Random Histogram Forest for Unsupervised Anomaly Detection

*__Andrian Putina__, *Mauro Sozio, +Dario Rossi, +José .M. Navarro

*Telecom ParisTech France

+Huawei France

Anomaly Detection

«an observation, which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism»

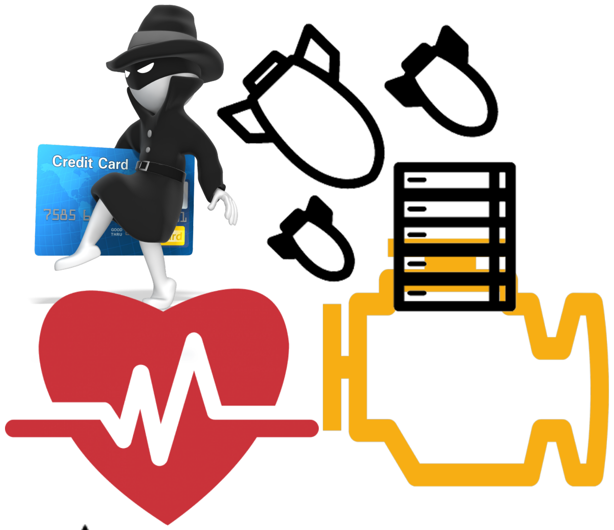
Hawkins

anomaly detection is the task of identifying data patterns or exceptions that are not inline with what expected

Applications and Characteristics

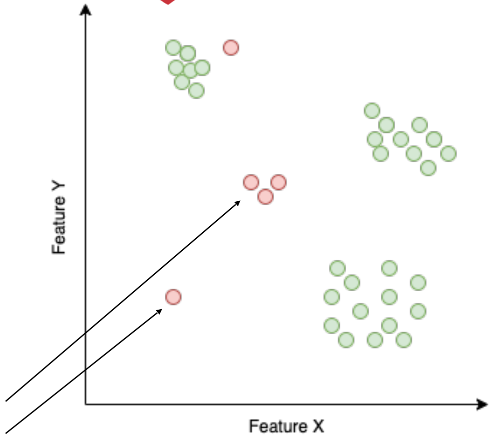
Applications

- Intrusion in computer networks
- Frauds in credit card transactions
- Faults in engines
- Cancerous Masses



Characteristics

- Rare (only small portion of dataset)
- Different from normal instances

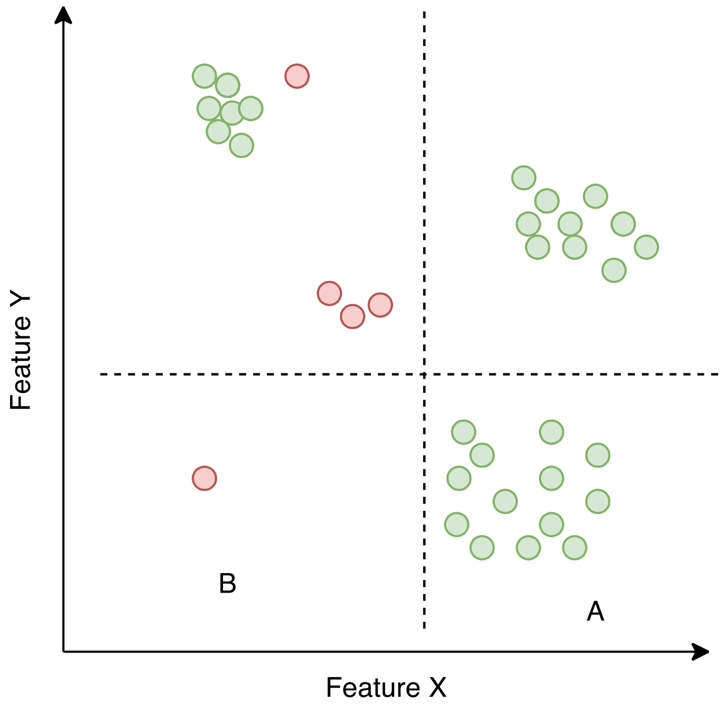


Methods

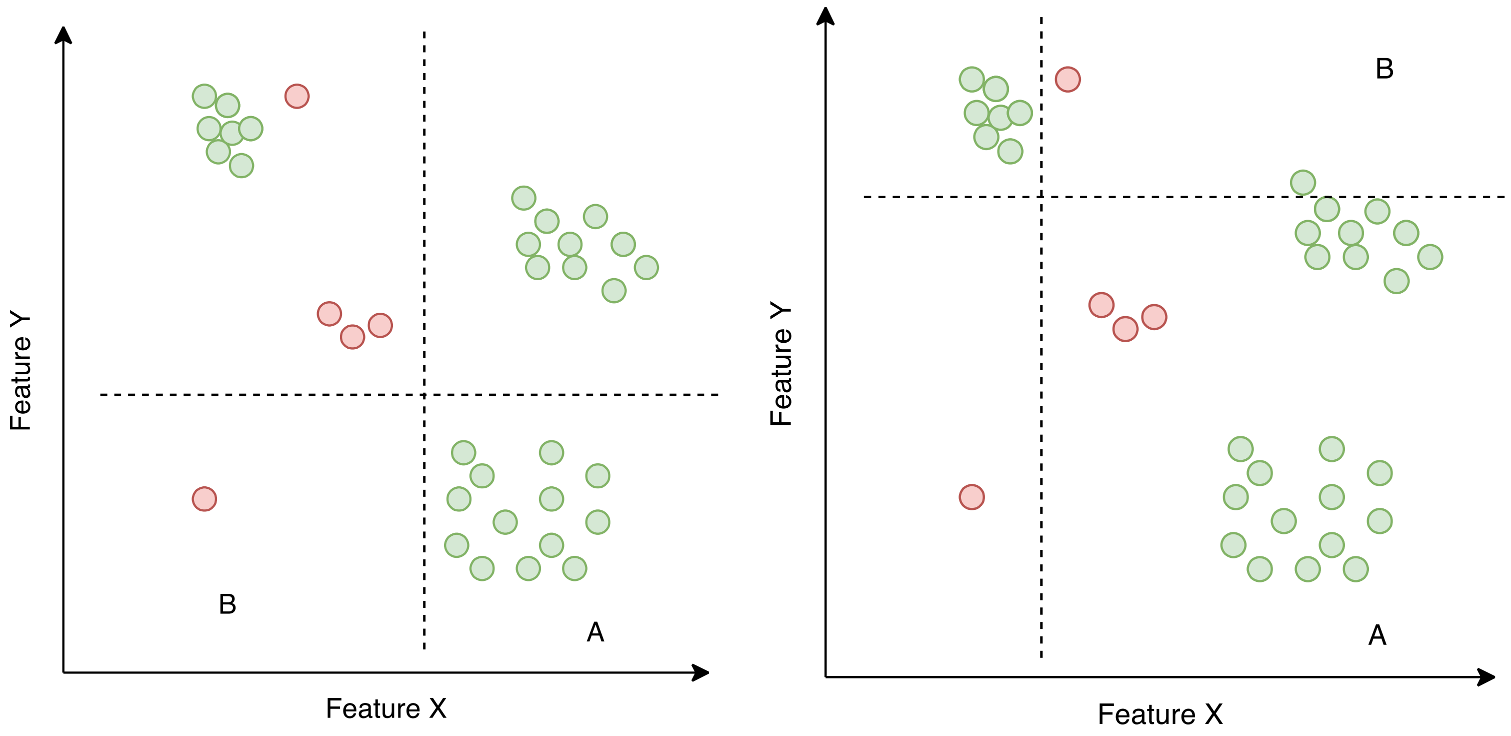
- Probabilistic/Linear (PPCA, OCSVM, etc.)
- Proximity (KNN, LOF, etc.)
- Ensemble (iForest, xStream)

Idea

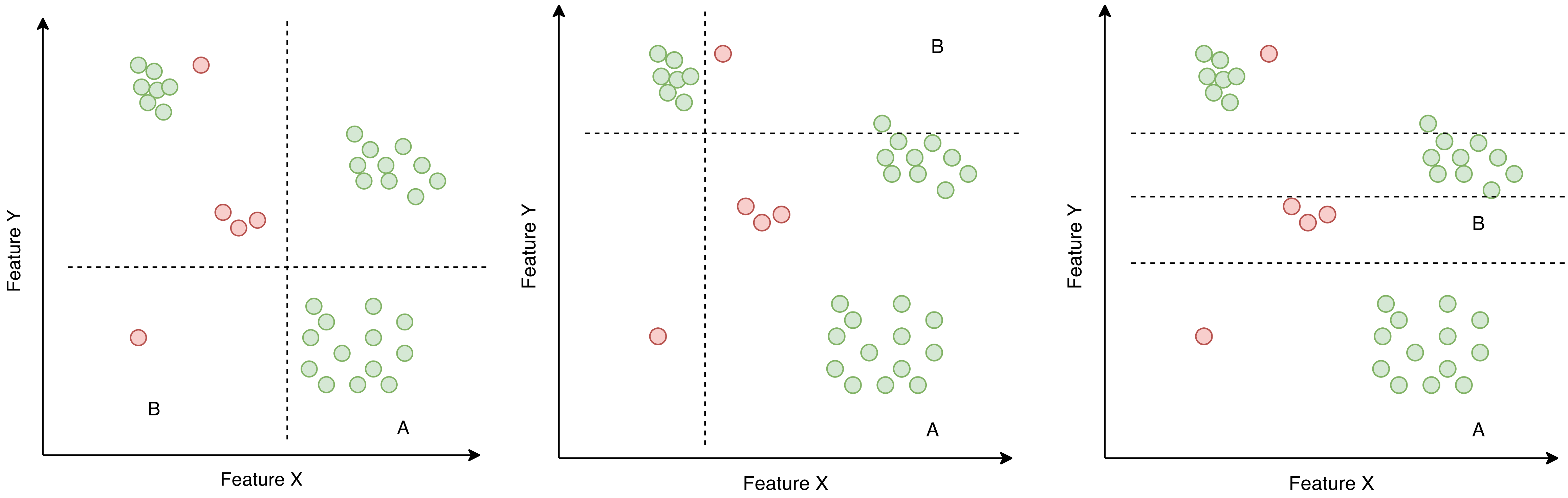
Idea



Idea

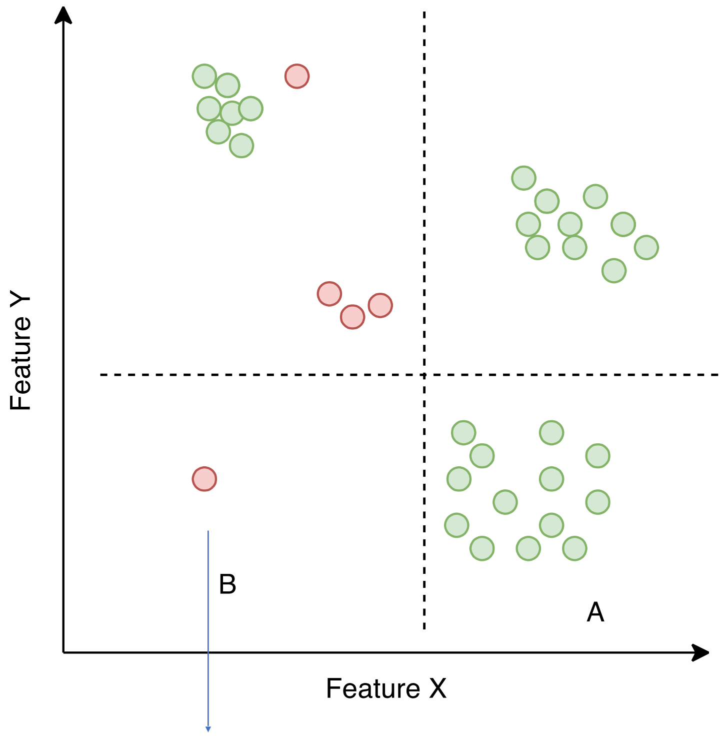


Idea

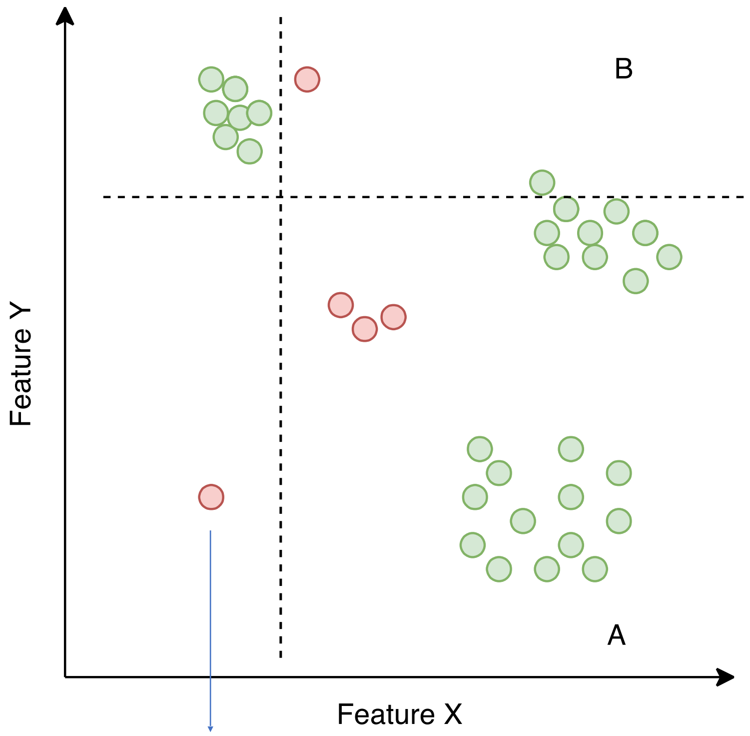


Feature Y
B
A
Feature X
Feature Y
B
A
Feature X
Feature Y
B
A
Feature X

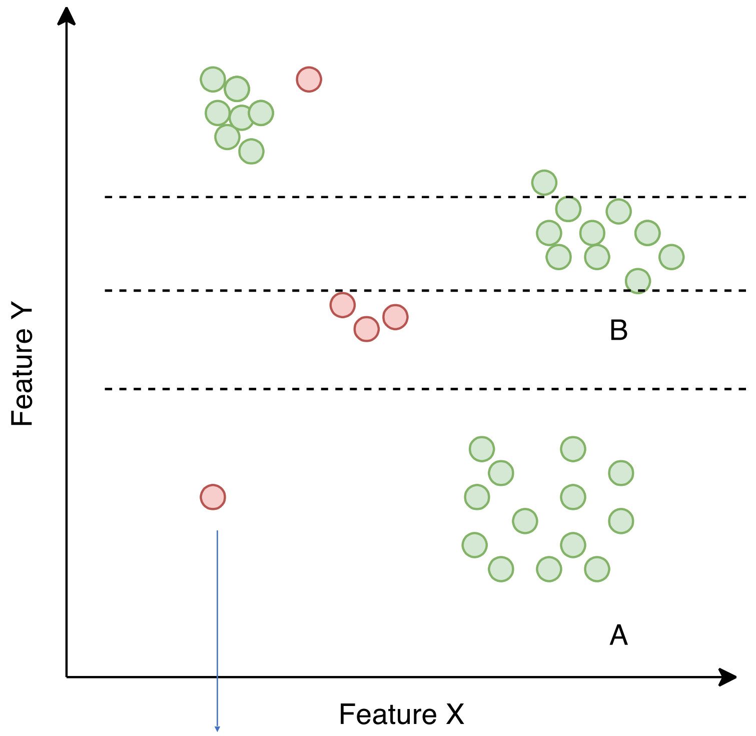
Idea



p = 1/40



p =1/40

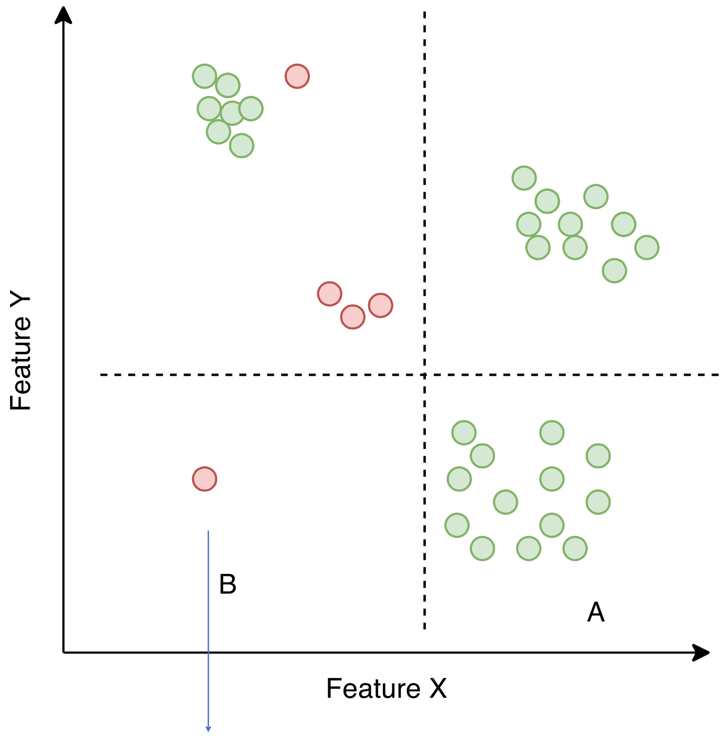


p = 13/40

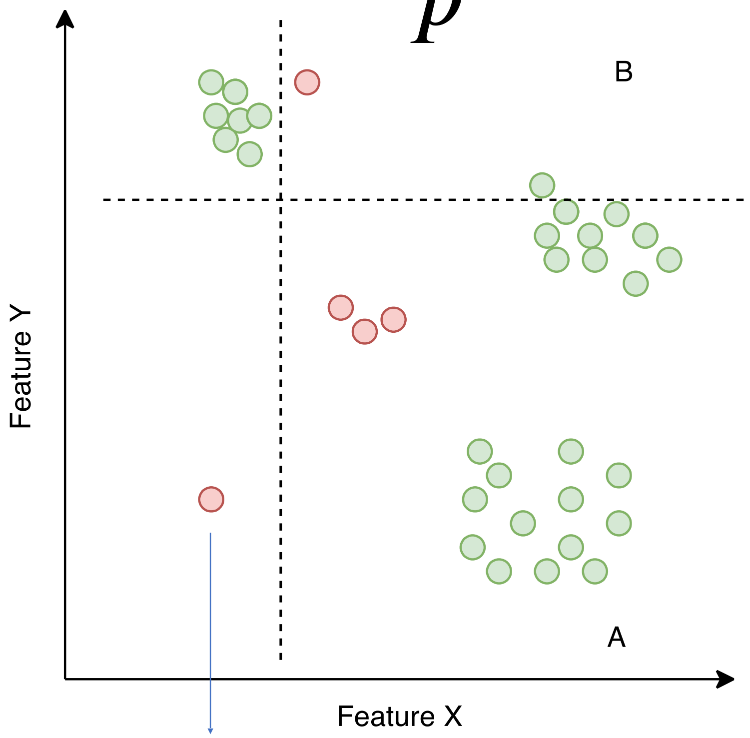
Idea

Information Content

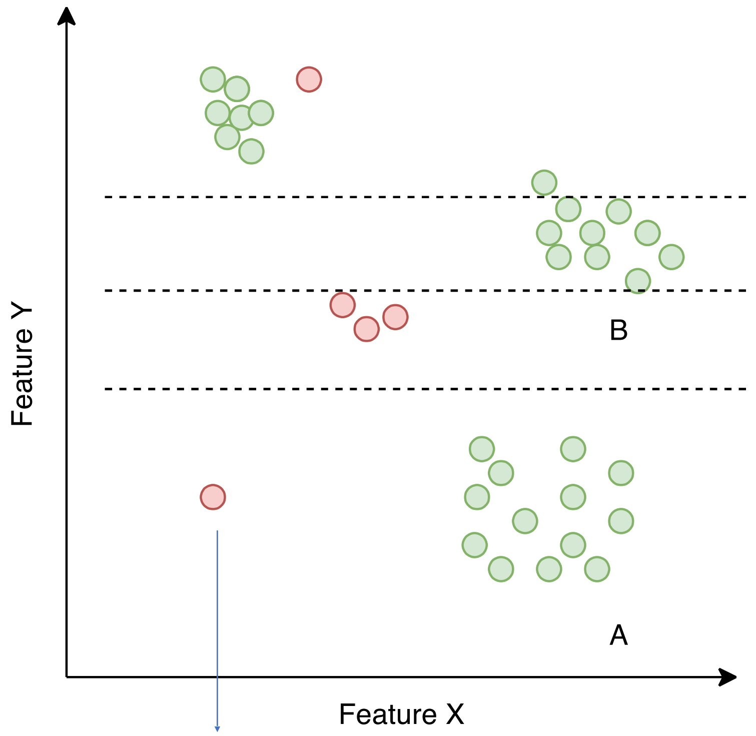
$$\log\frac{1}{p}$$



p = 1/40



p =1/40



p = 13/40

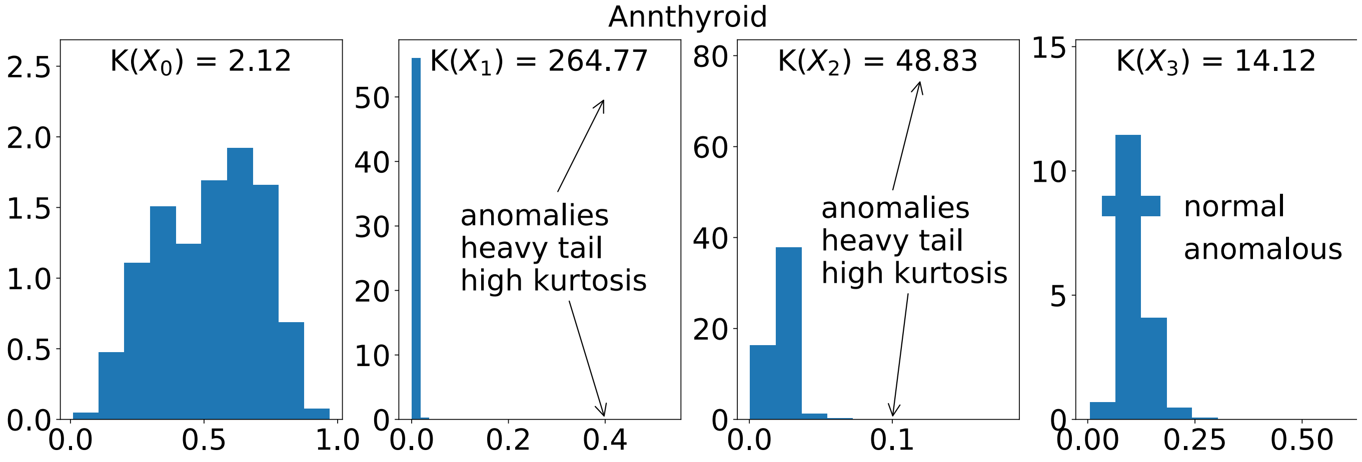
RHF characteristics – Kurtosis Split

- Kurtosis score (tailedness)

$$\mathrm{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

  - 4th moment (standardized data raised to the fourth power)
  - Only values outside the peak region contribute to the kurtosis score

- Features whose Kurtosis is higher are likely to contain separable anomalies.

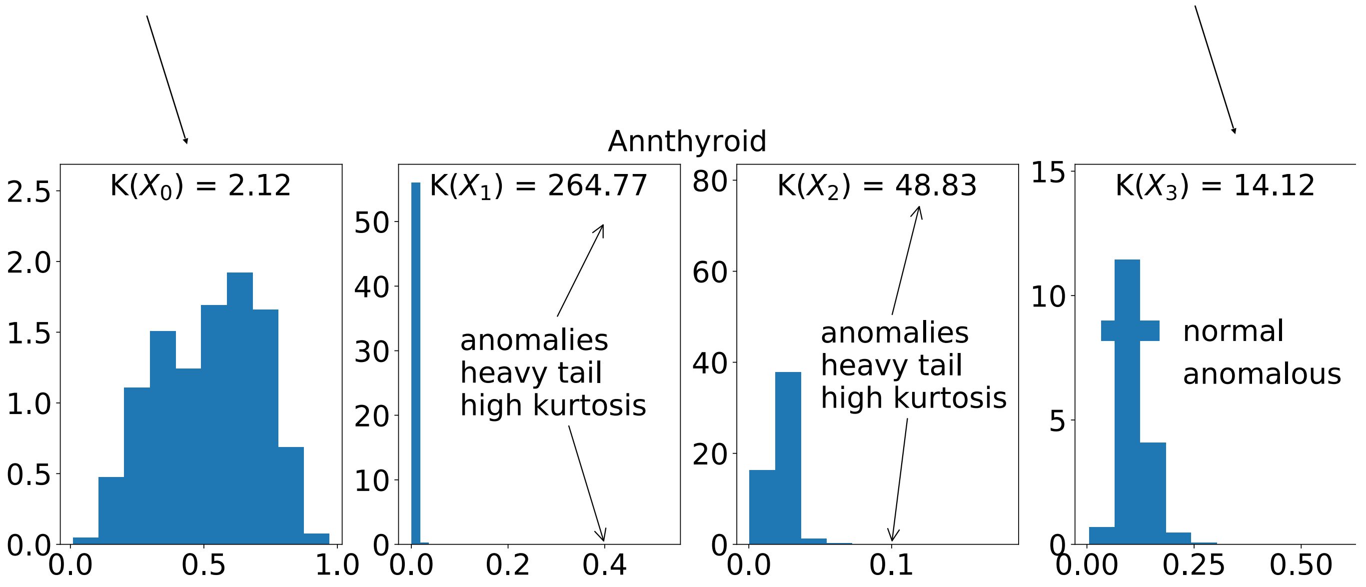


RHF characteristics – Kurtosis Split

- Kurtosis score (tailedness)

$$\mathrm{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

  - 4th moment (standardized data raised to the fourth power)
  - Only values outside the peak region contribute to the kurtosis score



- Features whose Kurtosis is higher are likely to contain separable anomalies.

RHF characteristics – Kurtosis Split

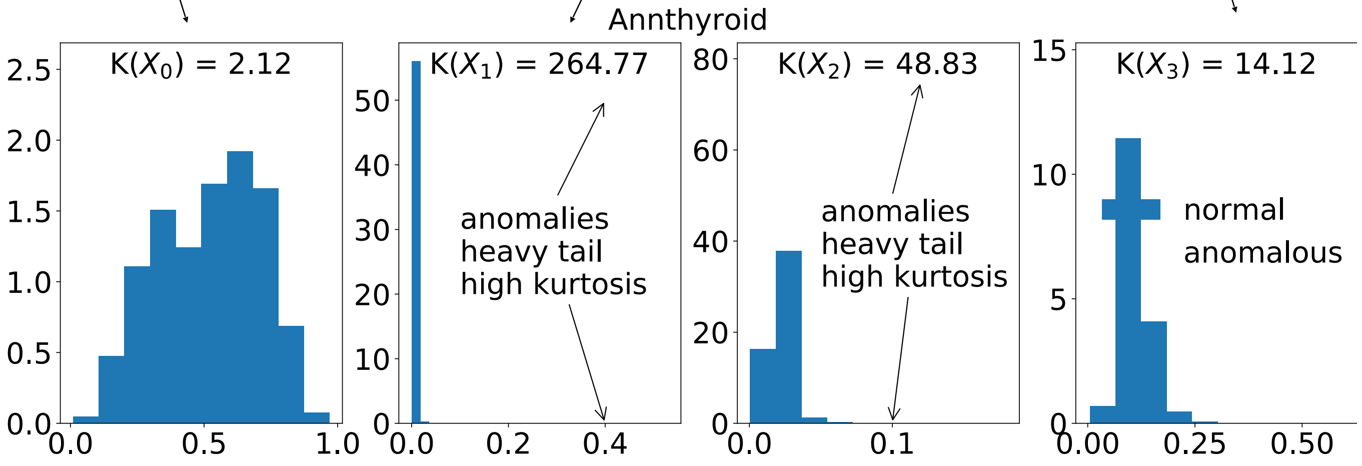
- Kurtosis score (tailedness)

$$\mathrm{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

  - 4th moment (standardized data raised to the fourth power)
  - Only values outside the peak region contribute to the kurtosis score

- Features whose Kurtosis is higher are likely to contain separable anomalies.

Higher kurtosis

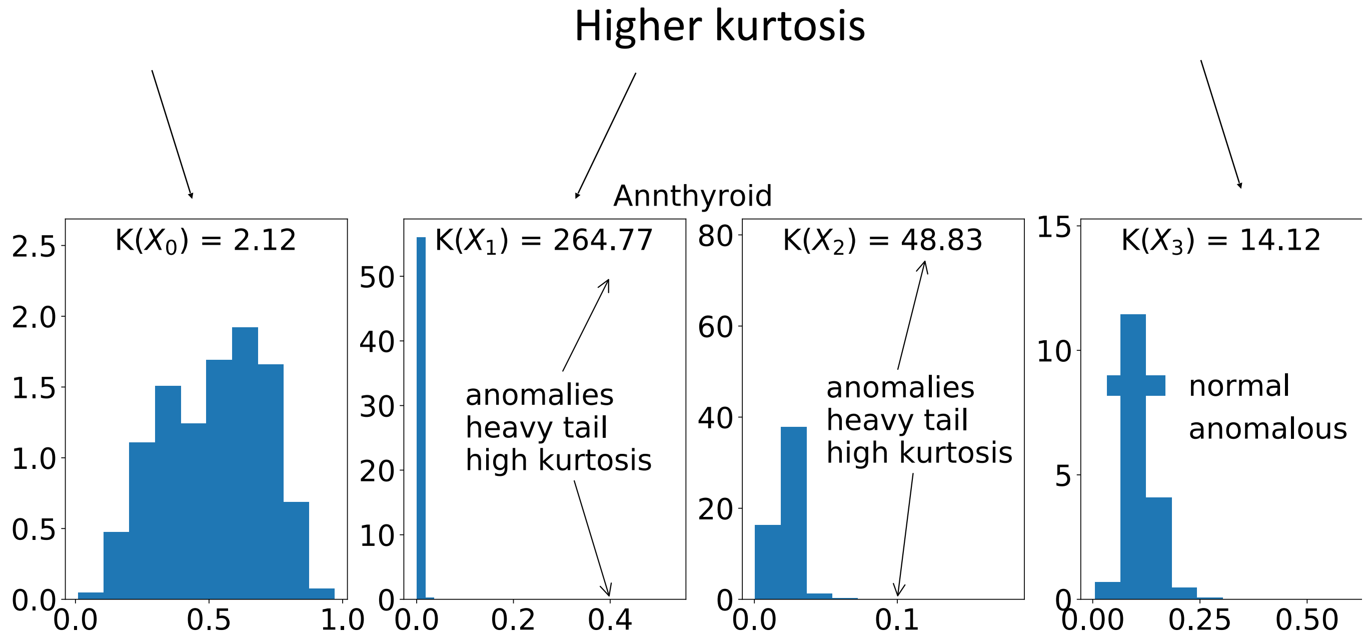


RHF characteristics – Kurtosis Split

- Kurtosis score (tailedness)

$$\mathrm{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

  - 4th moment (standardized data raised to the fourth power)
  - Only values outside the peak region contribute to the kurtosis score



- Features whose Kurtosis is higher are likely to contain separable anomalies.

Let kurtosis guide our search for anomalies!

RHF: Building a tree

Input: A set of points D, max height **h** of the tree T

Output: an anomaly score for each data point

- Compute the kurtosis k(A) of each feature A
- Select a feature A with probability proportional to k(A)
- Let be **a** value u.a.r between the min and max value of A
- Split the data into 2 sets: D_1 with values of A < **a**, D_2 withh values $\geq$ **a**

Recursively apply to D_1 and D_2 until height is **h** or impossible to split anymore

Anomaly Score of p: inversely proportional to # of points in the same leaf in T

RHF: Example

RHF: Example

Max height **h=2**

	A	B
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7
p_5	154	1.2

RHF: Example

Max height **h=2**

	A	B
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7
p_5	154	1.2

RHF: Example

Max height **h=2**

	A	B
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7
p_5	154	1.2

RHF: Example

Max height **h=2**

kur(A)=3.25 **kur(B)=1.72**

	A	**B**
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7
p_5	154	1.2

RHF: Example

Max height **h=2**

kur(A)=3.25 **kur(B)=1.72**

	A	**B**
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7
p_5	154	1.2

RHF: Example

Max height **h=2**

kur(A)=3.25 **kur(B)=1.72**

	A	B
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7
p_5	154	1.2

	A	B
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7

	A	B
p_5	154	1.2

RHF: Example

Max height **h=2**

kur(A)=3.25 **kur(B)=1.72**

	A	**B**
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7
p_5	154	1.2

kur(A)=2.28 **kur(B)=1.84**

	A	**B**
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7

	A	**B**
p_5	154	1.2

RHF: Example

Max height **h=2**

kur(A)=3.25 **kur(B)=1.72**

	A	**B**
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7
p_5	154	1.2

kur(A)=2.28 **kur(B)=1.84**

	A	**B**
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7

	A	**B**
p_5	154	1.2

RHF: Example

Max height **h=2**

kur(A)=3.25 **kur(B)=1.72**

	A	**B**
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7
p_5	154	1.2

kur(A)=2.28 **kur(B)=1.84**

	A	**B**
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7

	A	**B**
p_5	154	1.2

	A	**B**
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6

	A	**B**
p4	5.9	1.7

RHF: Example

Max height **h=2**

kur(A)=3.25 **kur(B)=1.72**

	A	B
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7
p_5	154	1.2

kur(A)=2.28 **kur(B)=1.84**

	A	B
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6
p_4	5.9	1.7

	A	B
p_5	154	1.2

	A	B
p_1	3.9	1.5
p_2	4.2	1.3
p_3	4.0	1.6

	A	B
p4	5.9	1.7

Large number of instances
=>
low anomaly score

$$log\frac{1}{\frac{3}{5}} = 0.5$$

Small number of instances
=>
high anomaly score

$$log\frac{1}{\frac{1}{5}} = 1.6$$

RHF: Overview

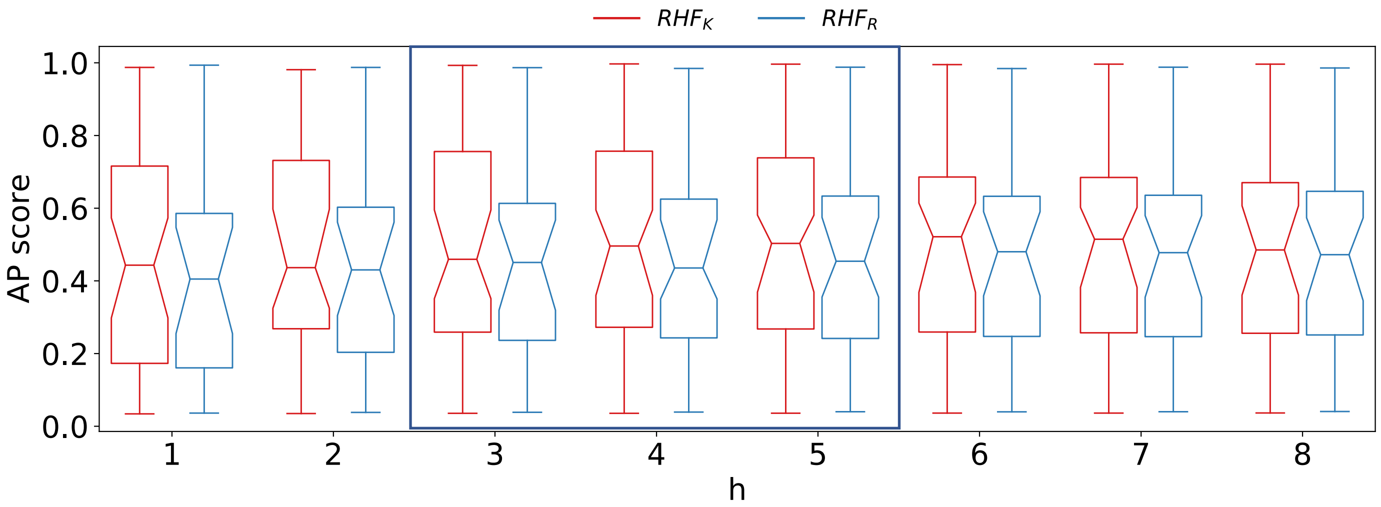
- Build a forest of **t** trees with max height **h**
- Each tree computes an anomaly score for each point in dataset.
- The Anomaly Score is the **Information Content**/**Shannon Information** measuring the level of surprise (rare events more surprising than common ones)
- The final score is aggregated across all the trees

Evaluation - Parameters

- 38 datasets publicly available
    - 240 to 623091 instances
    - 3 to 274 dimensions
    - 0.4% to 10% anomalies

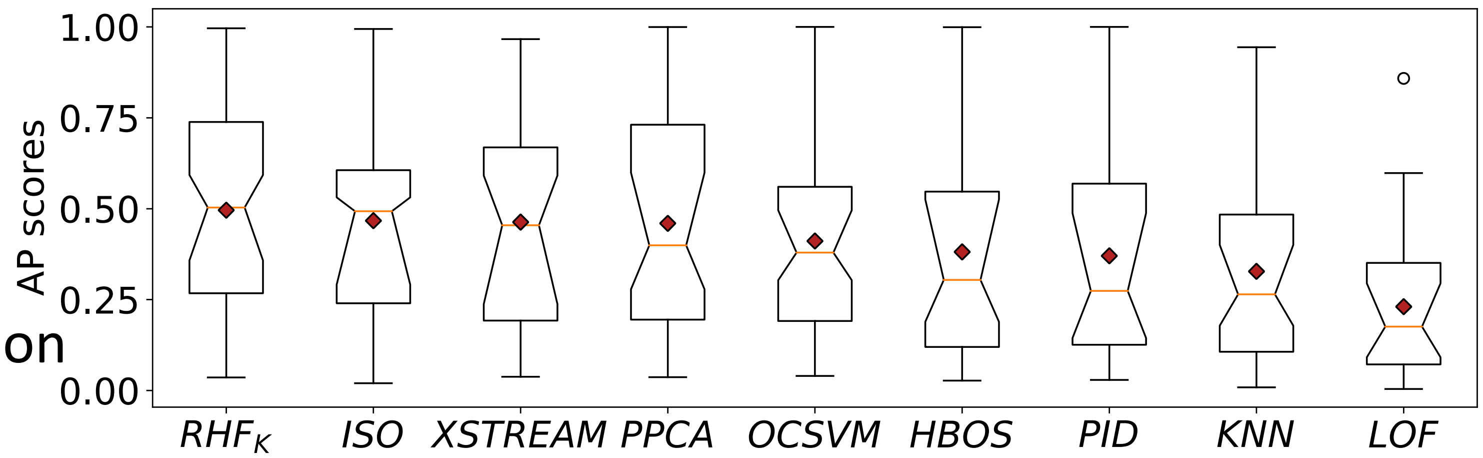
- Average Precision (AP) score:
    - $AP = \sum_n (R_n - R_{n-1}) P_n$
    - $P_n = \frac{tp}{tp + fp}, R_n = \frac{tp}{tp + fn}$ at nth threshold

- Parameters tuning
    - Kurtosis better than random split
    - Max height h produce consistently good results for different values
    - Max height in line with Sturge's formula k = 1 + log2(N)



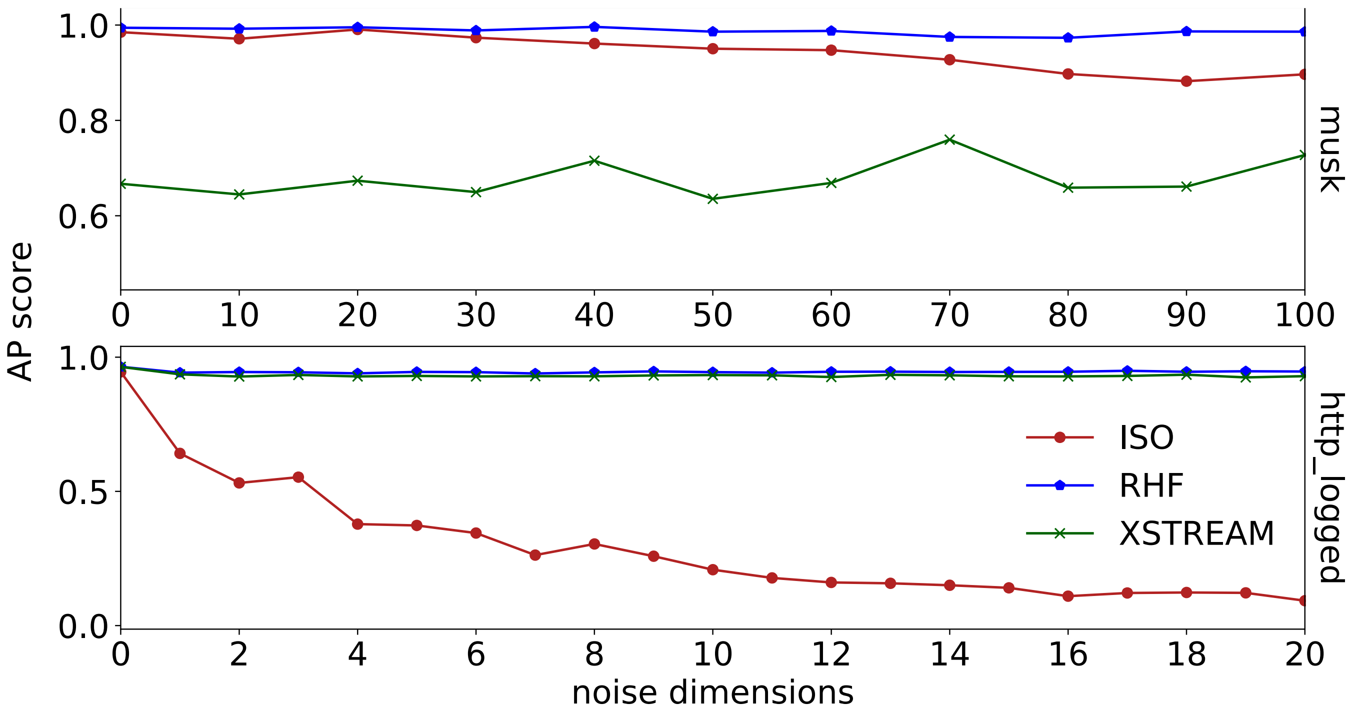
Evaluation - Comparison

- Methods
  - Probabilistic (PPCA, OCSVM, etc.)
  - Proximity (KNN, LOF, etc.)
  - Ensemble (iForest, xStream)
- Top performer
  - $\text{xStream} = 0.453 \pm 0.098$
  - $iForest = 0.463 \pm 0.098$
  - $RHF = \mathbf{0.513} \pm \mathbf{0.010}$
- High discrepancy wrt competitors on some datasets.
  - kdd_http_distinct 0.01 vs **0.74**
  - kdd99G 0.53 vs **0.77**
  - mulcross 0.56 vs **0.73**
  - Musk 0.65 **vs 0.99**



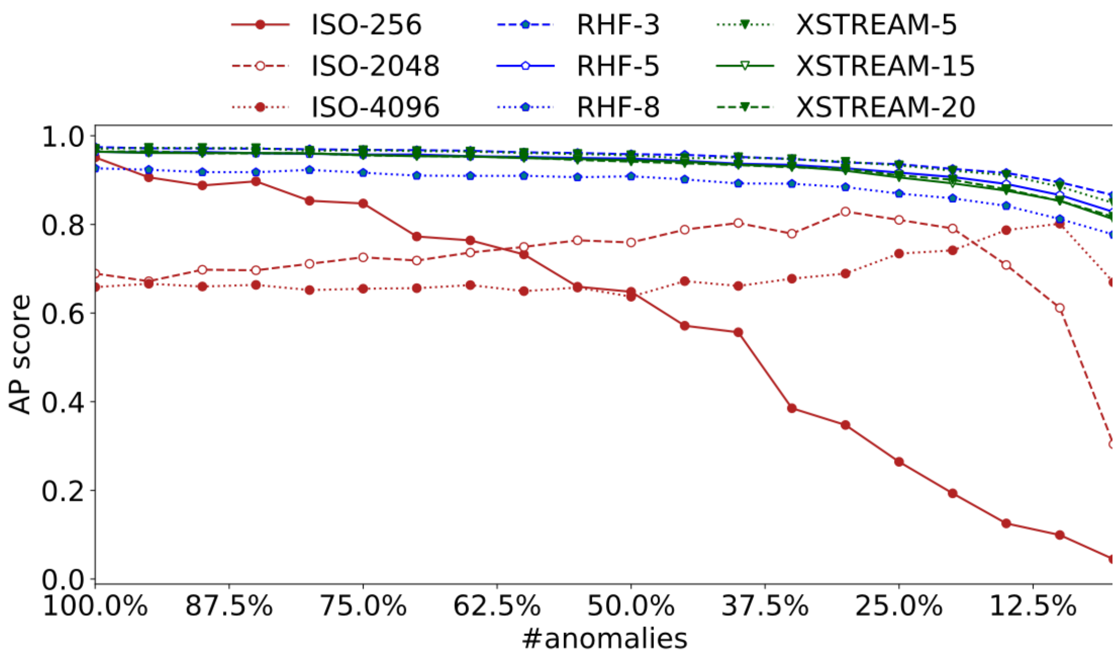
Evaluation – Irrelevant features

- High dimensional data
- Irrelevant dimensions
- Gaussian noise
- Robustness
  - **RHF** = Kurtosis
  - **xStream** = Random Projections

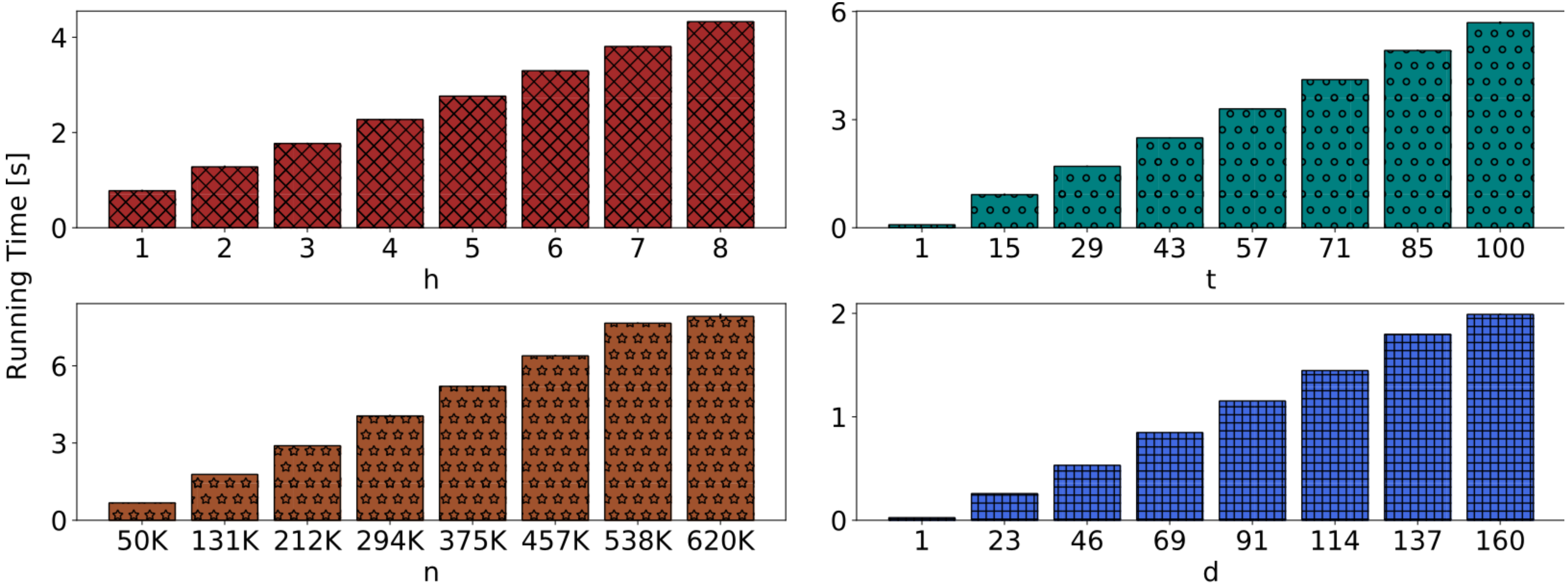


Evaluation – vary #anomalies

- Impact on input parameter
- Vary #anomalies into the dataset
  - 565287 normal instances
  - 2211 anomalous instances (100%)
  - 100 anomalous instances (5%)
- Isolation (2nd best performing) shows overfitting effects in the public benchmark dataset
- RHF (1st) and xStream (2nd) perform well also on **private** datasets



Running time



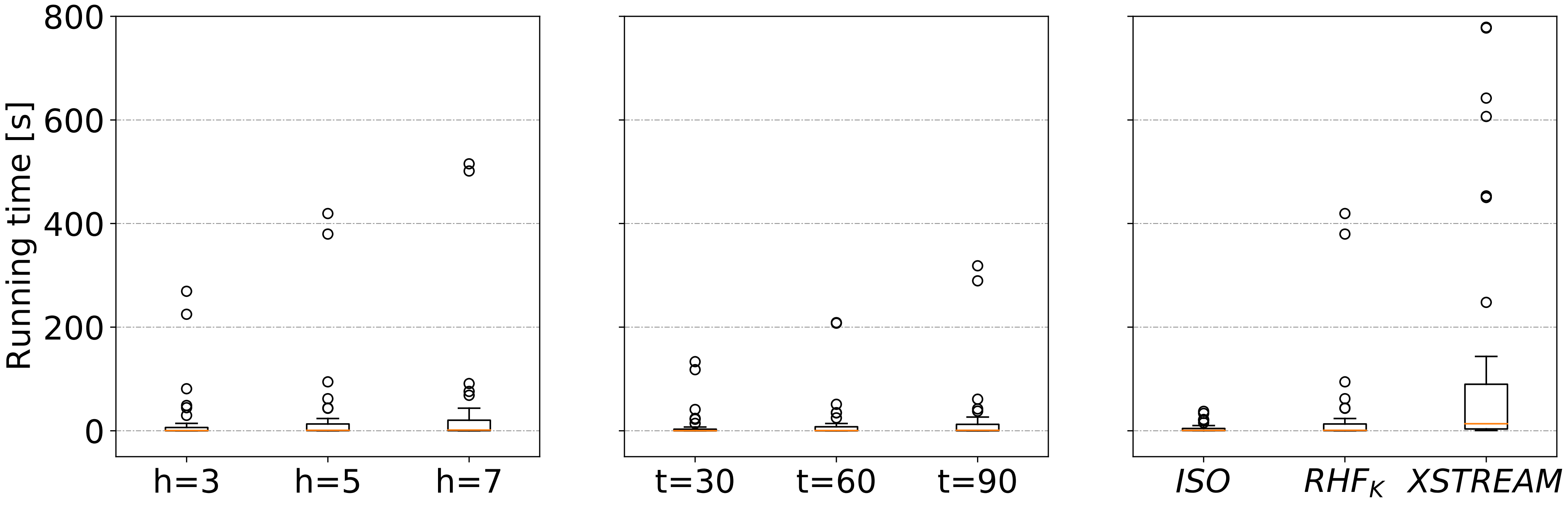
Linearly increasing in n, d, h, t

Conclusions

- Best performing one on 38 datasets
  - 10% better on avg/median
  - Better than a factor of 2 in many datasets
  - Large gap in some datasets (0.75 vs 0.01)
- Robust to inner parameter selection
- Robust to irrelevant features
- Linear running time in input size
- Produces results that are easy to interpret and explain

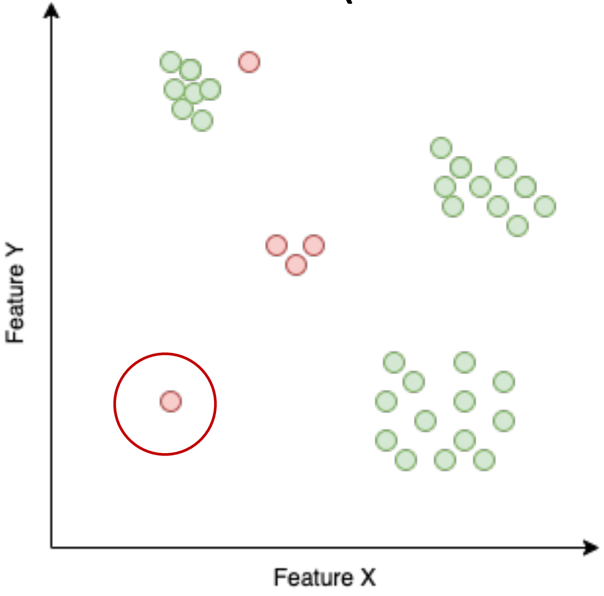
Backup Slides

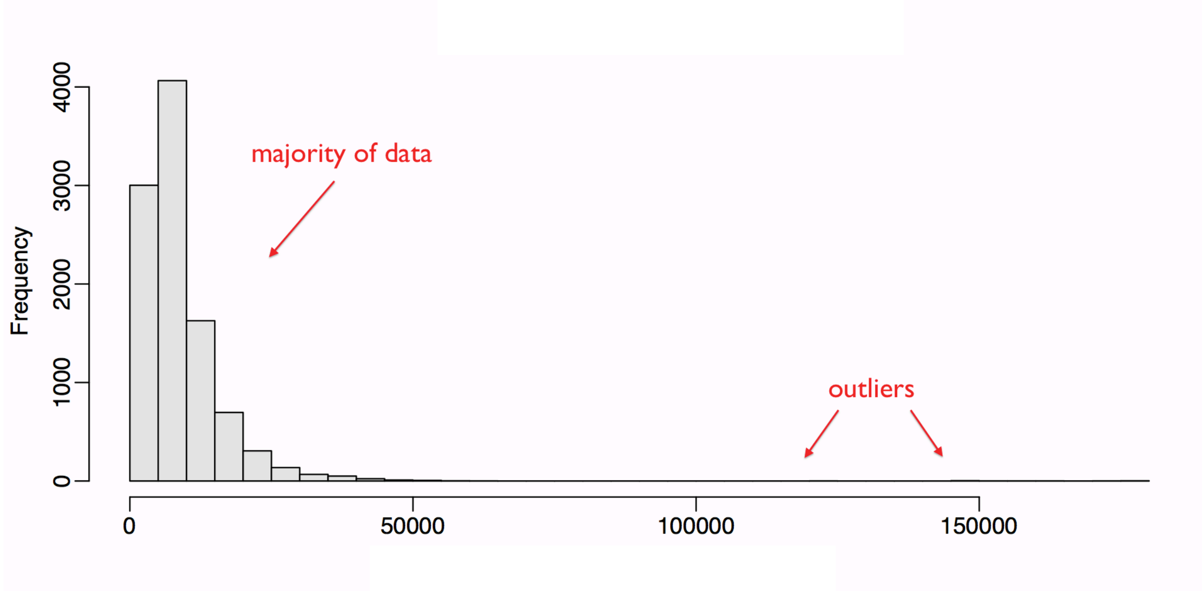
Running Time



Model characteristics

- Anomalies
  - Rare (low probability and high information)
  - Different (skewed data distribution)



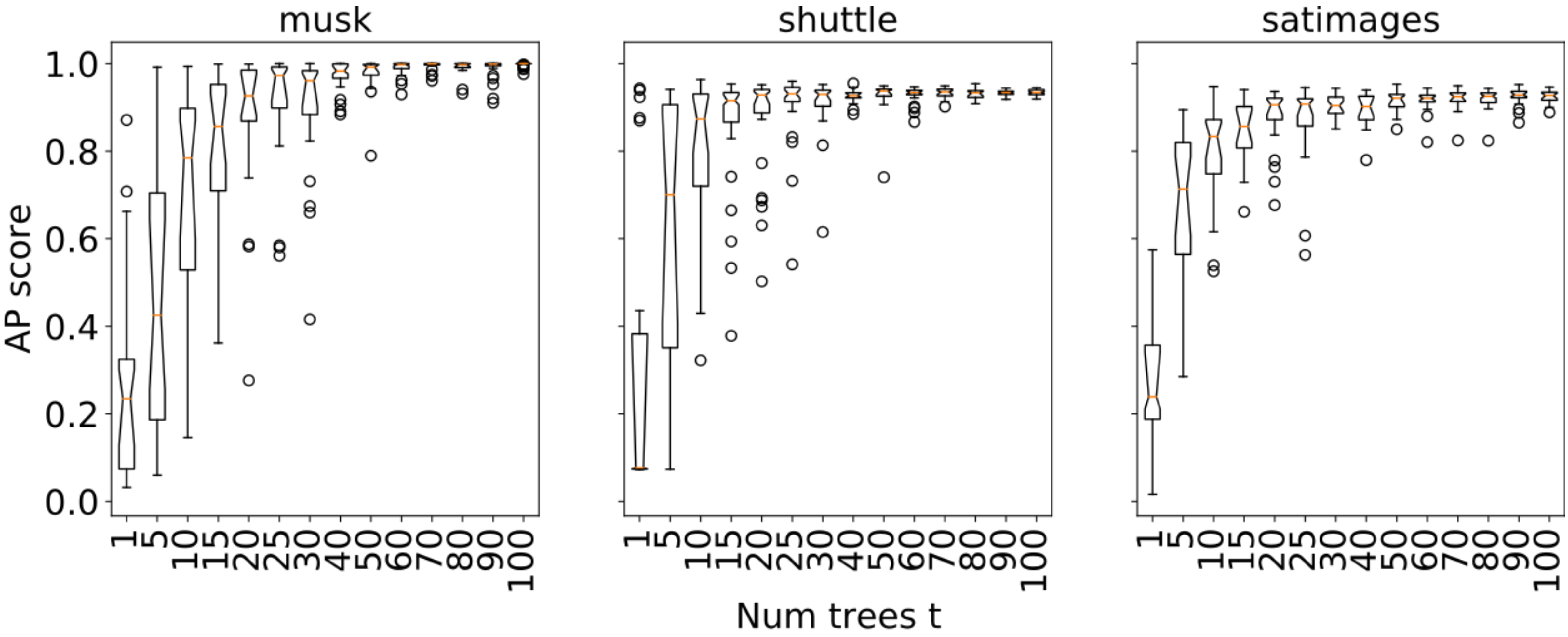


Kurtosis Split

$$K_s = \sum_{a=0}^{d} log\left[K\left(X_a\right) + 1\right]$$

$$r = \mathcal{X} \sim U[0, K_s]$$

$$a_s = argmin\left(i|\sum_{a=0}^{i} log\left[K\left(X_i\right) + 1\right] > r\right)$$


musk
shuttle
satimages
AP score
1.0
0.8
0.6
0.4
0.2
0.0
1
5
10
15
20
25
30
40
50
60
70
80
90
100
Num trees t