hypothesis space $\mathcal{H}$ i.e. set of linear separators
- **Goal:** Learn a hypothesis that will make small error with high probability



- Sample complexity is related to the complexity of $\mathcal{H}$: VC-dimension
- What is analogous for Anomaly Detection?

PAC-RPAD Framework



Definition 1. (**PAC-RPAD**) Detection algorithm $\boldsymbol{\mathcal{A}}$ is PAC-RPAD if for any $\boldsymbol{\mathcal{P}}$ and any $\boldsymbol{\tau}$, with probability at least $\mathbf{1} - \boldsymbol{\delta}$ (over draws of $\boldsymbol{\mathcal{D}}$), $\boldsymbol{\mathcal{A}}$ detects all $\boldsymbol{\tau}$-outliers and rejects all $(\boldsymbol{\tau} + \boldsymbol{\epsilon})$-commons.

Sample efficient : if $\mathcal{A}$ draws polynomial (in d, $\frac{1}{\delta}$ and $\frac{1}{\epsilon}$) number of training examples from $\mathcal{P}$

RarePatternDetect Algorithm

Input:

δ: Probability tolerance

ϵ: Error tolerance

τ: Detection threshold

1. Draw a training set $\mathcal{D}$ of $\mathcal{N}(\delta, \epsilon)$ instances from $\mathcal{P}$
2. Decision Rule for any x:

   "detect": If x has a τ-rare pattern

   "reject": Otherwise

Is RarePatternDetect Sample efficient?

Sample Complexity of RarePatternDetect

- For finite pattern space $\mathcal{H}$:

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

- For infinite pattern space $\mathcal{H}$, but bounded VC-dimension $\mathcal{V}_{\mathcal{H}}$:

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

- Polynomial in $\mathcal{V}_{\mathcal{H}}$, $\frac{1}{\delta}$ and $\frac{1}{\epsilon}$

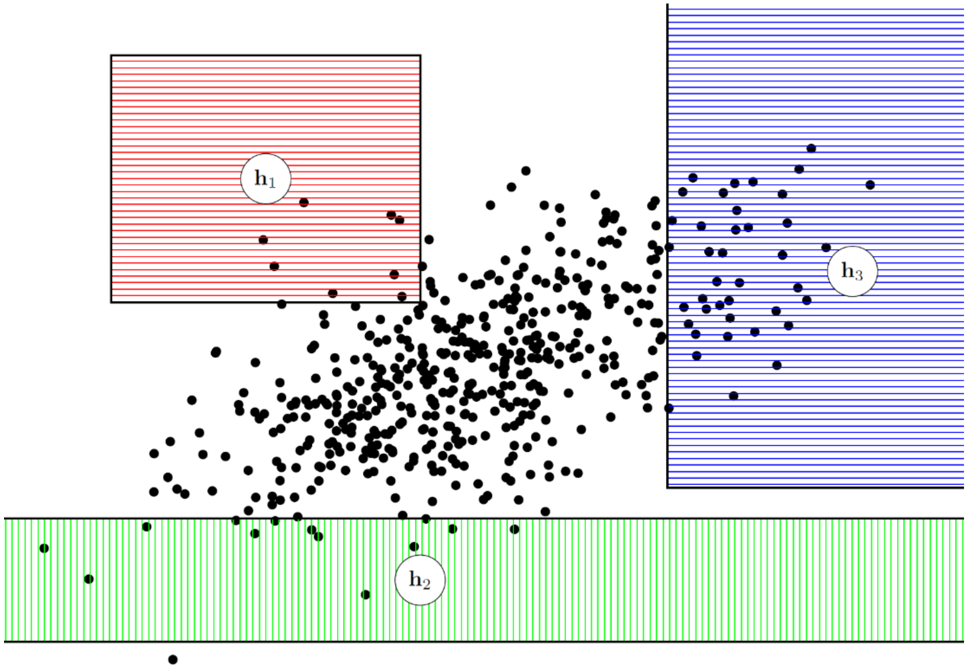
- For the example spaces, $\mathcal{V}_{\mathcal{H}}$ are polynomial in data dimension d
- Hence, $\mathcal{H}$ can be learned efficiently

Pattern Spaces for Anomaly Detectors

- Half-spaces
  - ✓The half-space mass algorithm [Chen et al. 2015]
- Axis aligned hyper rectangle
  - ✓Isolation Forest [Liu et al. 2008] and RS-Forest [Wu et al. 2014]
- Stripes
  - ✓Light weight online detectors of anomaly (LODA) [Tomas Pevny 2016]
- Ellipsoids and shells
  - ✓Density based detectors, for example, multivariate Guassians

Axis Aligned Hyper Rectangles

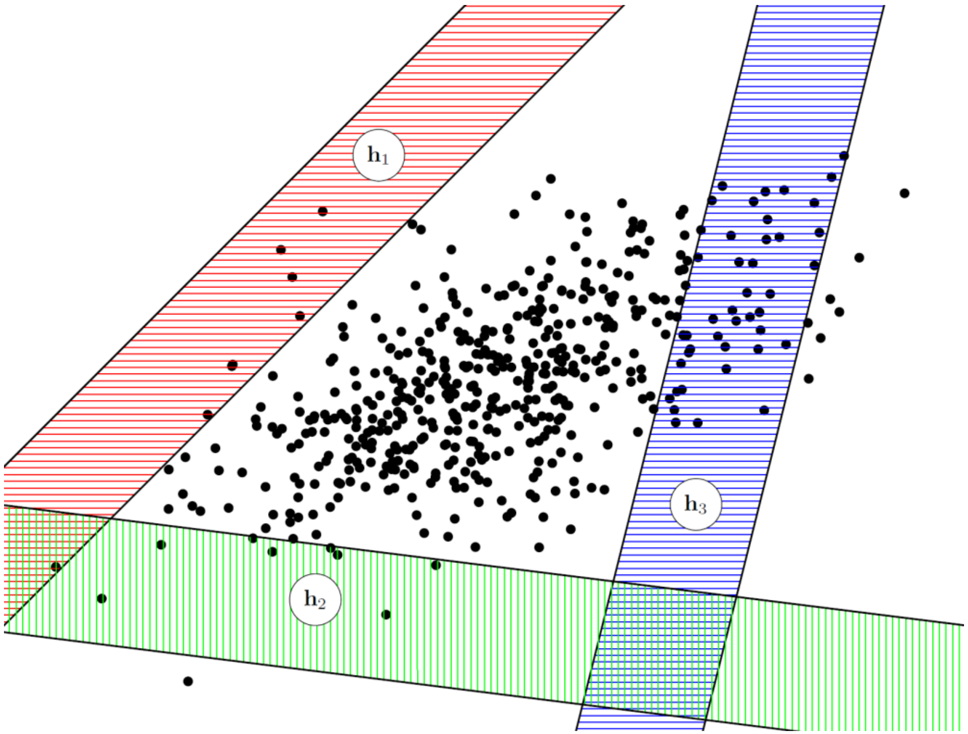
- An axis aligned hyper rectangle (bounded or unbounded) is defined by k boundaries in d-dimensional space
- Isolation Forest [Liu et al. 2008] and RS-Forest [Wu et al. 2014]



- VC-dimension = $O(d)$

Stripes

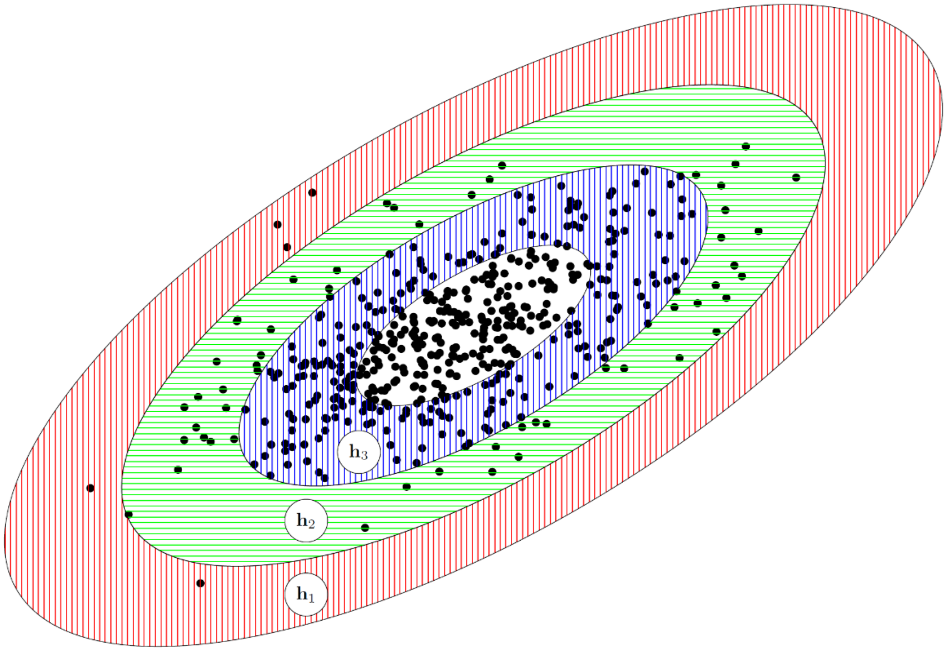
- A stripe pattern is an intersection of two parallel half-spaces with opposite orientations
- Light weight online detectors of anomaly (LODA) [Tomas Pevny 2016]



- VC-dimension = $O(d)$

Ellipsoidal Shells

- An Ellipsoidal shell is a subtraction between two ellipsoids with same center and shape but different volumes
- Density based detectors, for example, multivariate Gaussians



- VC-dimension = $O(d^2)$

Experiments

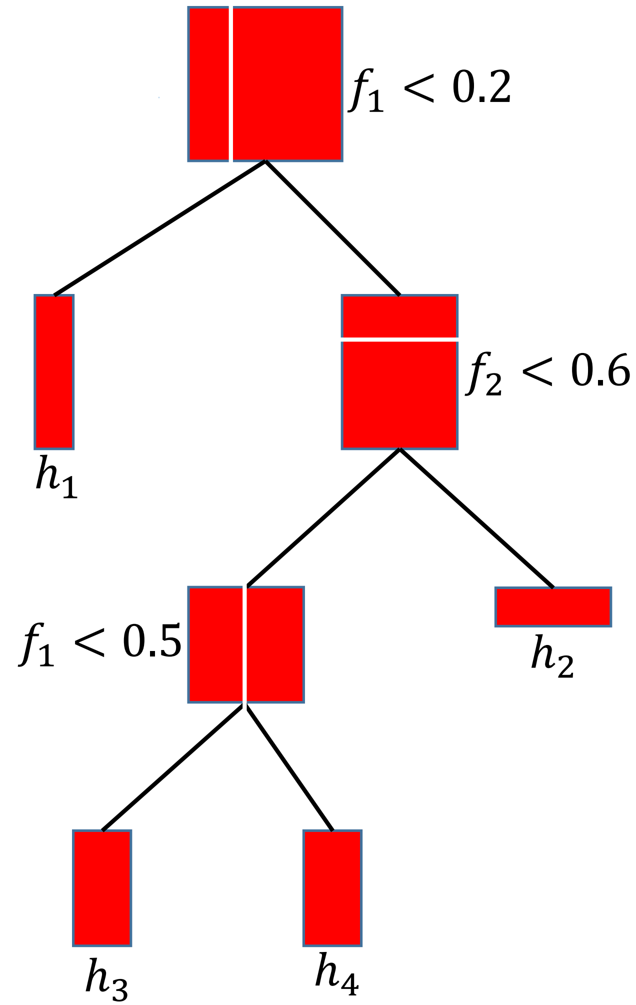
- What are the qualitative properties of the learning curves of RAREPATTERNDETECT?

- Is RAREPATTERNDETECT competitive?
  - ✓State-of-the-art anomaly detector Isolation Forest (IF)
  - ✓Pattern space: axis aligned hyper rectangles

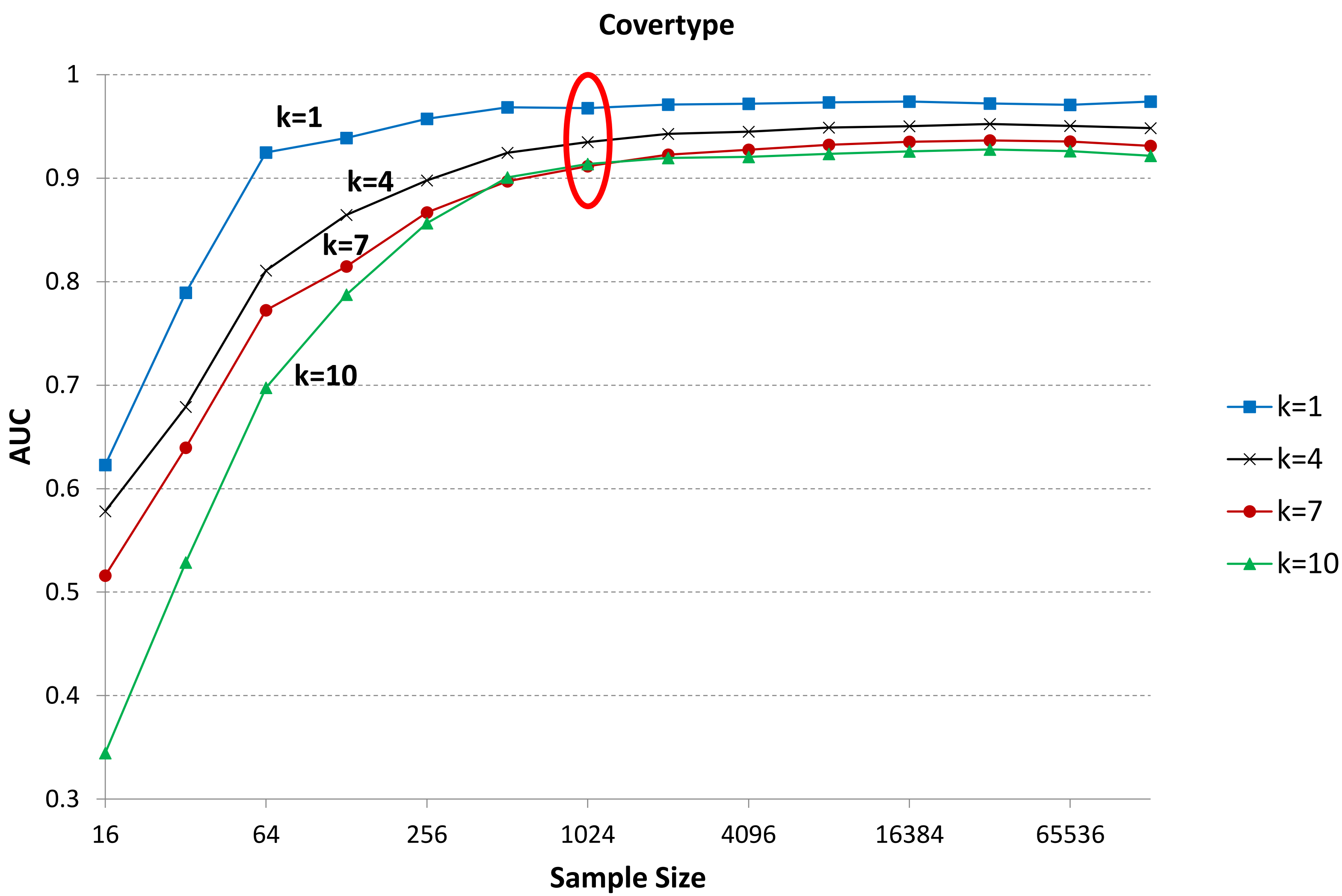
Dataset	Dimension	# Instances	% Anomaly
Covertype	10	286K	0.9%
Particle	50	130K	5%
Shuttle	9	58K	5%

Pattern Space Generation

- Construct a forest of 250 random decision trees
- Each internal node is a threshold test on a feature
- Each tree node is a pattern i.e. an axis aligned hyper rectangle
- depth (k) of the node determines the complexity of the pattern
- $\mathcal{H}_k$: Set of patterns up to k threshold tests, for example, $\mathcal{H}_2 = \{h_1, h_2\}$

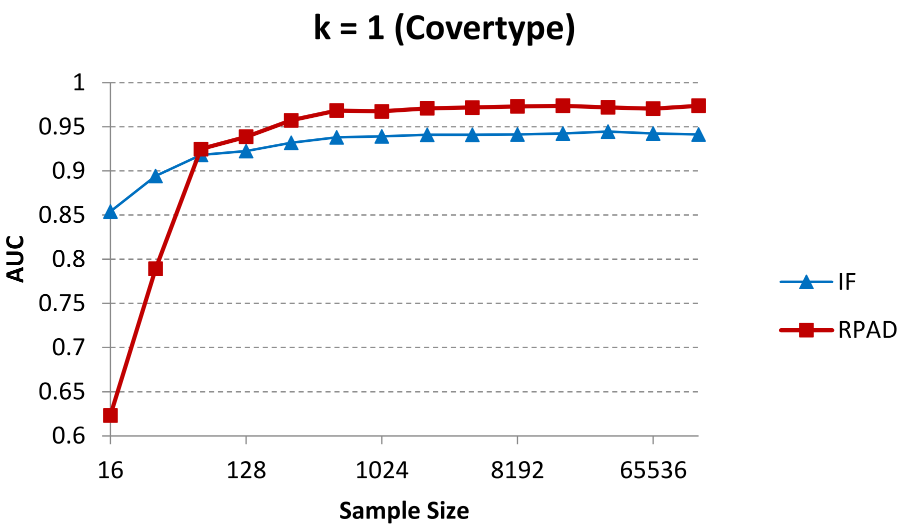


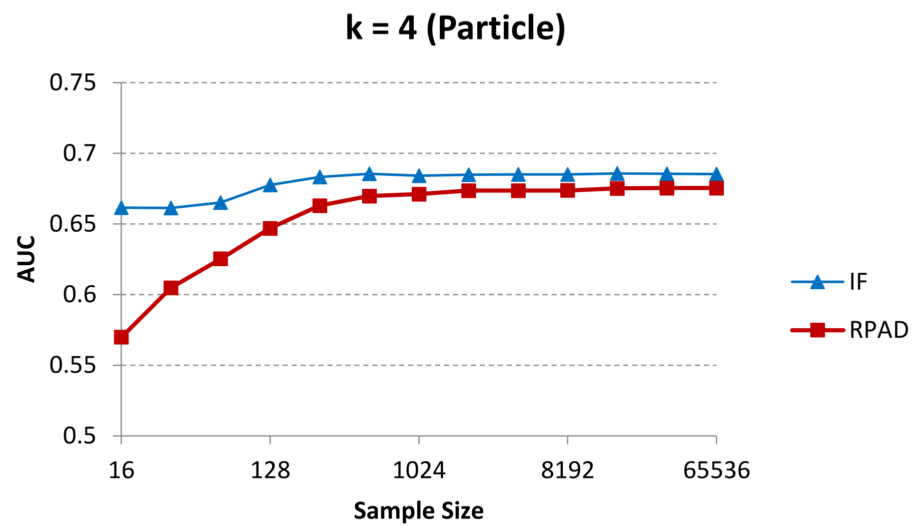
RarePatternDetect Learning Curve

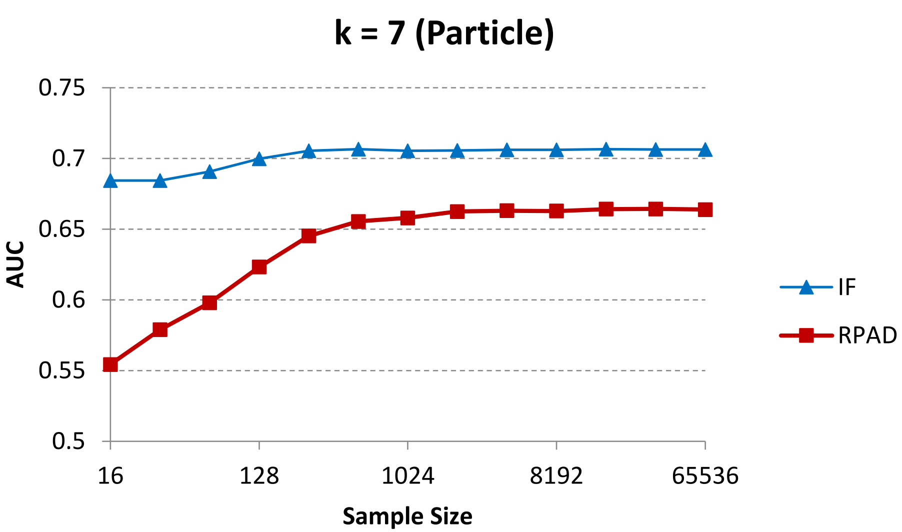


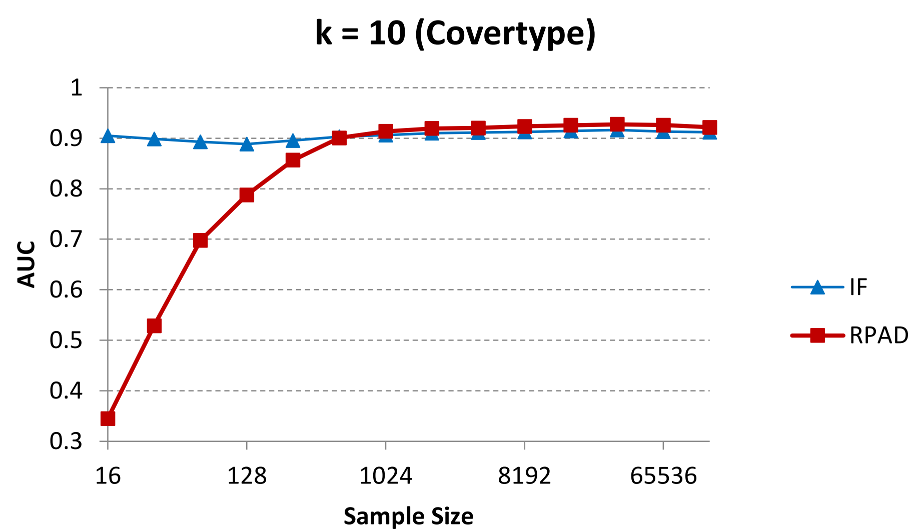
Oregon State University OSU

Comparison


k = 1 (Covertype)
AUC
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
16
128
1024
8192
65536
Sample Size
IF
RPAD


k = 4 (Particle)
AUC
0.75
0.7
0.65
0.6
0.55
0.5
16
128
1024
8192
65536
Sample Size
IF
RPAD


k = 7 (Particle)
AUC
0.75
0.7
0.65
0.6
0.55
0.5
16
128
1024
8192
65536
Sample Size
IF
RPAD


k = 10 (Covertype)
AUC
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
16
128
1024
8192
65536
Sample Size
IF
RPAD

Summary

- We developed a PAC framework to better understand the sample complexity of modern anomaly detection
- To the best of our knowledge, this is the first study of empirical learning curves for anomaly detection
- A simple PAC-RPAD algorithm is competitive with a state-of-the-art algorithm

Questions?

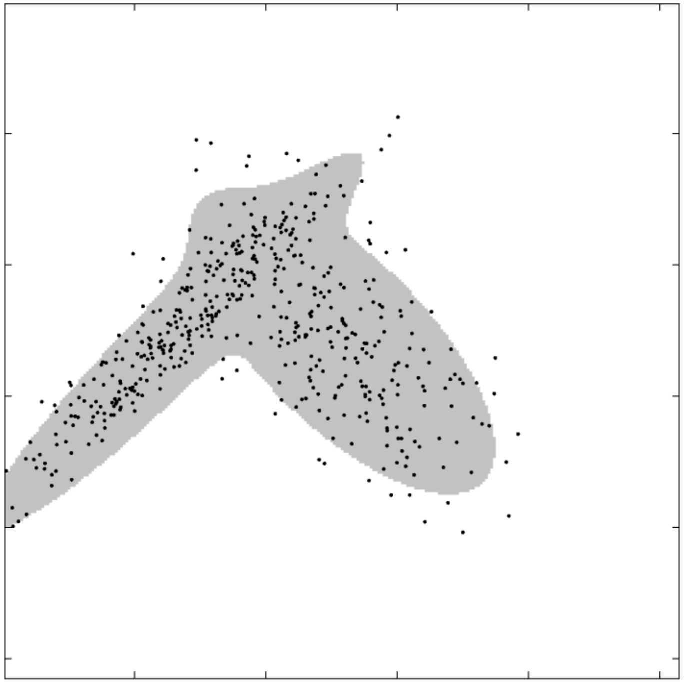
Extra Slides

Prior Work

- Sample Complexity for Anomaly Detection:

  ✓One Class SVM (Scholkopf et al. 2001)

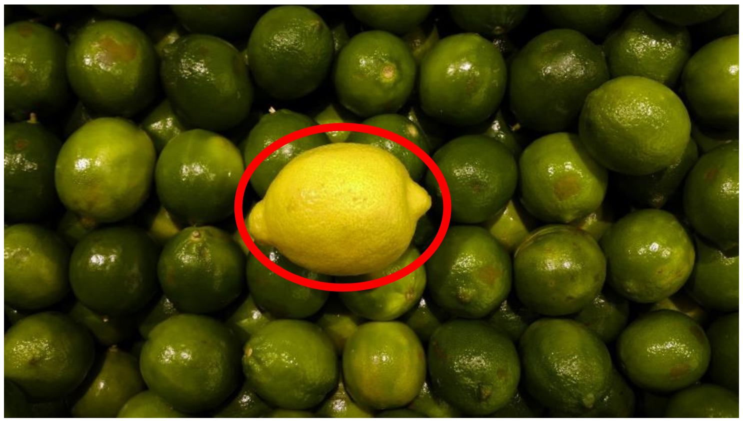
  ✓Learning Minimum Volume Sets
  (Scott & Nowak 2006)



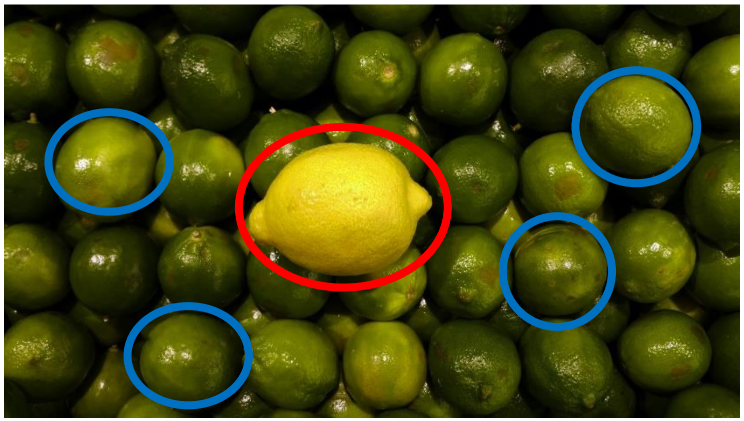
- Find a region in the input space that capture the normal points
- NOT competitive with pattern based approaches (Emmott et al. 2013)

Rare Pattern Anomaly Detection (RPAD)

- A pattern simply can be a specific color or size



Color

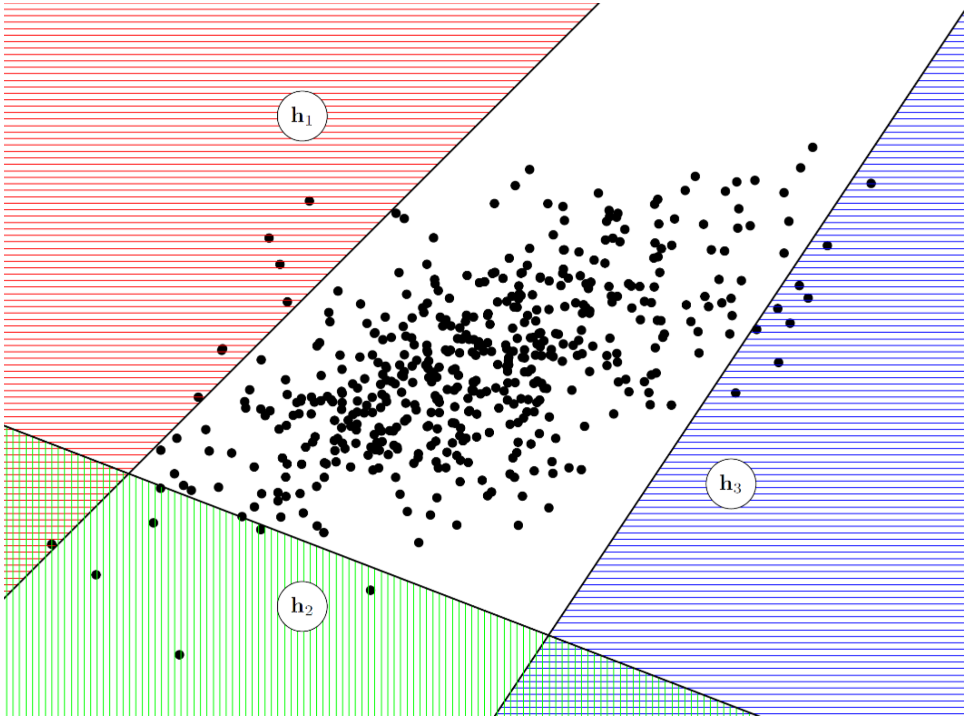


Size

- Identifies anomaly based on the characteristics of rare patterns

Half-spaces

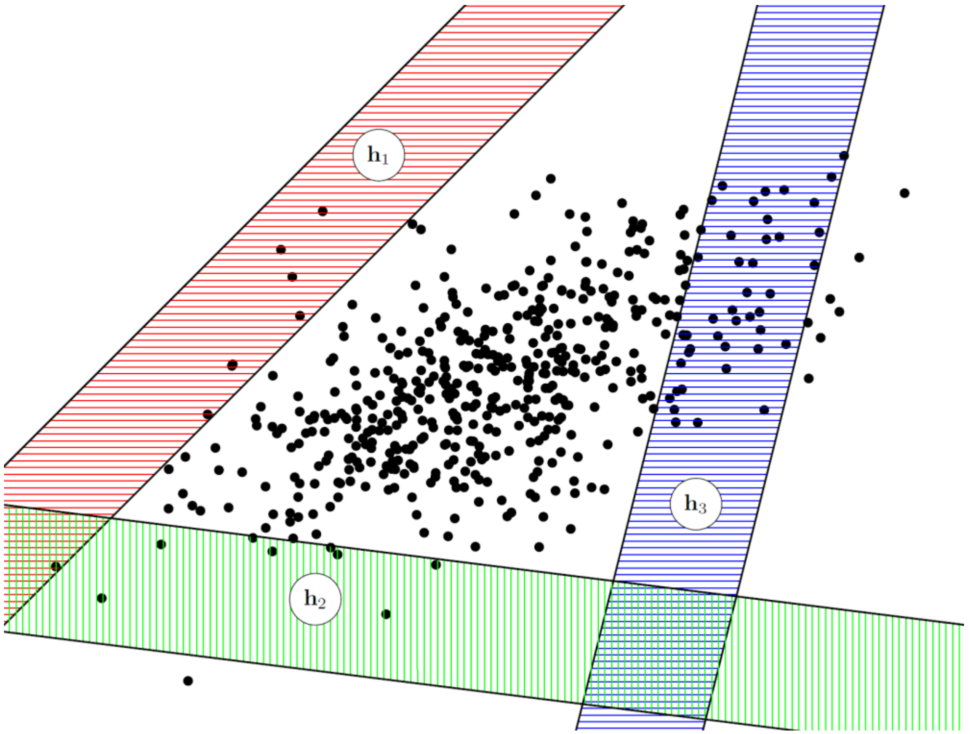
- A half-space pattern is an oriented d-dimensional hyperplane



- The half-space mass algorithm [Chen et al. 2015] operates in this pattern space
- **Anomaly score** : Mean frequency estimates of random half-spaces containing the query point x

LODA

- Construct T sparse random projections in of $\mathcal{R}^d$
- Each time, Estimate 1D histogram density from projected input data
- **Anomaly score:** geometric average of the T densities corresponding a query point



- Each bin of the histograms corresponds to a stripe in $\mathcal{R}^d$
- The perpendicular direction of the projection defines the orientation of the stripe
- Bin width corresponds to the width of the stripe