Anomaly detection and explanation

Martin Kopp

Czech Technical University in Prague

Cisco Systems, Cognitive Research Team in Prague

Institute of Computer Science, Academy of Sciences of the Czech Republic

March 12, 2015

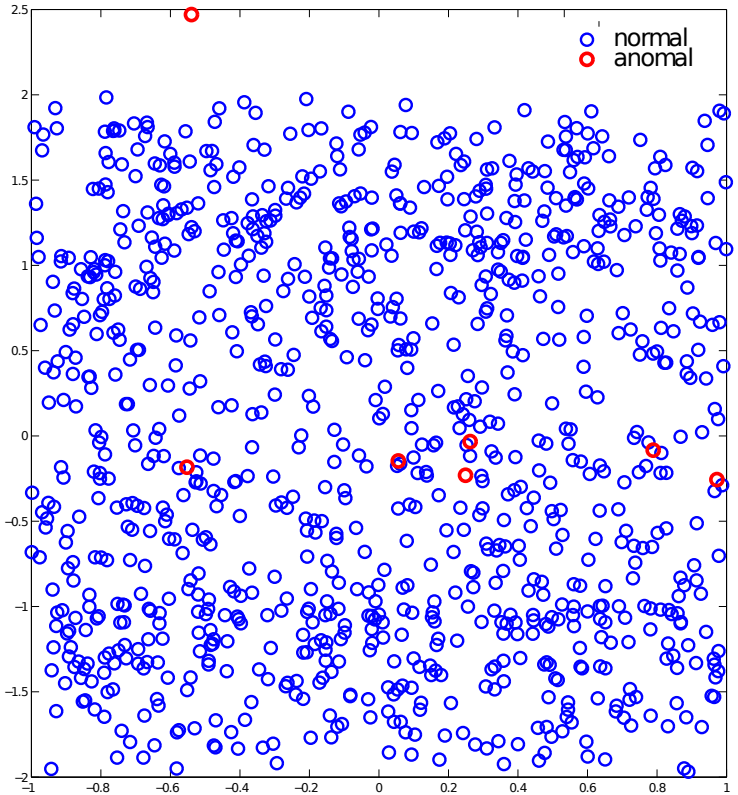
Outline

1. Anomaly detection
2. Anomaly explanation
   - Sapling random forests
   - minimal explanation
   - maximal explanation
   - rules aggregation
3. Clustering
   - voting vectors
   - feature deviations
   - evaluation

Outline

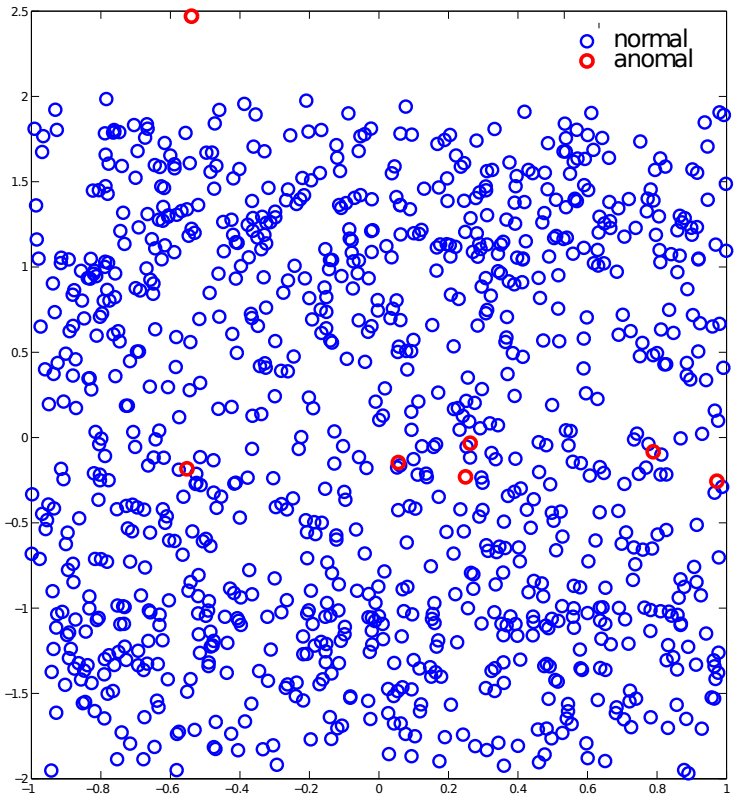
Anomaly detection
motivation

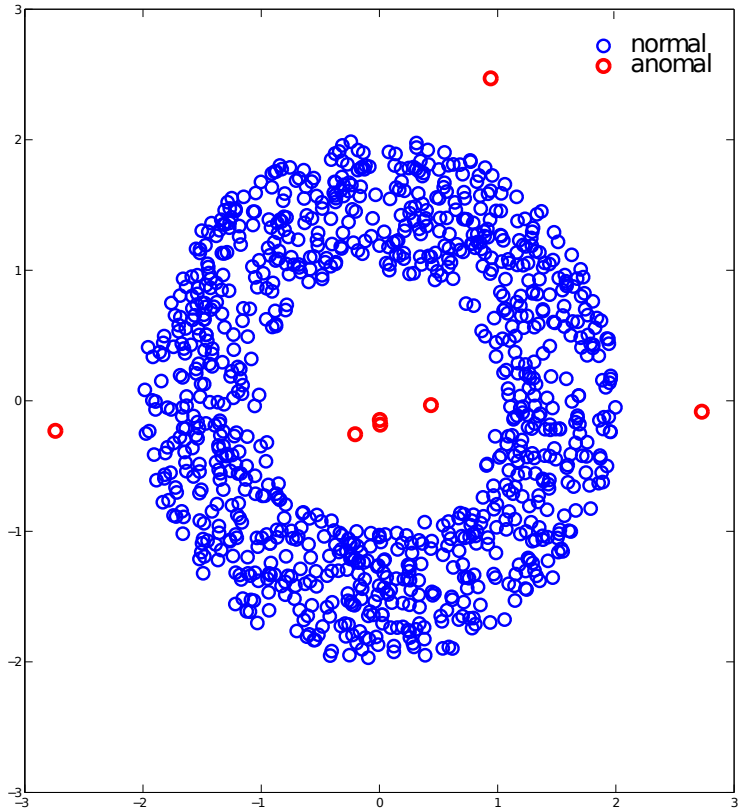
Anomaly detection is about ...



Anomaly detection
motivation

... point of view.





Anomaly detection
motivation

Anomaly in crowd

[1] www.svcl.ucsd.edu/projects/anomaly/

Anomaly detection
motivation

Network security

- typical proportion of anomalies is $1 - 0.1\%$
- 0.5 million data points $\rightarrow$ 1000 anomalies

Particle physics

- typical proportion of anomalies is $10^{-3} - 10^{-4}\%$
- 2 million data points $\rightarrow$ 100 anomalies

Anomaly detection
motivation

Network security

- typical proportion of anomalies is $1 - 0.1\%$
- 0.5 million data points $\rightarrow$ 1000 anomalies

Particle physics

- typical proportion of anomalies is $10^{-3} - 10^{-4}\%$
- 2 million data points $\rightarrow$ 100 anomalies

Anomaly detection
problem statement

" *An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.*"

[1] Hawkins 1980 - Identification of outliers

Anomaly detection
problem statement

- Defining a normal region for every possible normal behaviour is very difficult.
- The boundary between normal and anomalous behaviour is often not precise.
- Some anomalous events often adapt to appear normally.
- Even normal behaviour may evolve over time.
- Obtaining labelled data for training and validation of models is usually a major issue.
- Often the data contains noise that tends to be similar to the actual anomalies and hence is difficult to distinguish and remove.

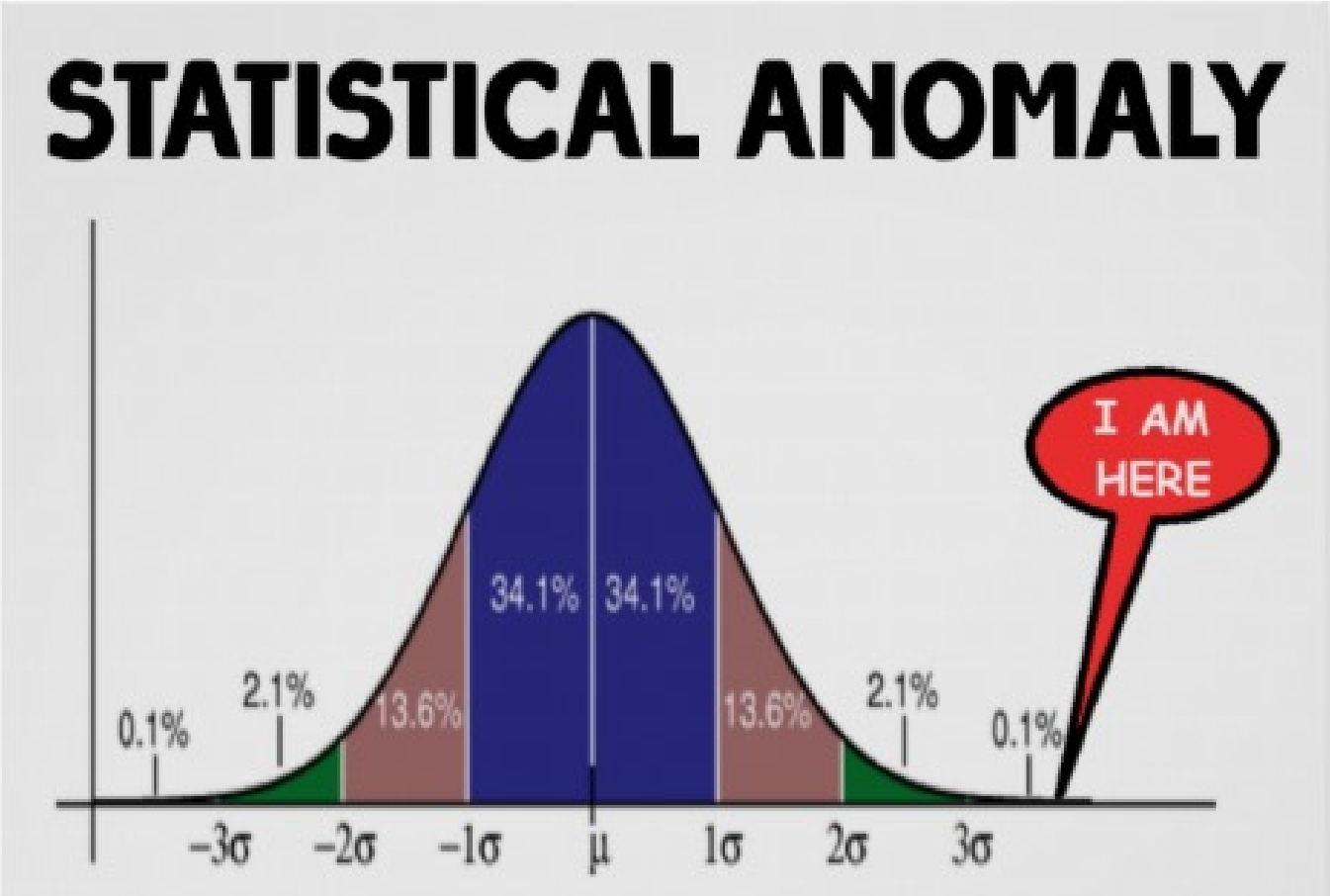
Anomaly detection
type of detectors

Anomaly detectors

- Statistical
- Linear
- Proximity based
  - cluster
  - distance
  - density
- domain specific

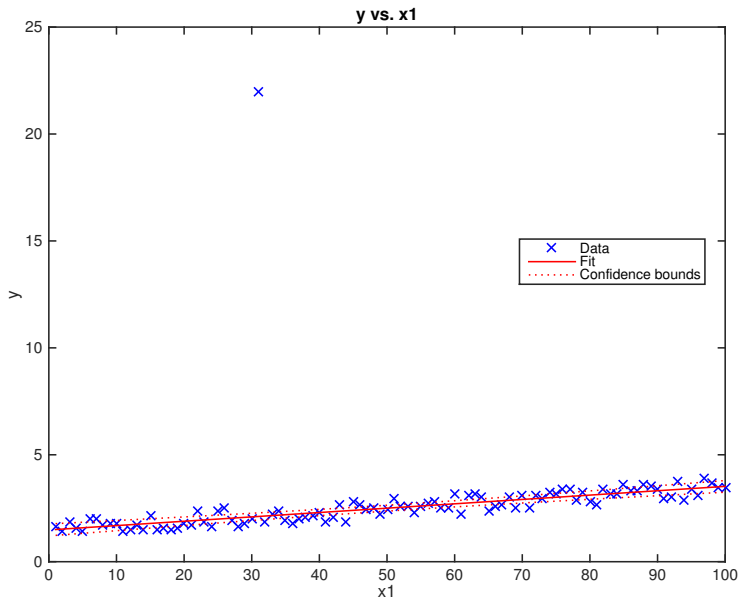
Anomaly detection
type of detectors

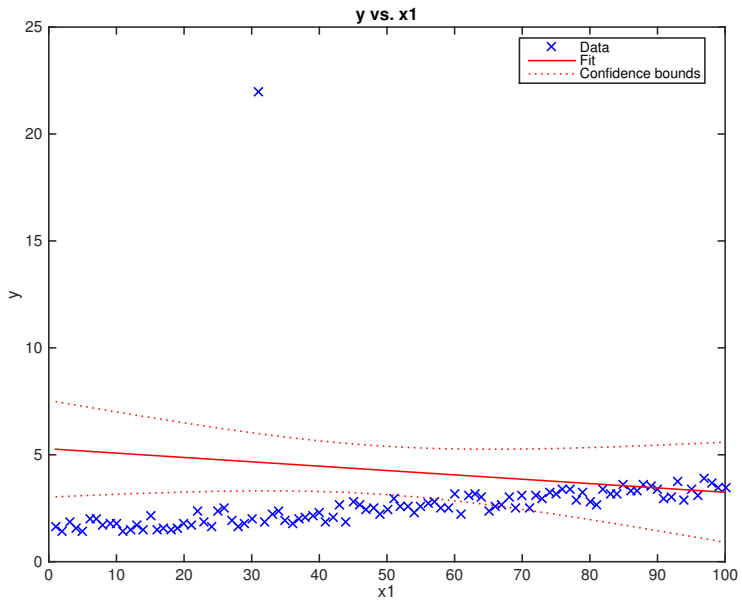
Statistical anomaly detectors



Anomaly detection
type of detectors

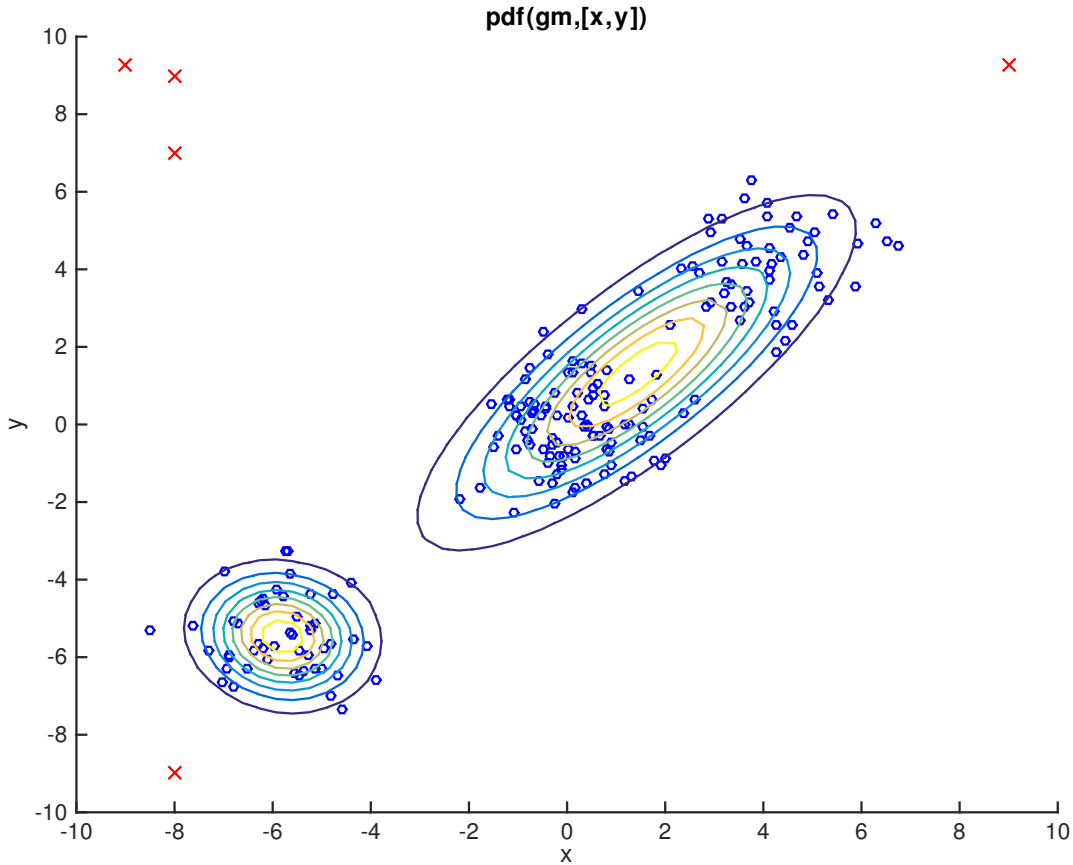
Linear model based detectors





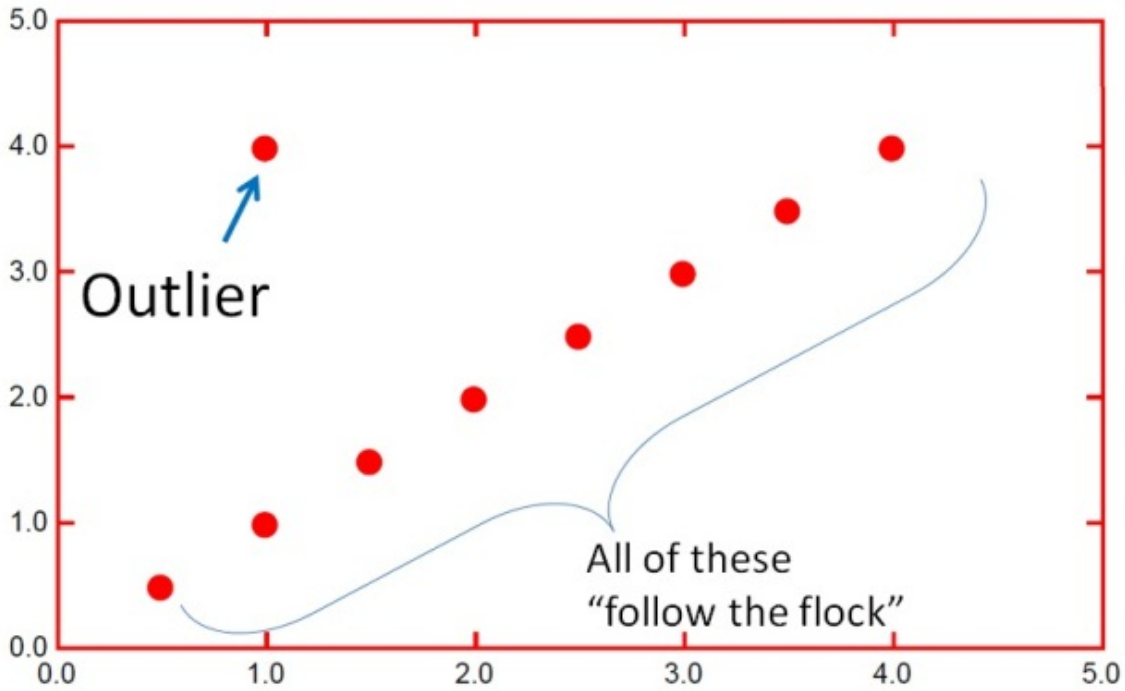
Anomaly detection
type of detectors

Cluster based detectors



Anomaly detection
type of detectors

Distance based detectors



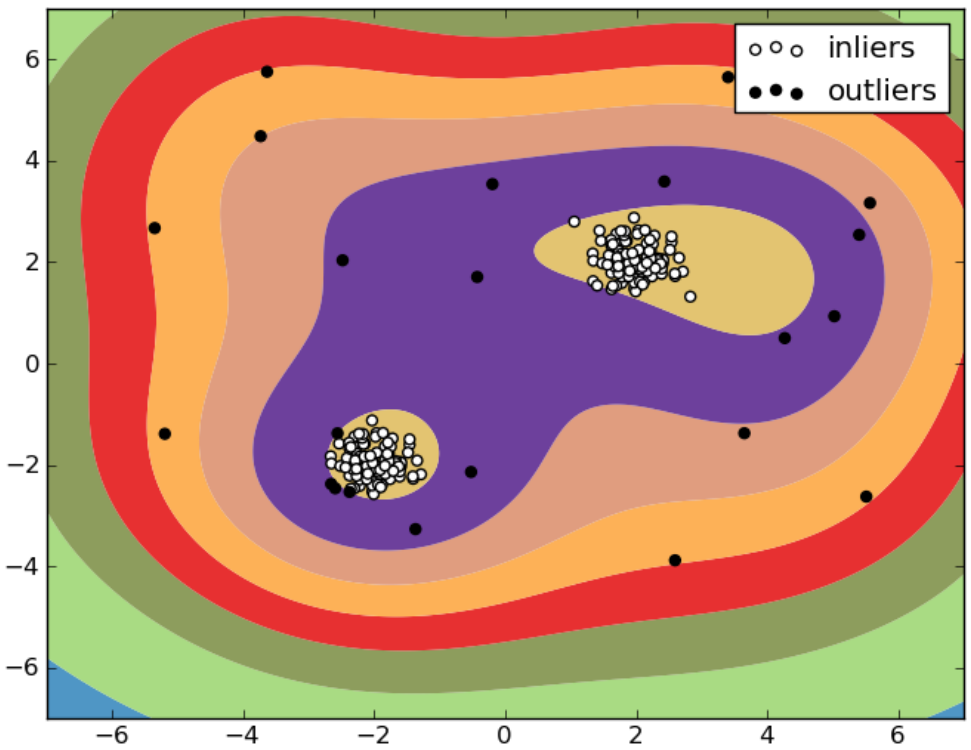
Never mind what the axes mean...

---

https://baldscientist.wordpress.com/2013/02/02/is-free-will-a-matter-of-being-a-conscious-outlier/

Anomaly detection
type of detectors

Density based detectors



http://scikit-learn.org/stable/modules/outlier_detection.html

Outline

Anomaly explanation
history

- Grubbs 1950 - Anomaly detection [1]
- Knorr 1999 - Question [2]
- Dang 2013 - Answer [3]

[1] Grubbs 1950 - *Sample criteria for testing outlying observations.*

[2] Knorr, Edwin M., and Raymond T. Ng. 1999 - *Finding intensional knowledge of distance-based outliers.*

[3] Dang, Xuan Hong, et al. 2013 - *Local outlier detection with interpretation.*

Anomaly explanation
motivation

- Network security
  - attack vs. unscheduled backup
- Particle physics
  - Higgs boson vs. misconfiguration of equipment
- Astronomy
  - cosmic microwave background vs. pigeon nest
- Fraud detection
  - holiday vs. credit card fraud

Anomaly explanation
problem definition

We have:

- dataset
- anomaly detection algorithm
- labelled suspicious samples

We want:

- examine the suspicious samples
- interpret them clearly
  - as a small subset of features
  - as human readable set of rules

Anomaly explanation
problem definition

We have:

- dataset
- anomaly detection algorithm
- labelled suspicious samples

We want:

- examine the suspicious samples
- interpret them clearly
  - as a small subset of features
  - as human readable set of rules

Sapling Random Forest

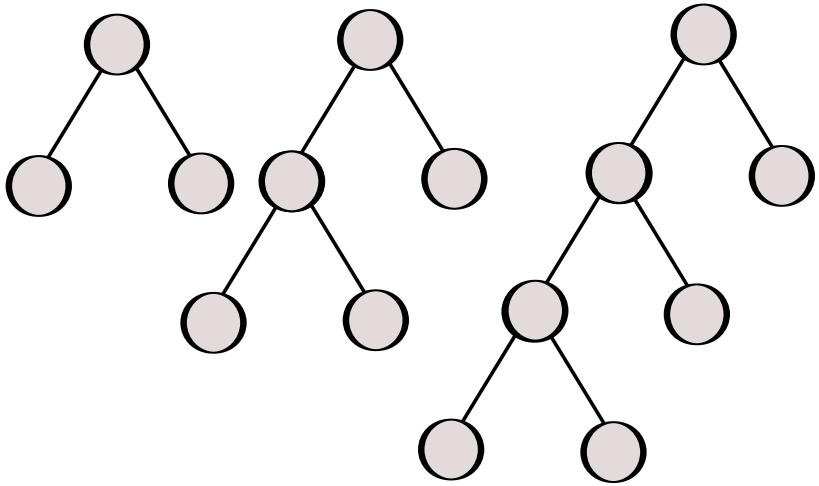
sapling

(a) In nature

Sapling Random Forest

sapling

(a) In nature



(b) In theoretical informatics

Sapling Random Forest
summary

- ensembles of specifically trained CARTs
- multiple trees per anomaly
- specifically made training sets -> grow sets
- trees are quite small -> saplings

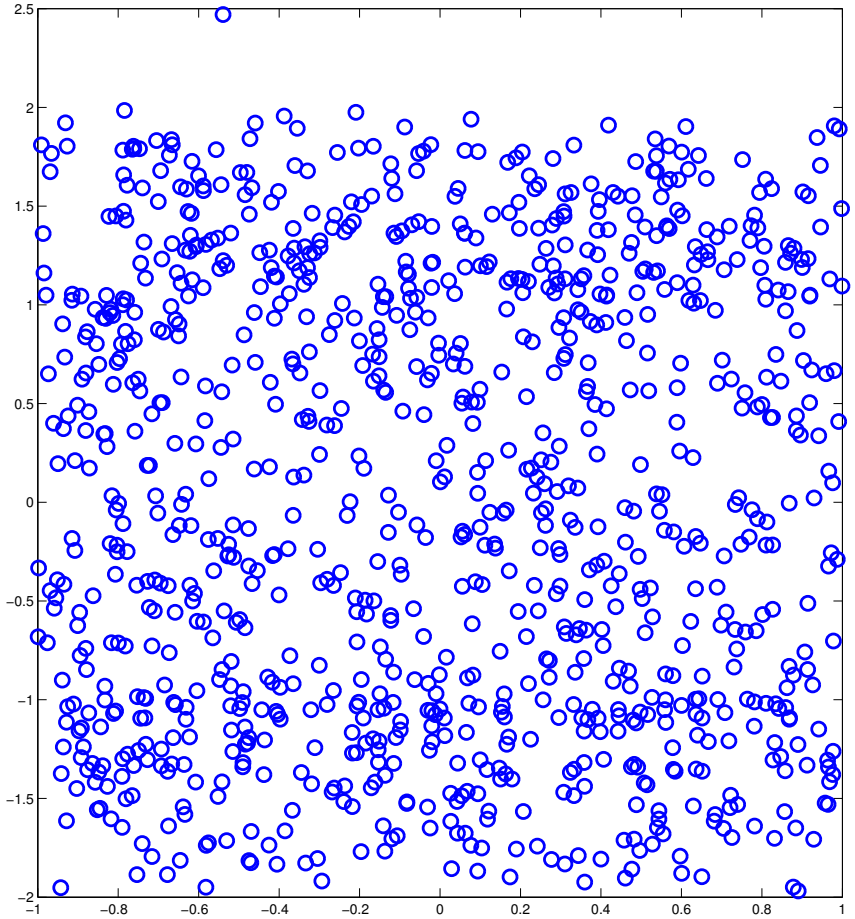
Sapling Random Forest
summary

Summary of the SRF for minimal explanation

Input: $data$
$y \leftarrow anomalyDetector(data)$
for all data($y ==$anomaly) **do**
$G \leftarrow createGrowSet(size, method)$
$T \leftarrow trainTree(G)$
$SRF \leftarrow T$
end for
$extractRules(SRF)$

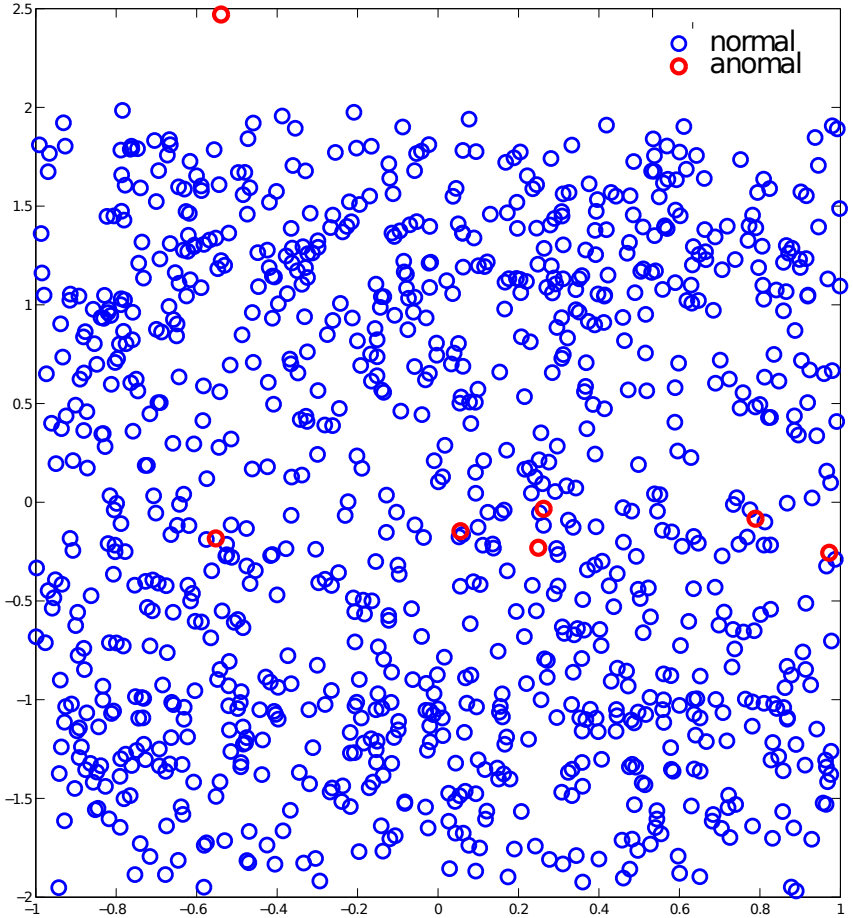
Sapling Random Forest
algorithm

Input: *data*



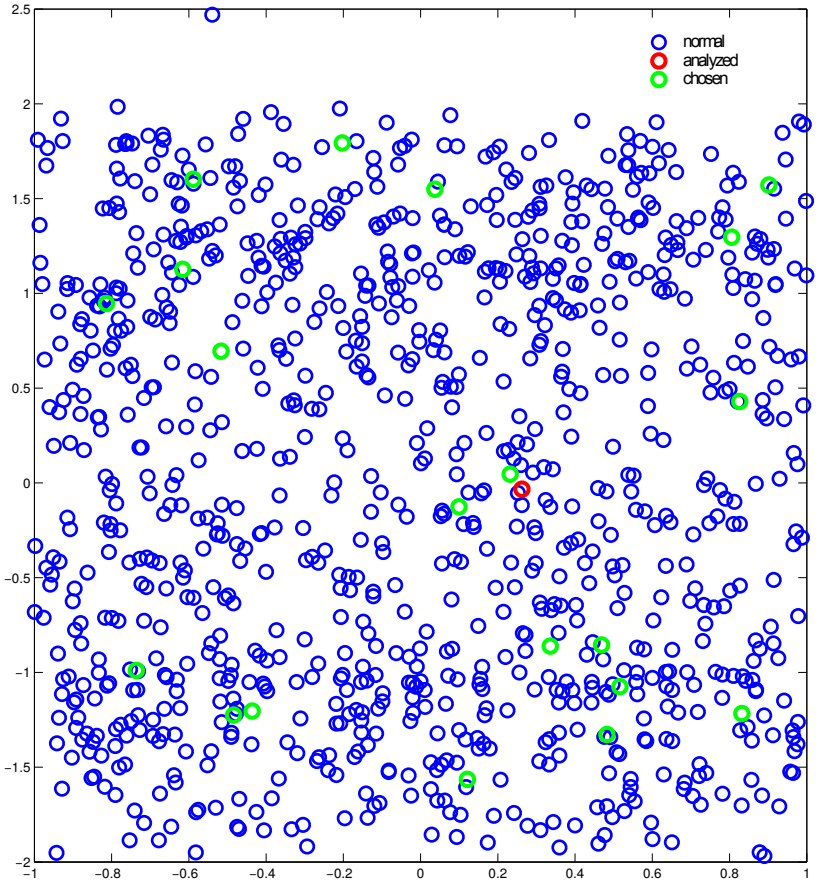
Sapling Random Forest
algorithm

$$y \leftarrow anomalyDetector(data)$$

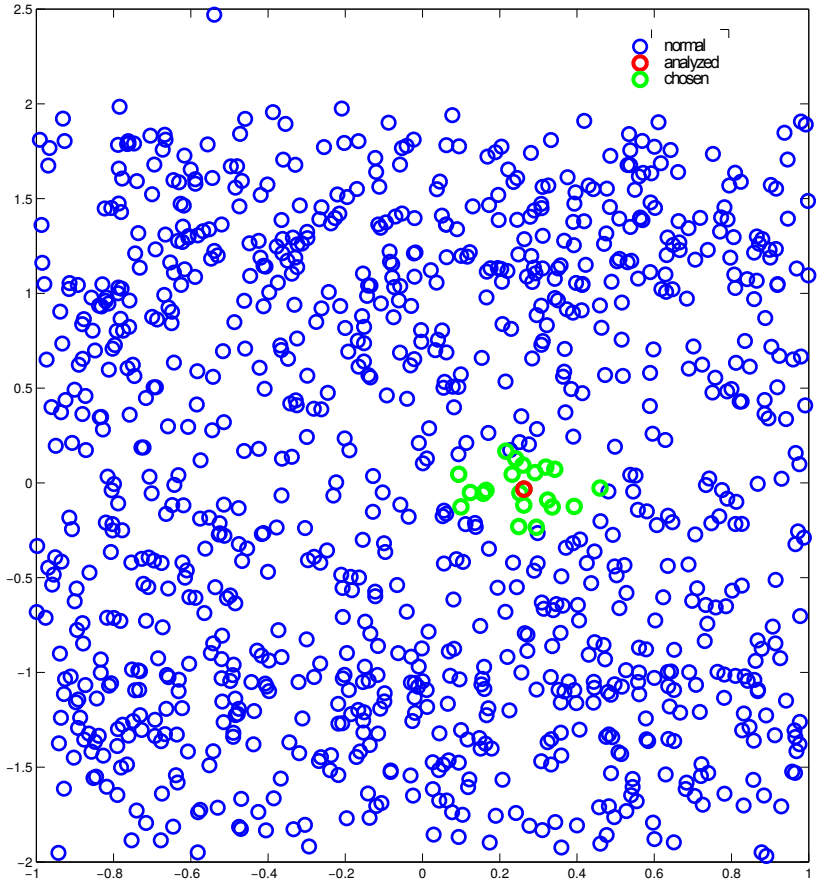


Sapling Random Forest
algorithm

$$G \leftarrow createGrowSet(size, method)$$



(a) random selection



(b) k-nn selection

Sapling Random Forest

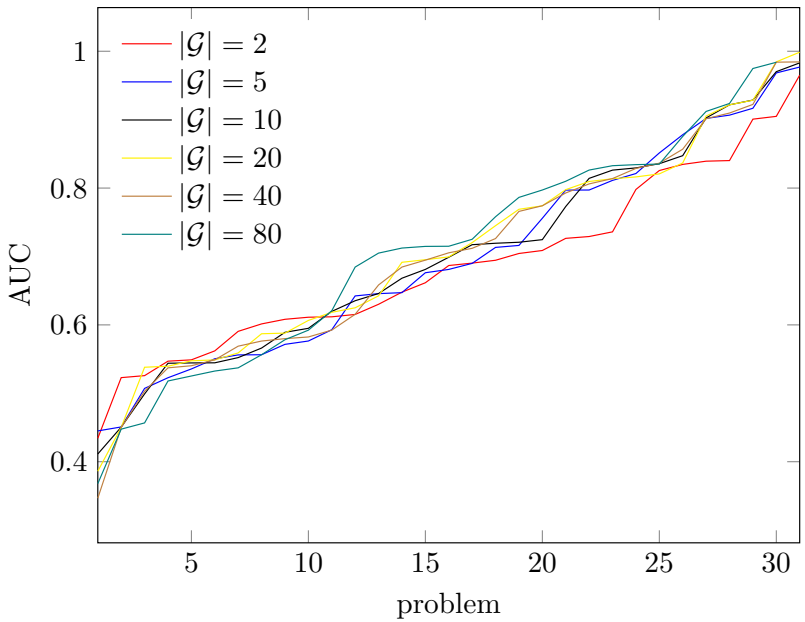
Grow set selection

A grow set $\mathcal{G}$ contains an anomaly x^a and several normal samples $x^n \subseteq \mathcal{X}^n$.

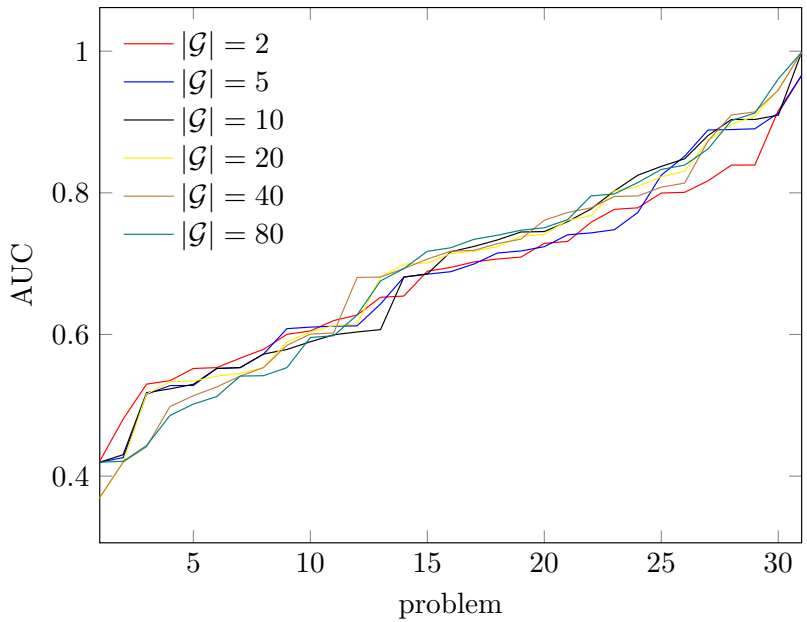
- typical size $|\mathcal{G}| = 100$
- random selection
  - fast even in high dimensions
  - multiple trees can be grown -> robust
- k-nn selection
  - deterministic - more trees are useless
  - slow in high dimensions
  - superior in low dimensions

Sapling Random Forest

Grow set selection



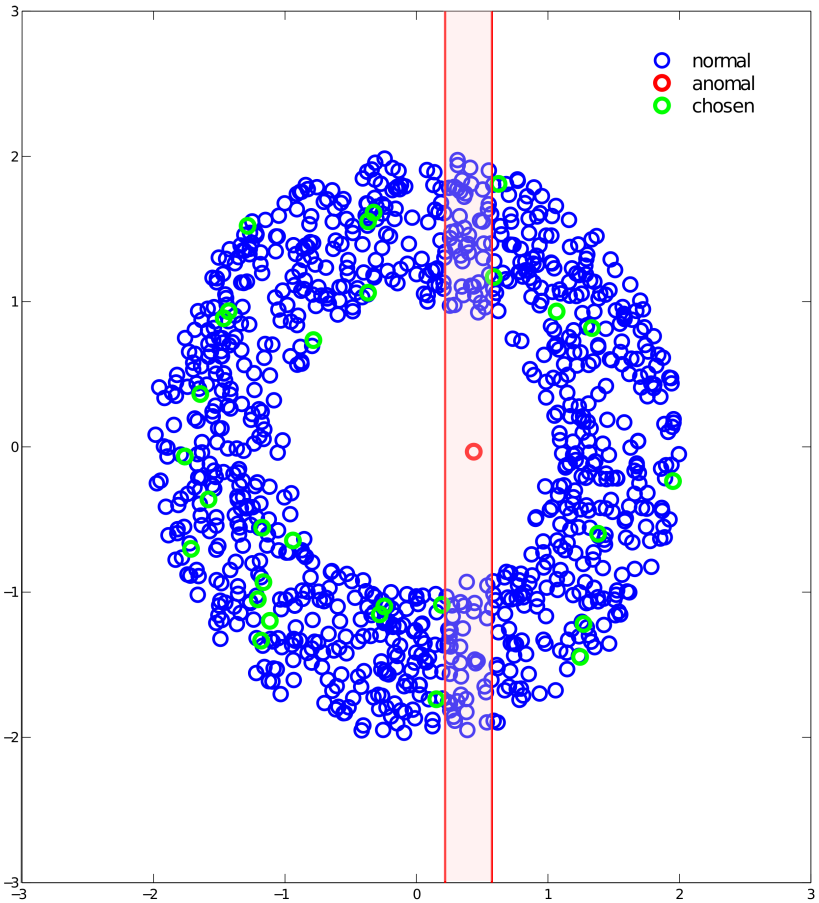
(a) random selection

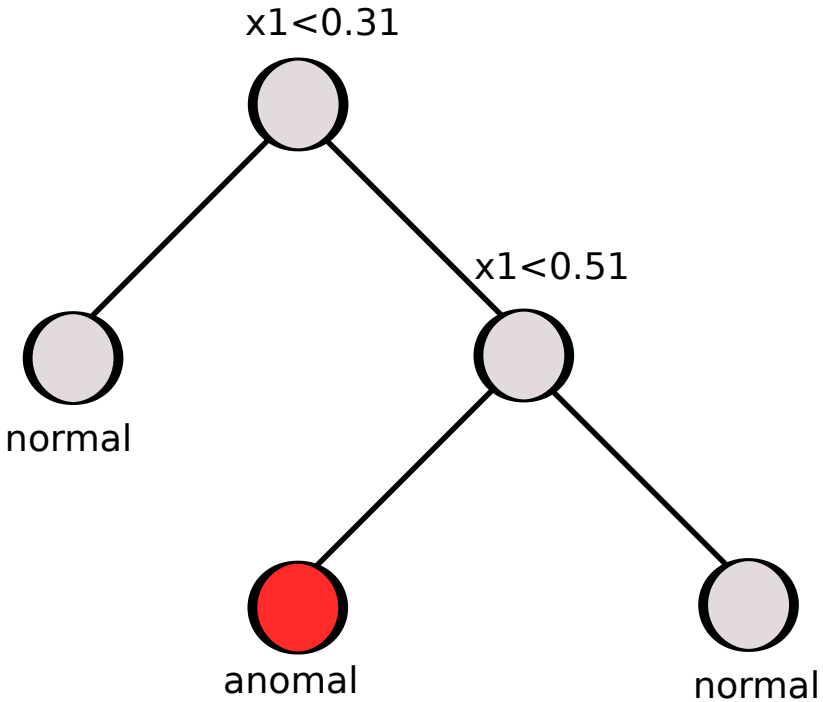


(b) k-nn selection

Sapling Random Forest
tree training

$$T \leftarrow trainTree(G)$$





Sapling Random Forest
splitting criterion

Gini's index

$$\mathrm{G_i} = 1 - p_a^2 - p_n^2,$$

Information gain

$$\arg\max_{h \in \mathcal{H}} \quad -\sum_{b \in \{L,R\}} \frac{|\mathcal{S}^b(h)|}{|\mathcal{S}|} H(\mathcal{S}^b(h)),$$

Sapling Random Forest
splitting criterion

Simplified criterion

$$\arg\min_{h \in \mathcal{H}} |\mathcal{S}^a(h)|,$$

Sapling Random Forest
splitting criterion

Simplified criterion

$$\arg\min_{h\in\mathcal{H}} |\mathcal{S}^a(h)|,$$

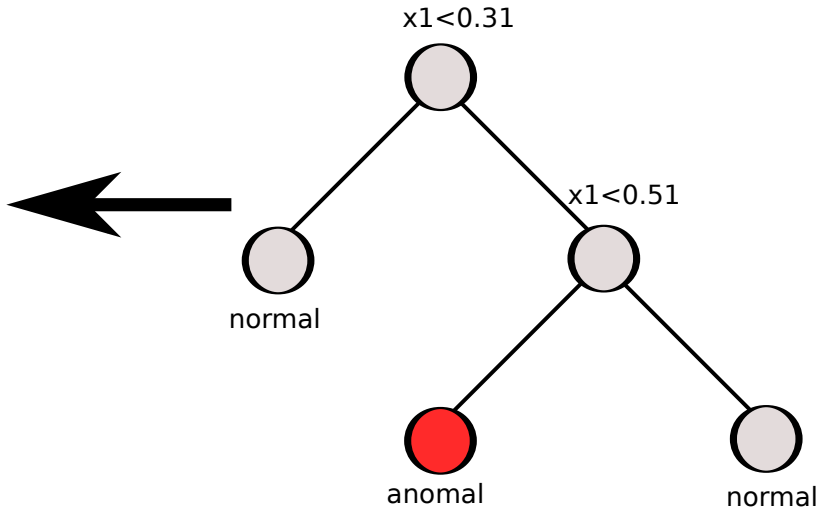
Maximal margin

$$\arg\max_{d\in\mathcal{D}} \max \min \mathcal{S}_d^n - x_d^a$$

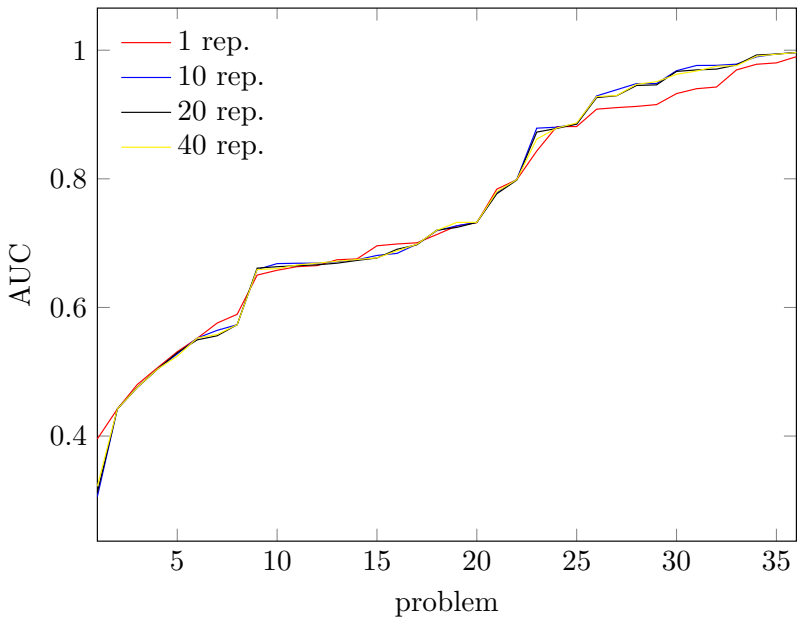
$$\arg\max_{d\in\mathcal{D}} \inf \mathcal{S}_d^n - x_d^a$$

Sapling Random Forest
algorithm

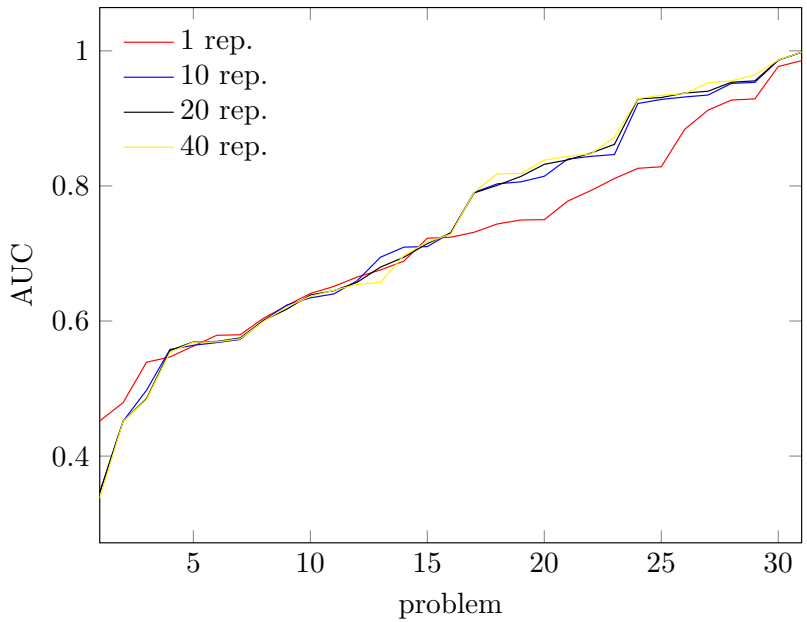
$$SRF \leftarrow T$$



Sapling Random Forest
tree training



(a) ground truth

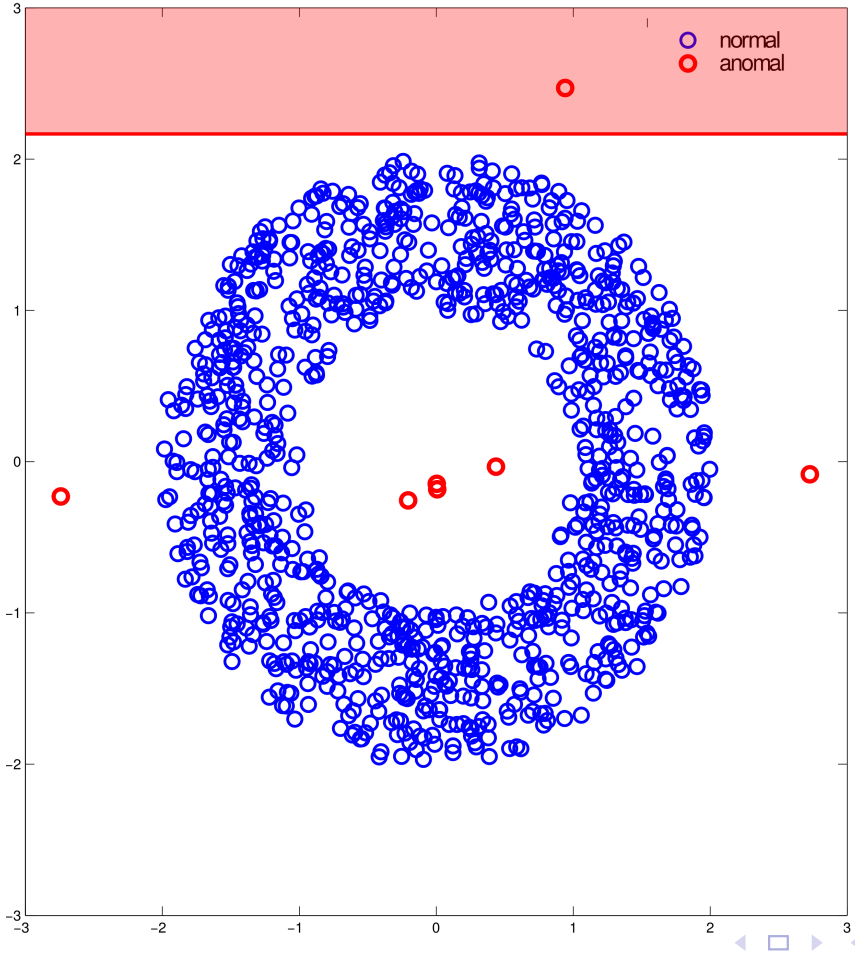


(b) Local Outlier Factor

Sapling Random Forest
explaining an anomaly

$$extractRules(SRF)$$

$$C = x_2 > 2.2$$

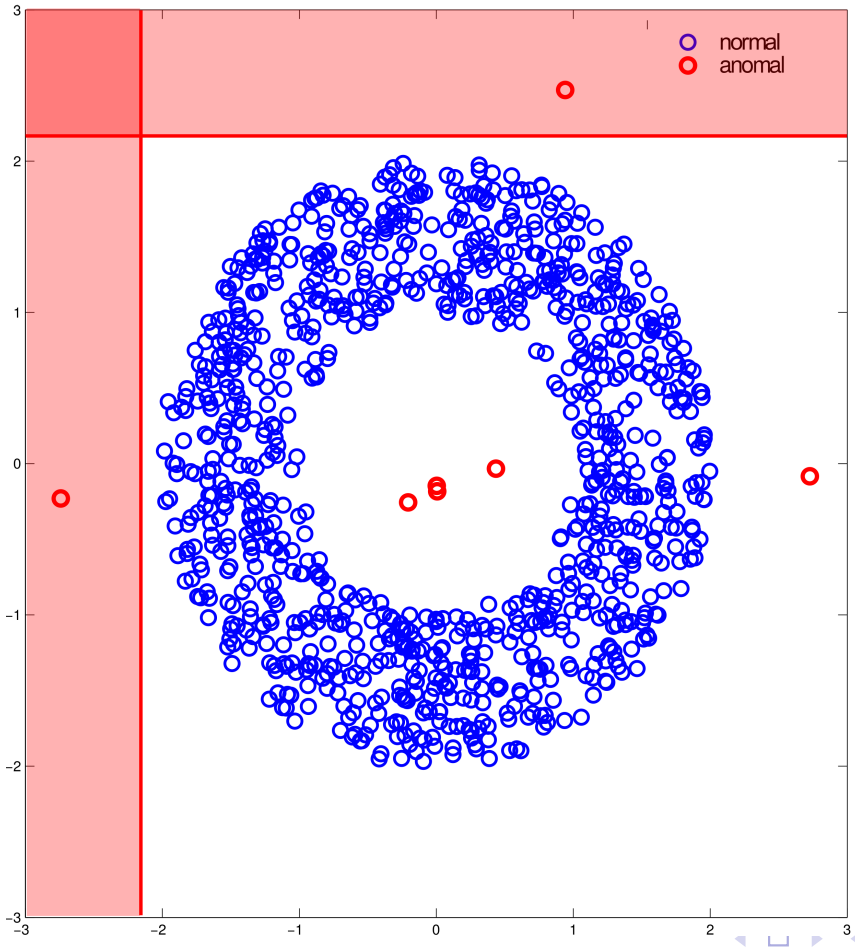


Sapling Random Forest

explaining an anomaly

$$extractRules(SRF)$$

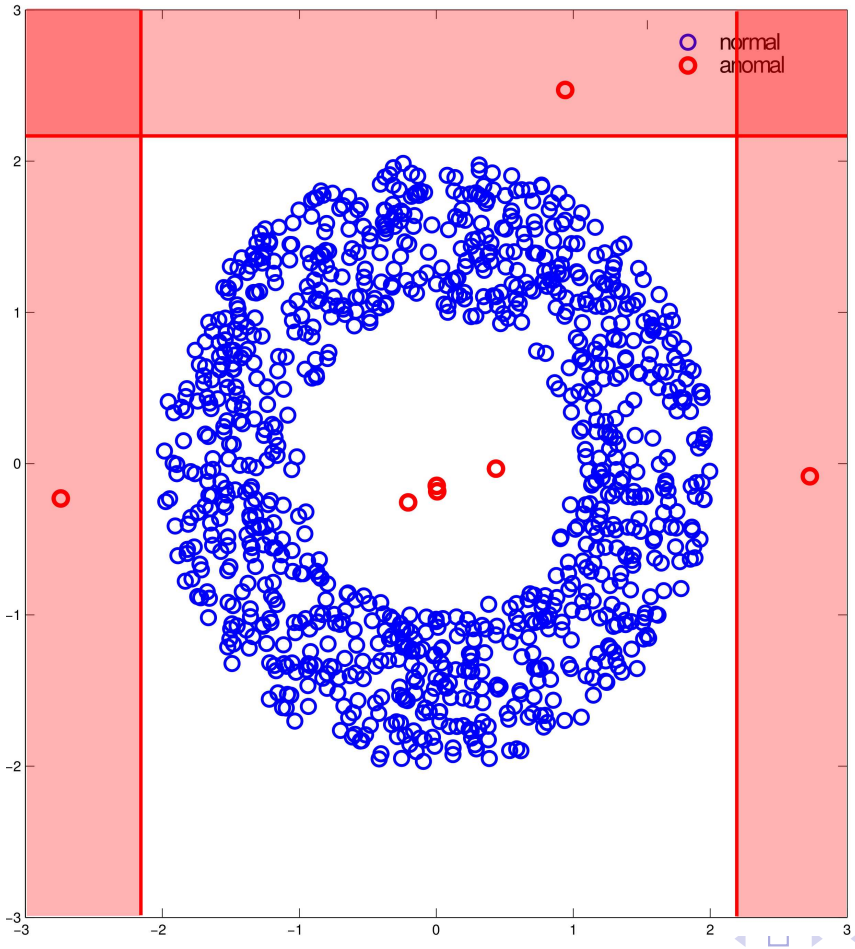
$$C = (x_2 > 2.2) \wedge (x_1 < -2.1)$$



Sapling Random Forest
explaining an anomaly

$$extractRules(SRF)$$

$$C = (x_2 > 2.2) \wedge (x_1 < -2.1) \wedge (x_1 > 2.2) \wedge \ldots$$



Sapling Random Forest
explaining an anomaly

The set of all possible rules is defined as $\mathcal{H} = \{h_{j,\theta} | j \in \{1, \ldots, d\}, \theta \in \mathbb{R}\}$ where

$$h_{j,\theta}(x) = \begin{cases} +1 & \text{if } x_j > \theta \\ -1 & \text{otherwise} \end{cases}$$

- d . . . number of features
- θ . . . inner node threshold
- x_j . . . jth feature of sample x

Sapling Random Forest
explaining an anomaly

The set of all possible rules is defined as $\mathcal{H} = \{h_{j,\theta} | j \in \{1, \ldots, d\}, \theta \in \mathbb{R}\}$ where

$$h_{j,\theta}(x) = \begin{cases} +1 & \text{if } x_j > \theta \\ -1 & \text{otherwise} \end{cases}$$

Let $h_{j_1,\theta_1}, \ldots, h_{j_t,\theta_t}$ be the set of decisions taken in inner nodes on the path from the root to the leaf with the anomaly x^a. Then x^a is explained as conjunction of atomic conditions

Sapling Random Forest
rules extraction

Rules in form:

$$C = (x_{j_1} > \theta_1) \wedge (x_{j_2} < \theta_2) \wedge \ldots \wedge (x_{j_t} > \theta_t)$$

We calculate groups sizes

$$r_{2j} = \sum_{C \in \mathcal{D}} \sum_{h \in C} I(j \in h, L)$$

$$r_{2j-1} = \sum_{C \in \mathcal{D}} \sum_{h \in C} I(j \in h, R)$$

$$I(j \in h) = \begin{cases} +1 & \text{if } < \textit{rule} \\ -1 & \text{otherwise} \end{cases}$$

Sapling Random Forest
rules extraction

and chose only k-most frequent, where

$$k = \arg\min_{k} \frac{1}{\sum_{j=1}^{2d} r_j} \sum_{j=1}^{k} r_j > \tau$$

Then we aggregate similar rules and chose the most strict thresholds.

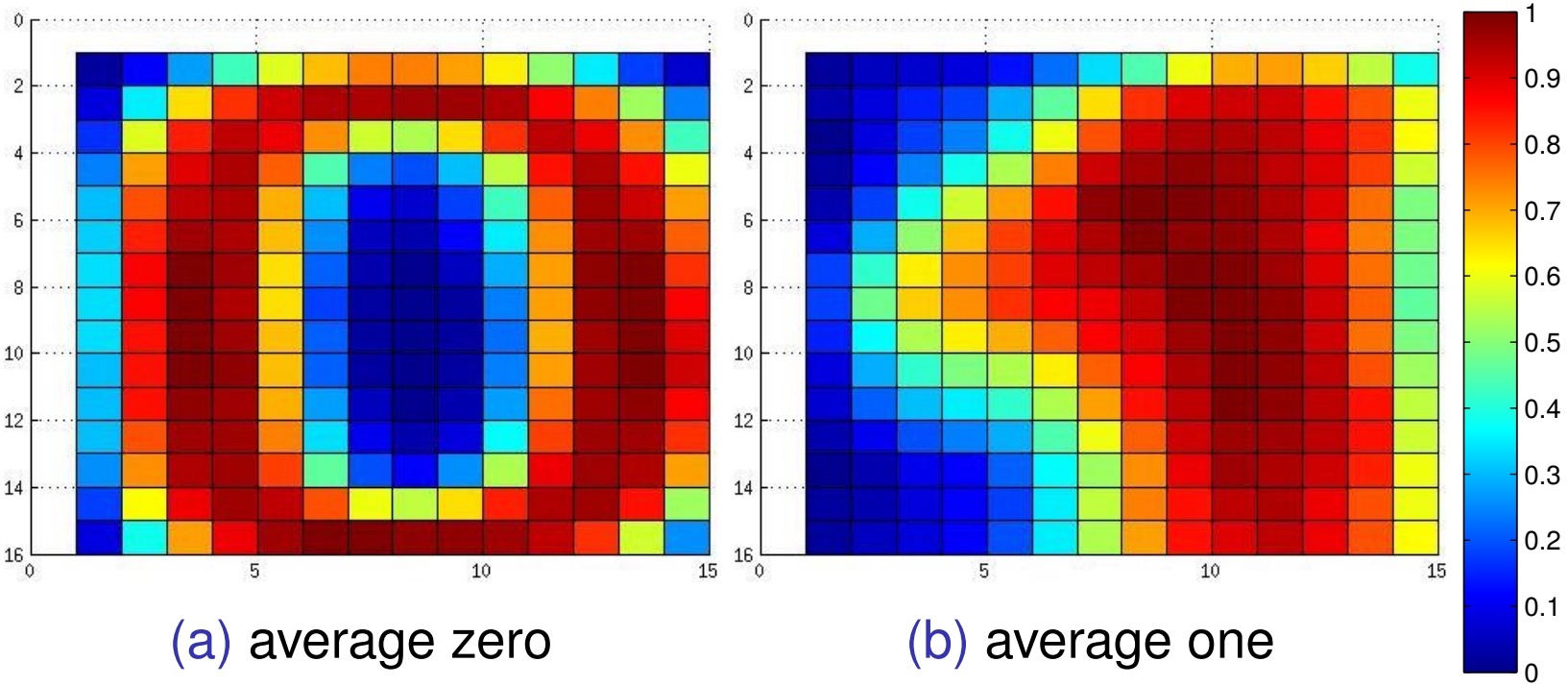
$$h_j^R = \arg\min_{h \in \mathcal{H}_j^R} \theta_h$$

Sapling Random Forest
summary

Summary of the SRF for maximal explanation

$y \leftarrow anomalyDetection(data)$
for all data($y ==$anomaly) **do**
  $f \leftarrow allFeatures$
  while $d < \tau$ **do**
    $G \leftarrow createGrowSet(size, f)$
    $t \leftarrow trainTree(G)$
    $SRF \leftarrow SRF + t$
    $f \leftarrow f - topSplitFeature(t)$
    $D = nnDistance(G)$
    $d = D(anomaly)/max(D)$
  end while
end for
$extractRules(SRF)$

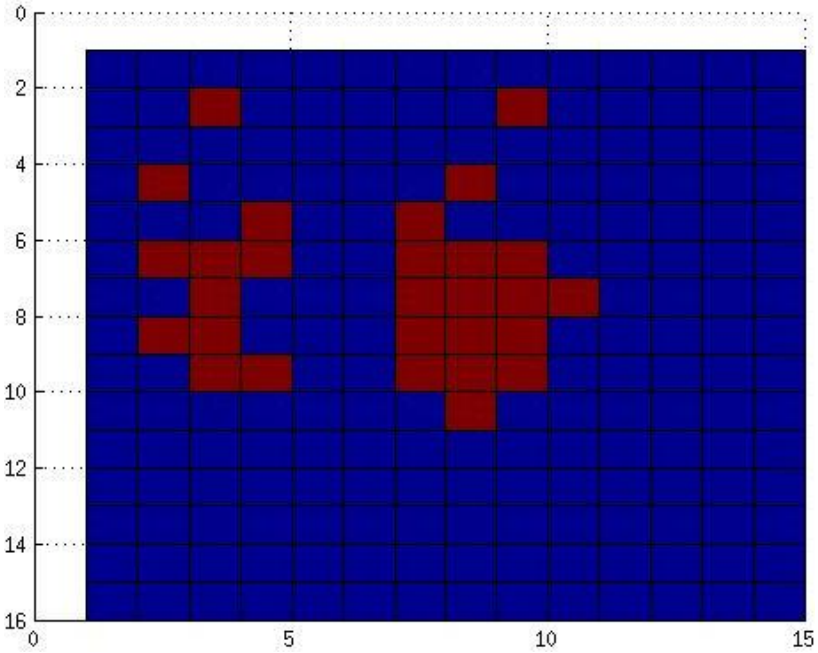
Sapling Random Forest
max vs min



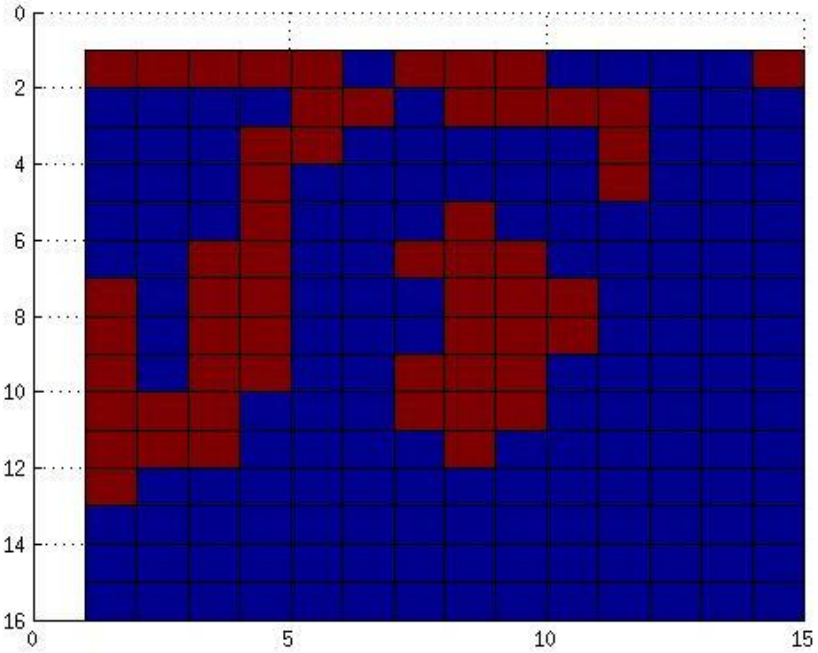
(a) average zero

(b) average one

Sapling Random Forest
max vs min

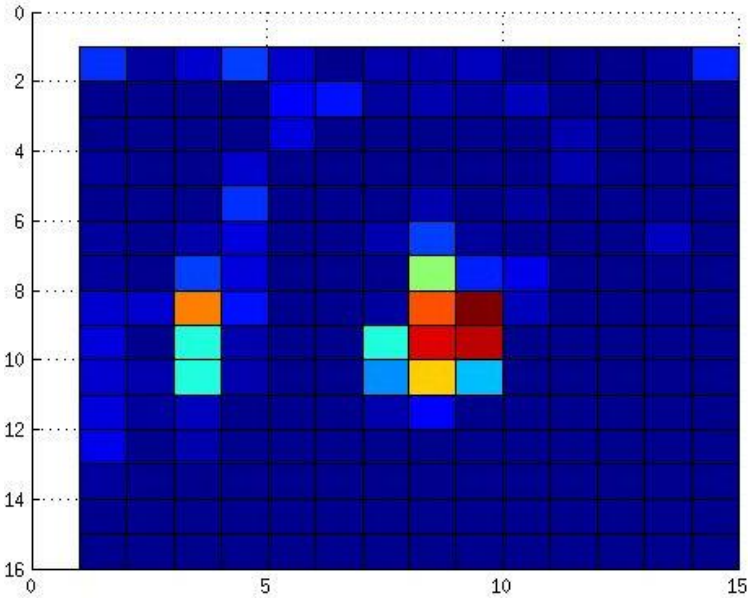


(a) minimal explanation

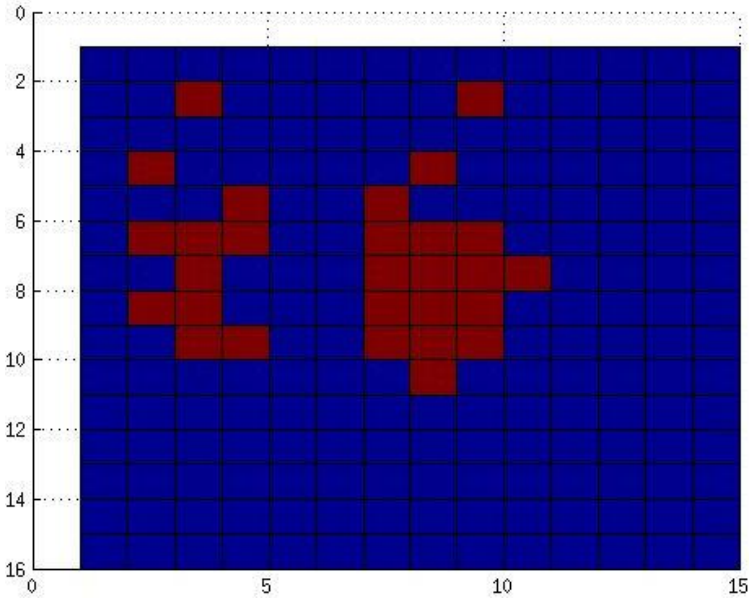


(b) maximal explanation

Sapling Random Forest
max vs min



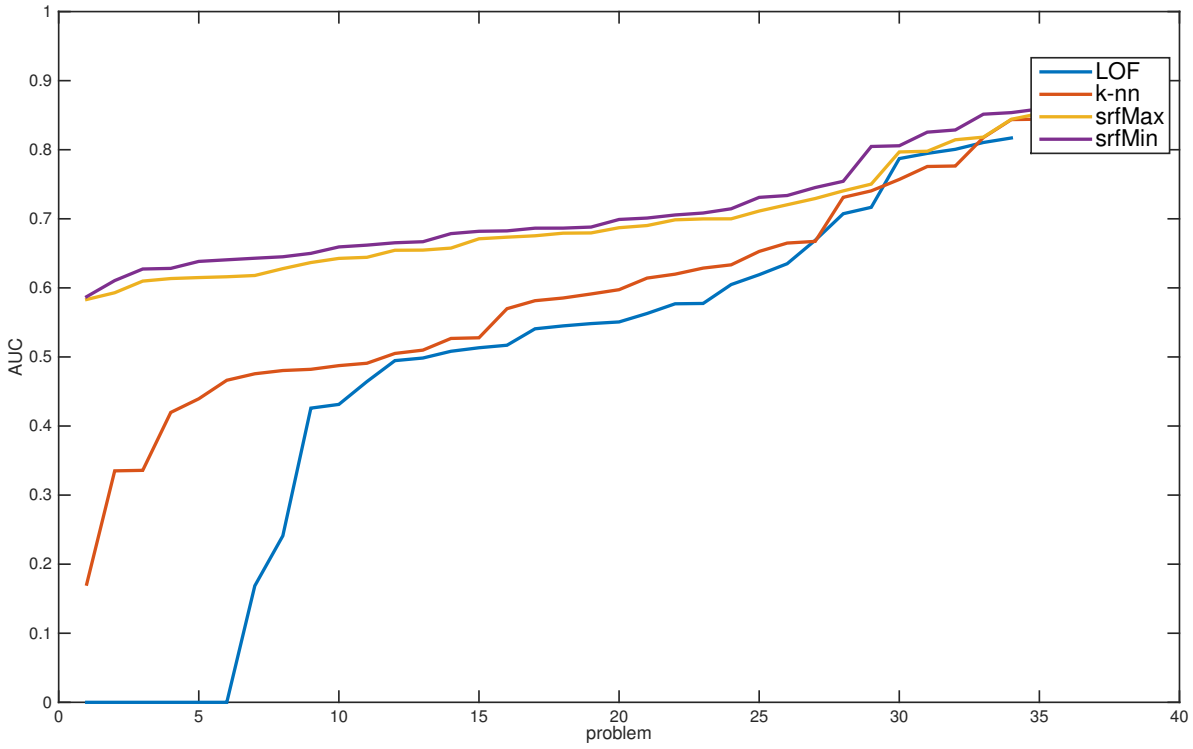
(a) maximal explanation relevance



(b) minimal explanation

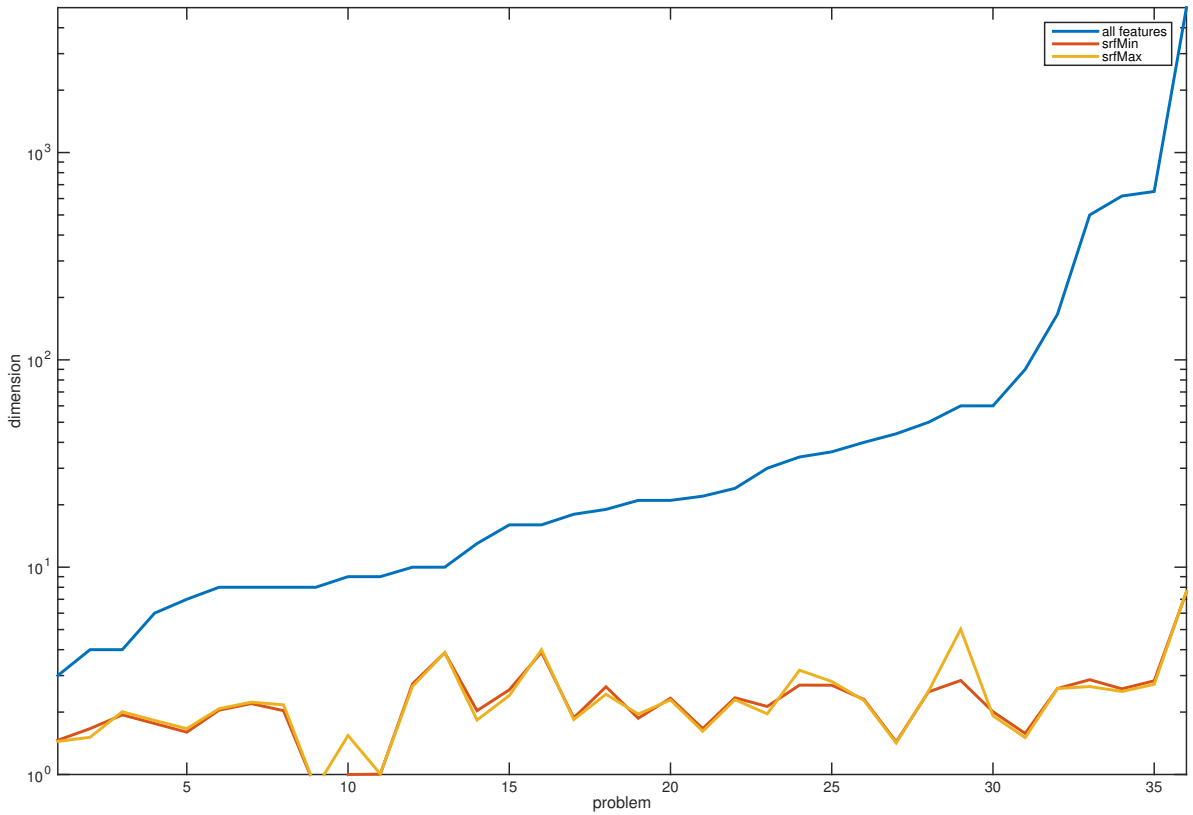
Sapling Random Forest
results

Anomaly explanation as feature selection



Sapling Random Forest
results

Anomaly explanation as feature selection



Outline

Clustering
motivation

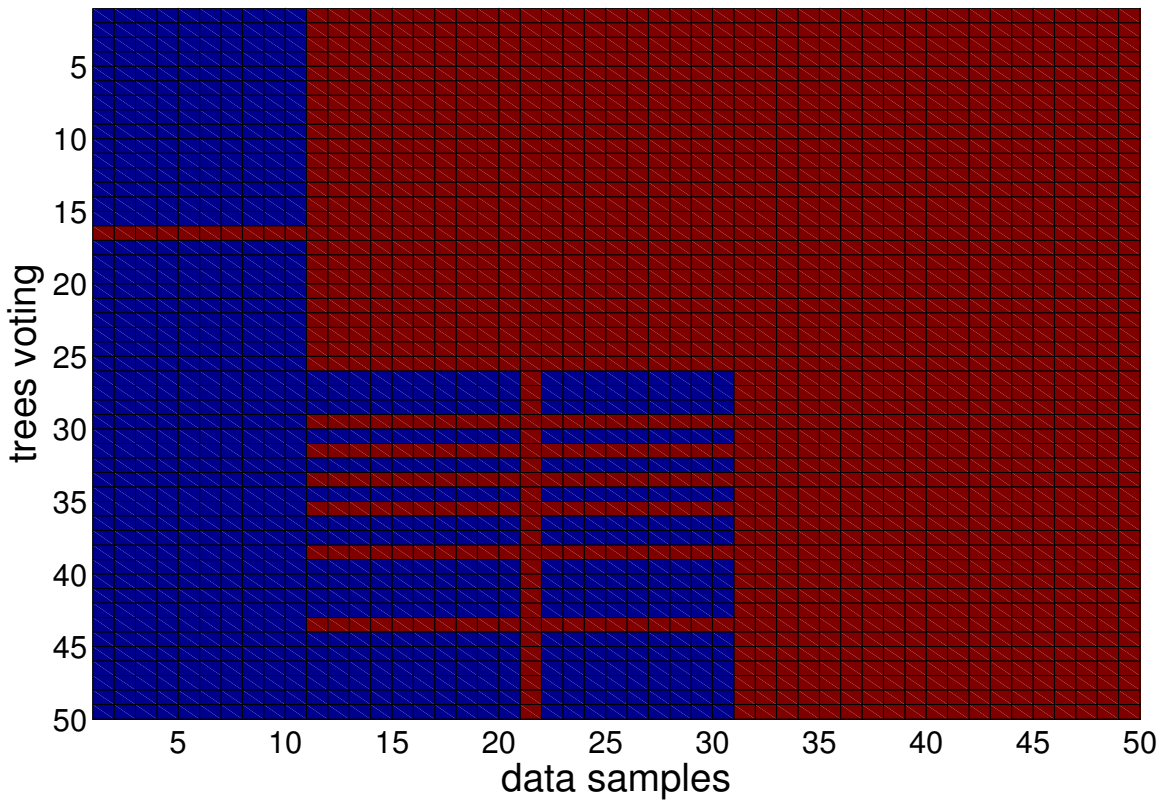
- Investigation of multiple anomalies at once
- Generalized anomaly groups
- Discovery of large scale anomalies
- Domain knowledge

Clustering
Voting vectors

- binary vector
- tree voting
- TxA matrix
- sapling are anomaly specific
- sapling votes for similar anomalies

Clustering
Voting vectors

Example of voting vectors

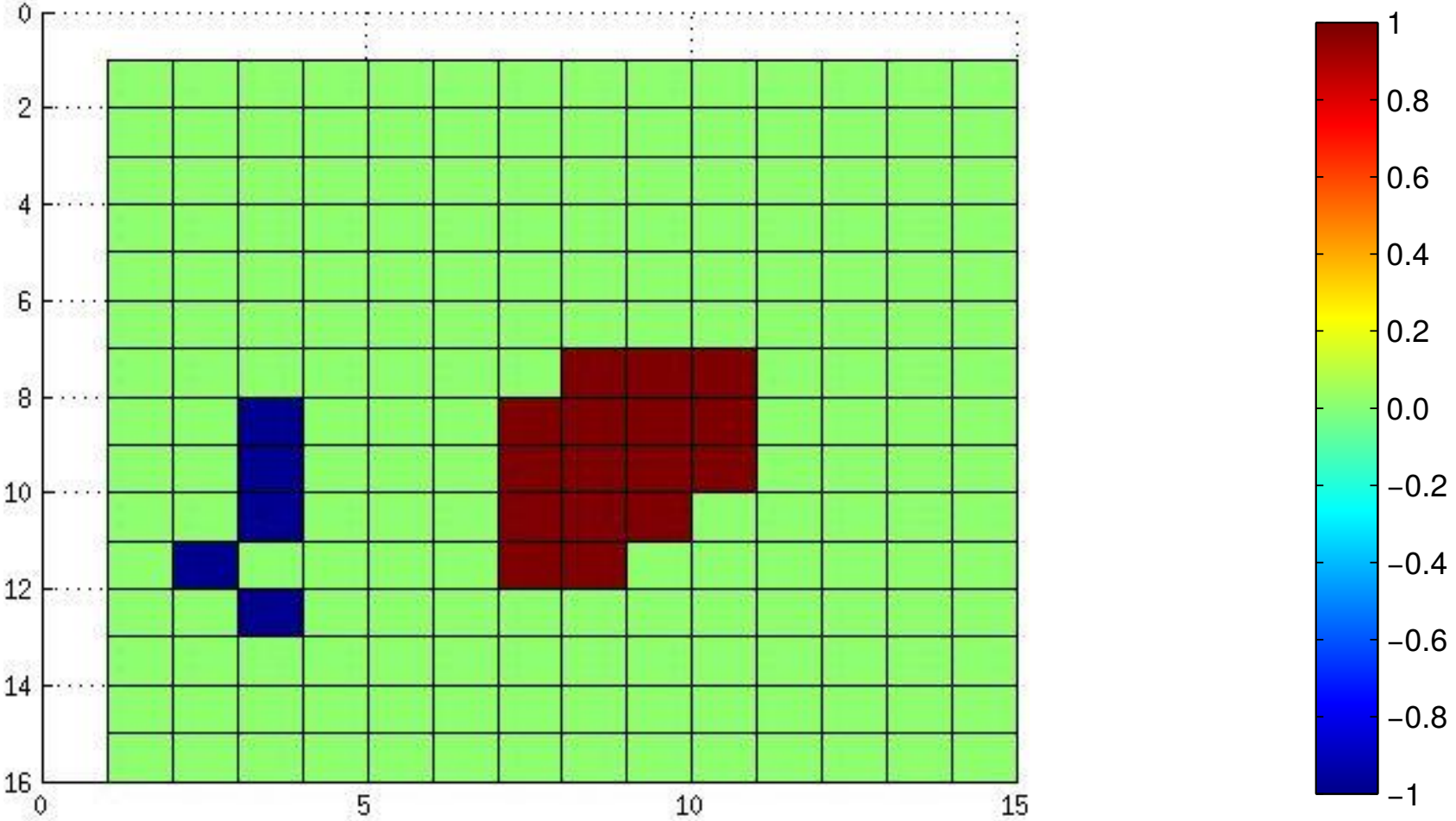


Clustering
Features deviation matrix

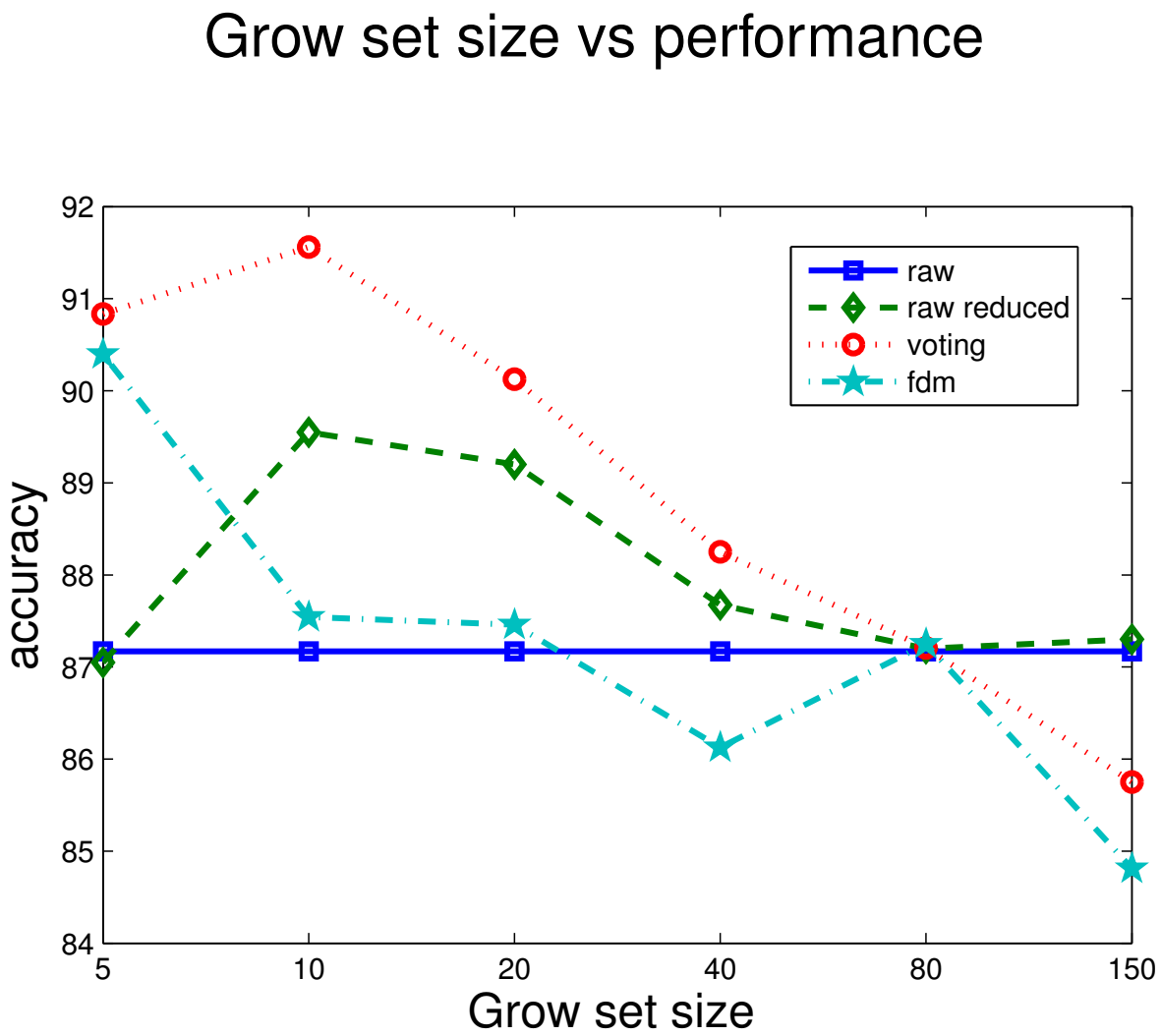
- deviation in feature ranges
- the most strict threshold is stored
- lower and upper boundary
- Tx2d matrix, but can be reduced

Clustering
Voting

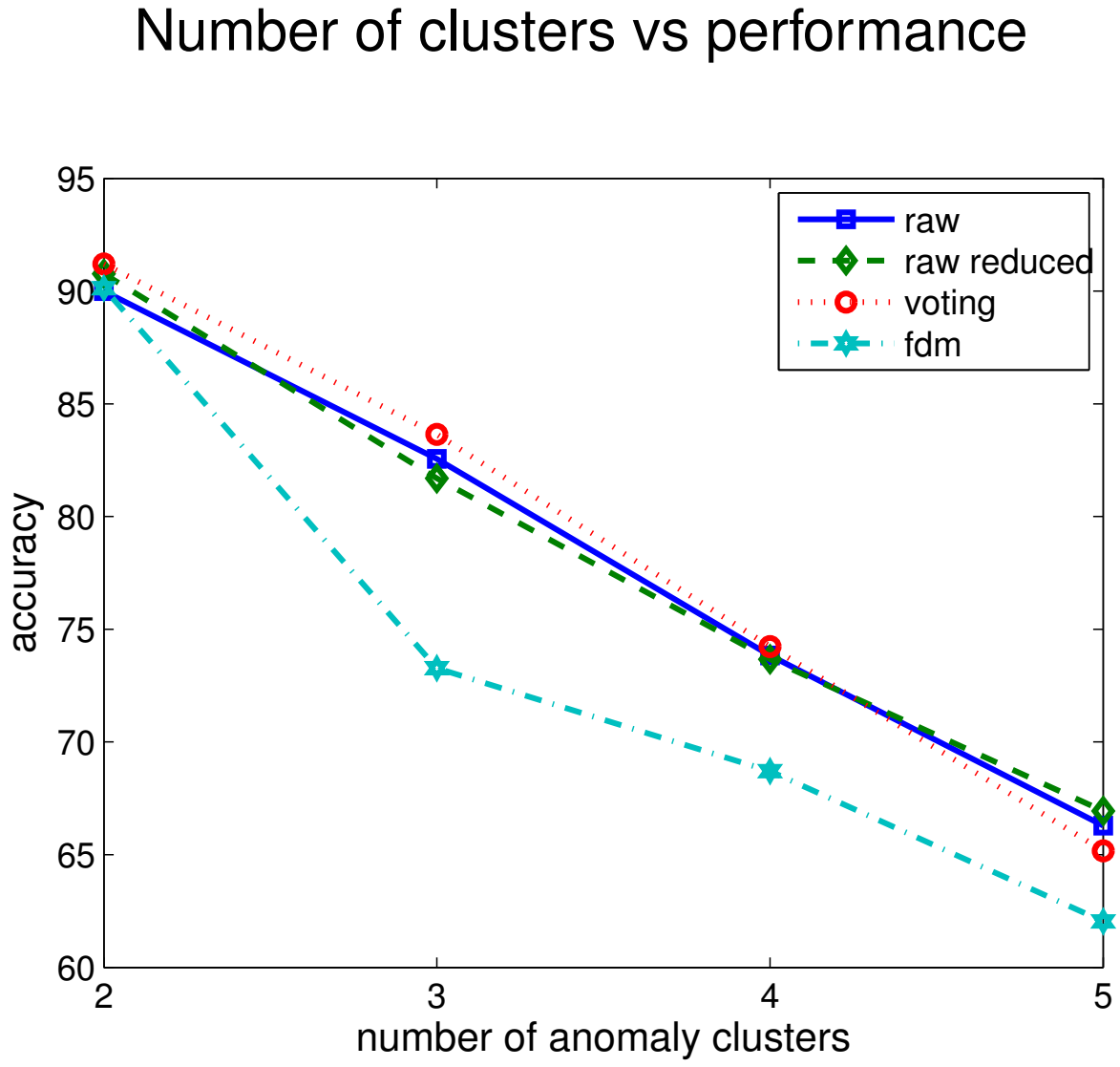
Example of features deviation matrix



Clustering
results



Clustering
results



Conclusion and future work

Conclusion

- anomaly explanation
  - most important features
  - human readable rules
- arbitrary anomaly detector
- real time/data streams

Future work

- multi-dimensional anomalies
- cluster rules aggregation
- fuzzy rules

Thank you for your attention.