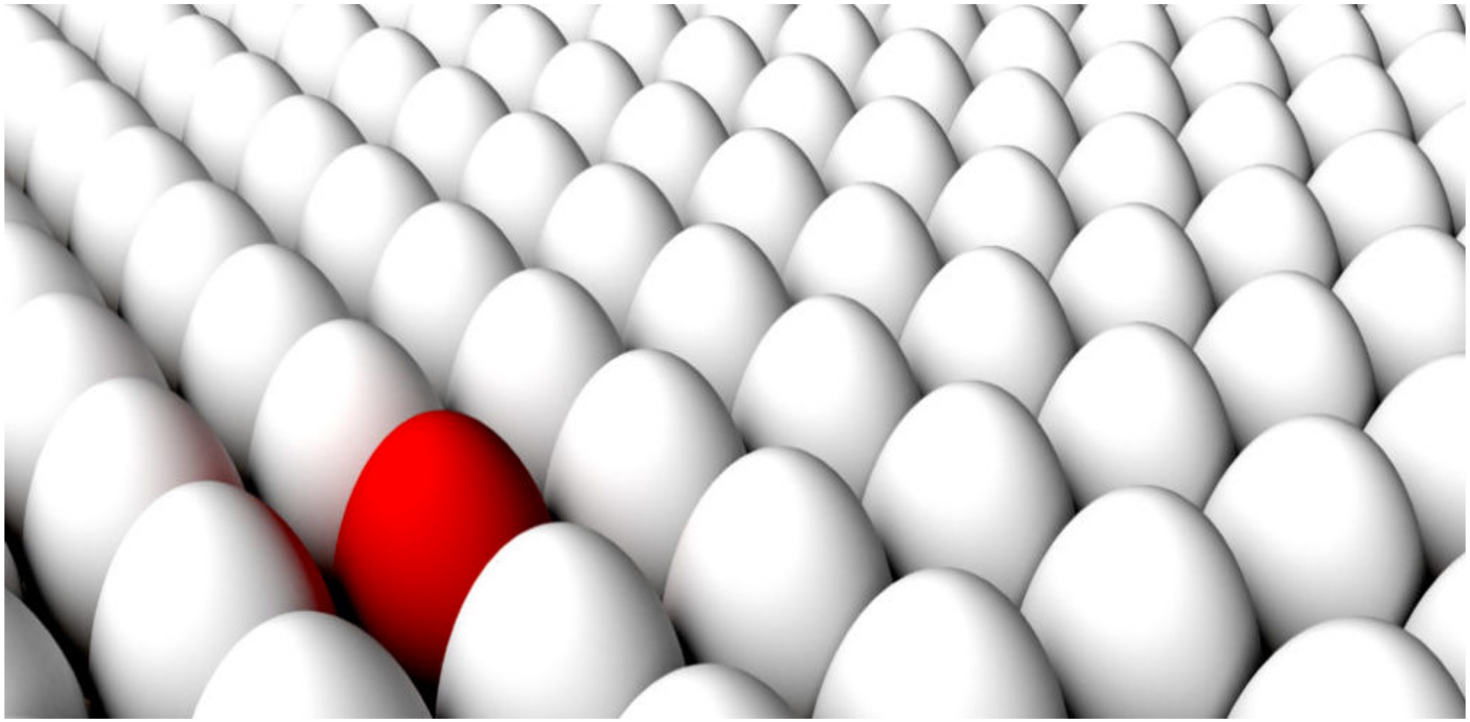
Anomaly Detection



What is anomaly?

- Hawkins' definition of outliers: [1980]

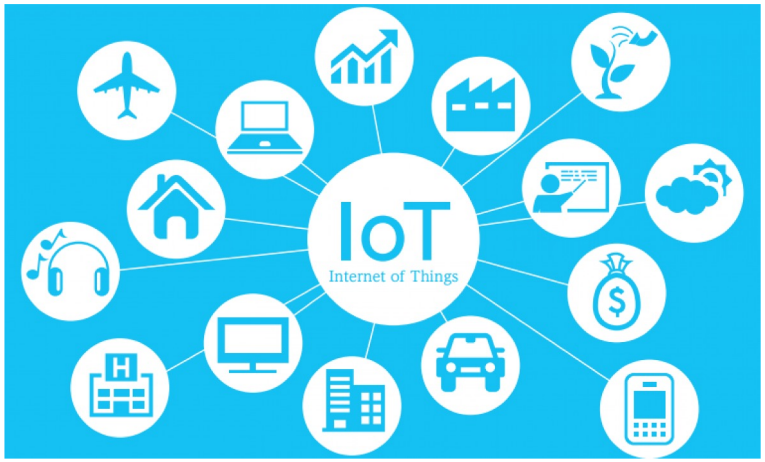
"an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"

- In practice:
  Outliers are determined narrowly by model assumptions
  - One person's noise is another person's signal
  - Most [unsupervised] anomaly detectors report statistical noise as outliers and/or anomalies

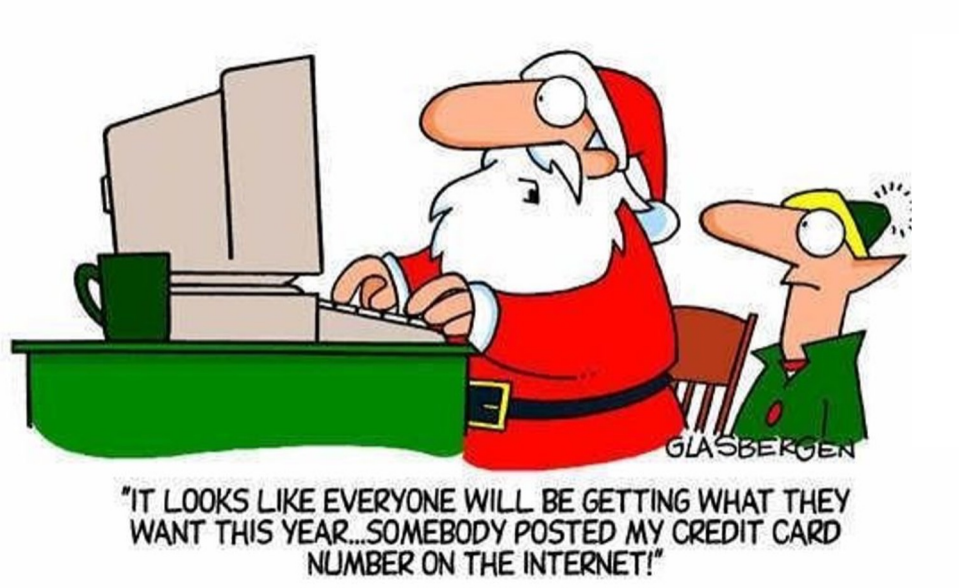
Agenda

- What is anomaly?
- Do we need it?
- Applicability
- Types of anomalies
- Types of detection settings
- Local outlier factor
- Isolation forest
- Loda
- Deep learning based anomaly detectors

Why ‘anomaly detection’ is important?



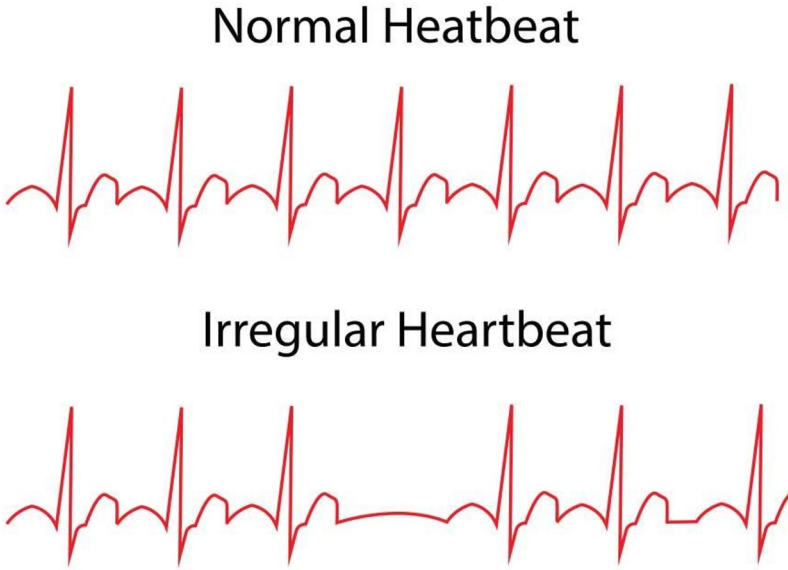
Motivating App #1: Credit Card Fraud

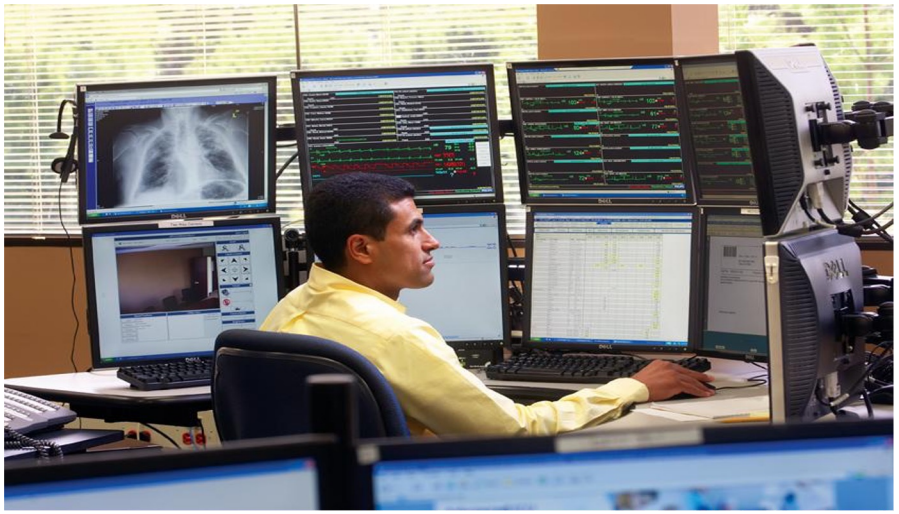


May-22	1:14 pm	FOOD	Monaco Café	$1,127.80	→ Point Anomaly
May-22	2:14 pm	WINE	Wine Bistro	$28.00	
...					
Jun-14	2:14 pm	MISC	Mobil Mart	$75.00	
Jun-14	2:05 pm	MISC	Mobil Mart	$75.00	
Jun-15	2:06 pm	MISC	Mobil Mart	$75.00	
Jun-15	11:49 pm	MISC	Mobil Mart	$75.00	
May-28	6:14 pm	WINE	Acton shop	$31.00	→ Collective Anomaly
May-29	8:39 pm	FOOD	Crossroads	$128.00	
Jun-16	11:14 am	MISC	Mobil Mart	$75.00	
Jun-16	11:49 am	MISC	Mobil Mart	$75.00	

- Unusual transactions
- False alarms are common
- High false positive rate □ wasted human effort

Motivating App #2: ICU Patient Monitoring





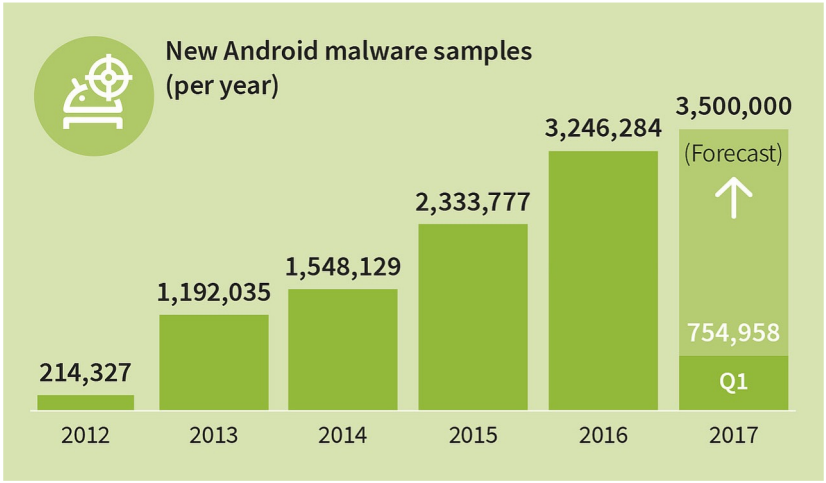
- Monitor patients' condition
- False alarms are common

Motivating App #3: Android Malware

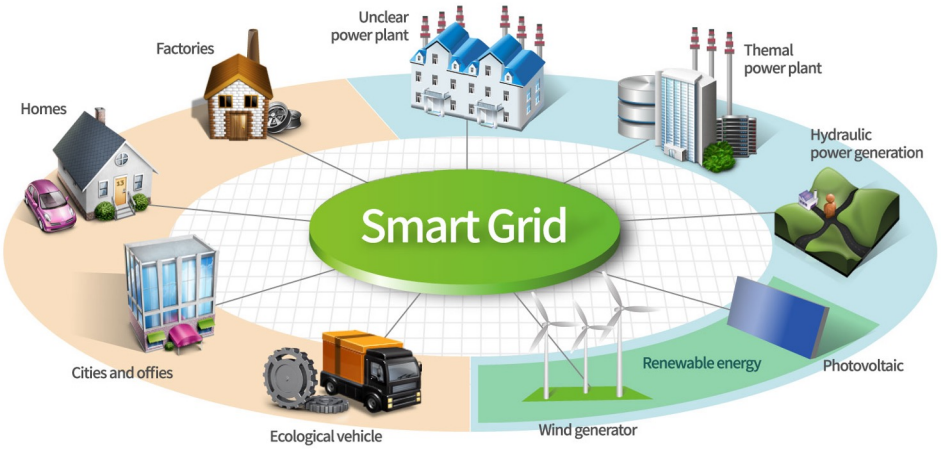
Malwares can potentially

- Spy
- Steal personal information
- Make fake calls
- ….

Anti-Malware Industry

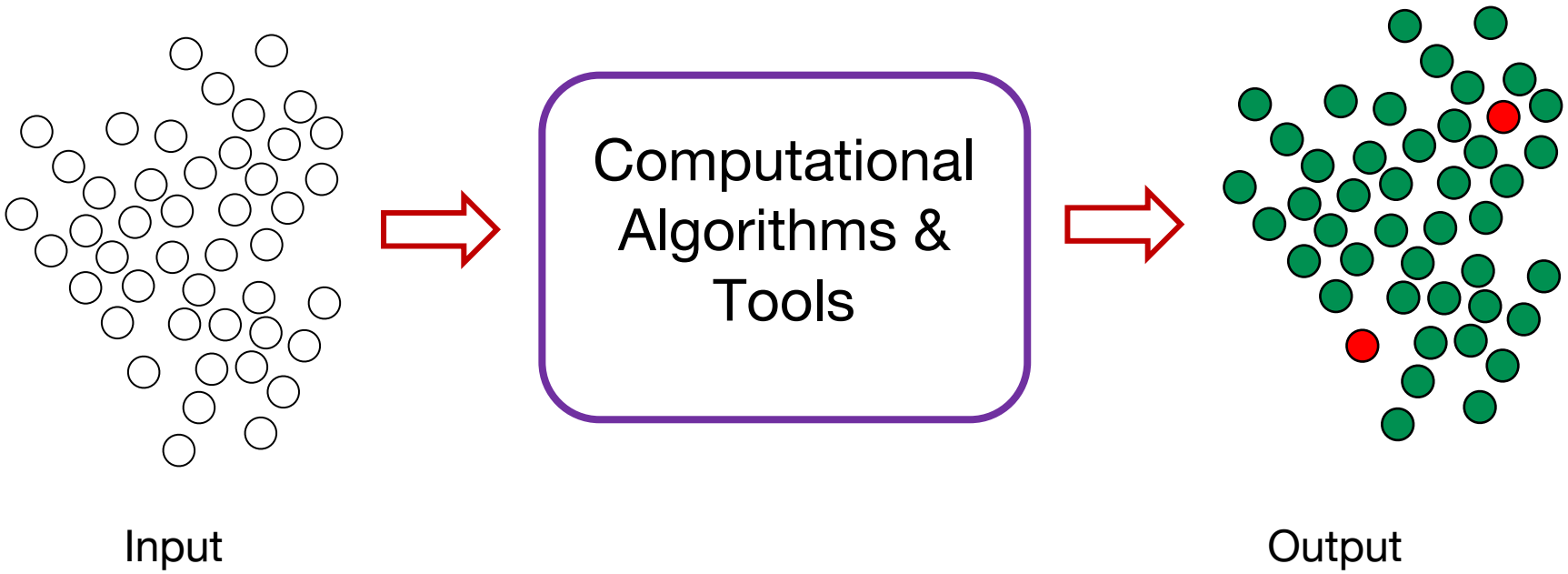


Motivating App #4: Smart Grid Security

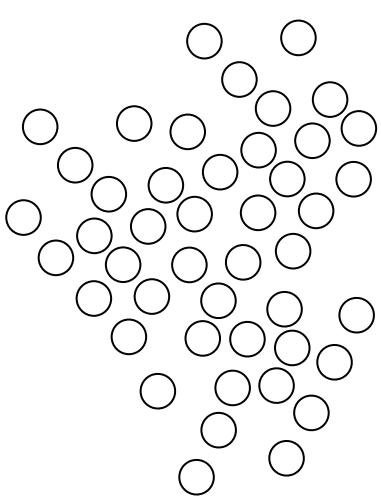


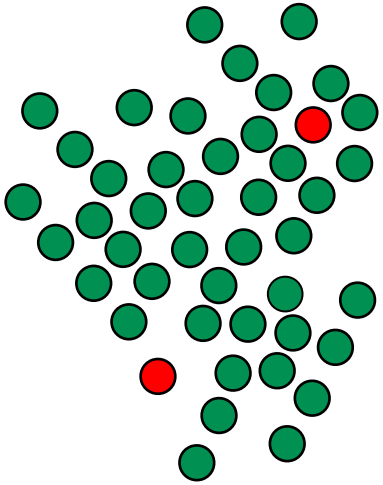
- Complete blackout
- False alarms are common
- High false-positive rate ⟹ wasted human effort

Anomaly Detection: Problem



Anomaly Detection: Challenges





- # of anomalies << # of nominals
- Knowledge discovery task involving humans
- High false-positive rate → wasted human effort

Anomaly Detection: Challenges

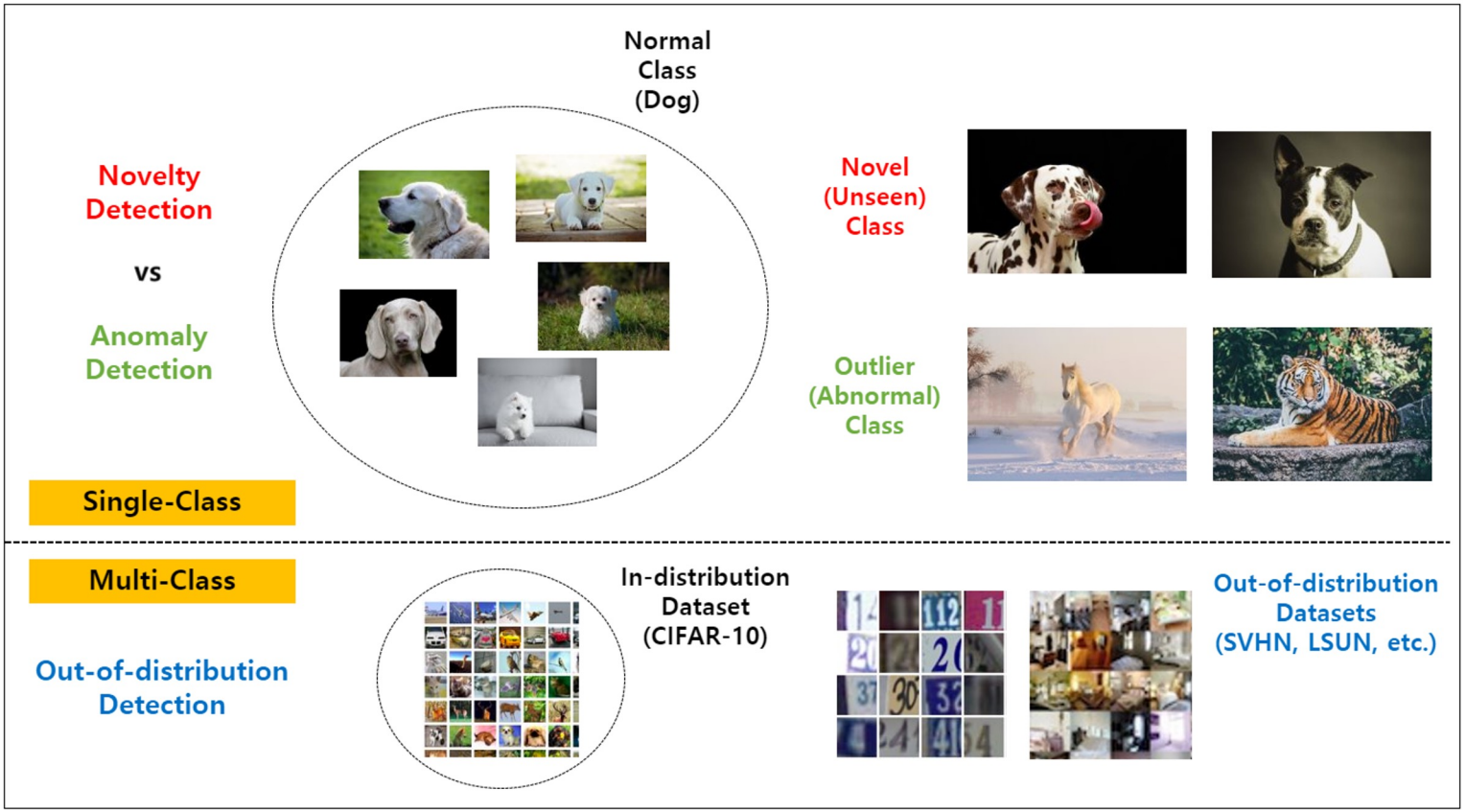
- Noise and anomalies are not same.
- Boundary between normal and anomalies is often undefined.
- More deeper challenges includes:
  - Explaining why it's anomaly.
  - What decision we can take based on the detected anomaly.

How are they produced?

- Objects from different class/underlying mechanism
  - Disease vs non-disease
  - Fraud vs not fraud
- Natural Phenomenon
  - Tail of a gaussian distribution
- Measurement/observation error
  - Faulty sensors
  - Needs replacement

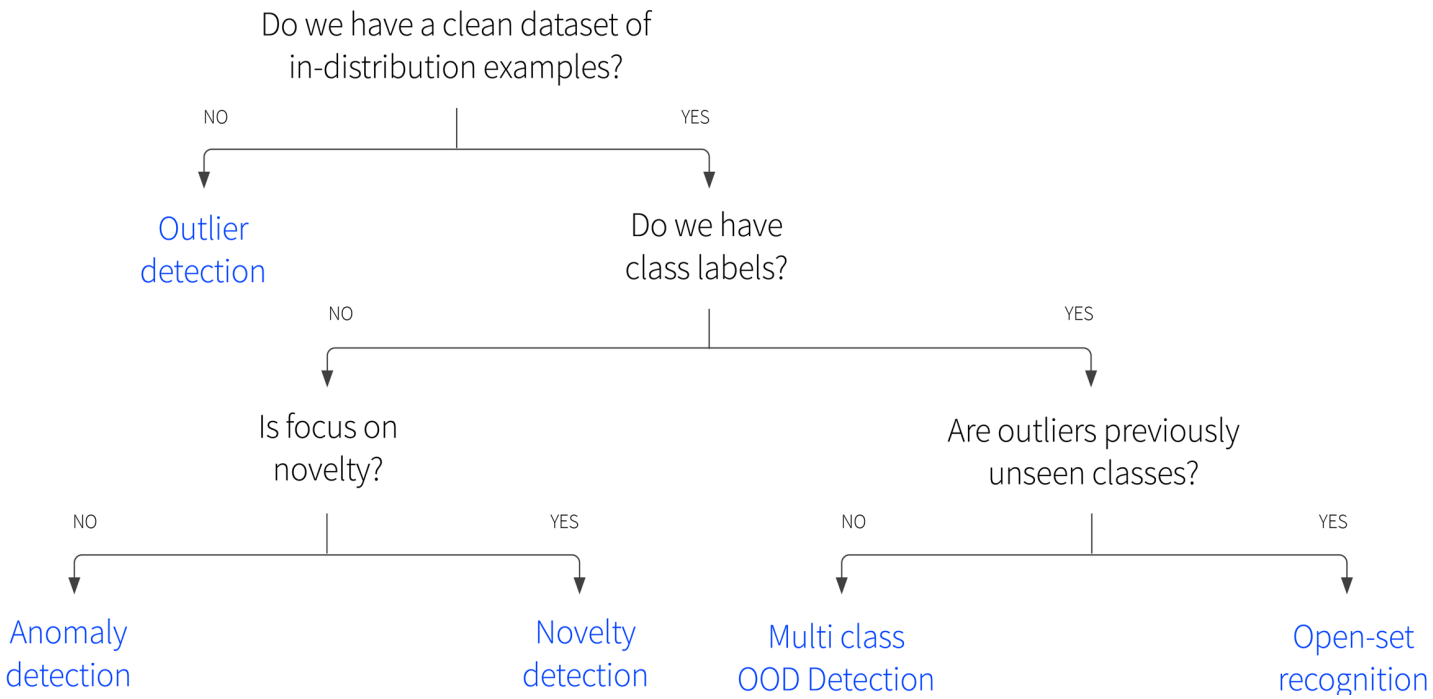
Terms related to anomaly detection

- Outlier analysis/Novelty Detection
- Change point (in time series)
- Some more..



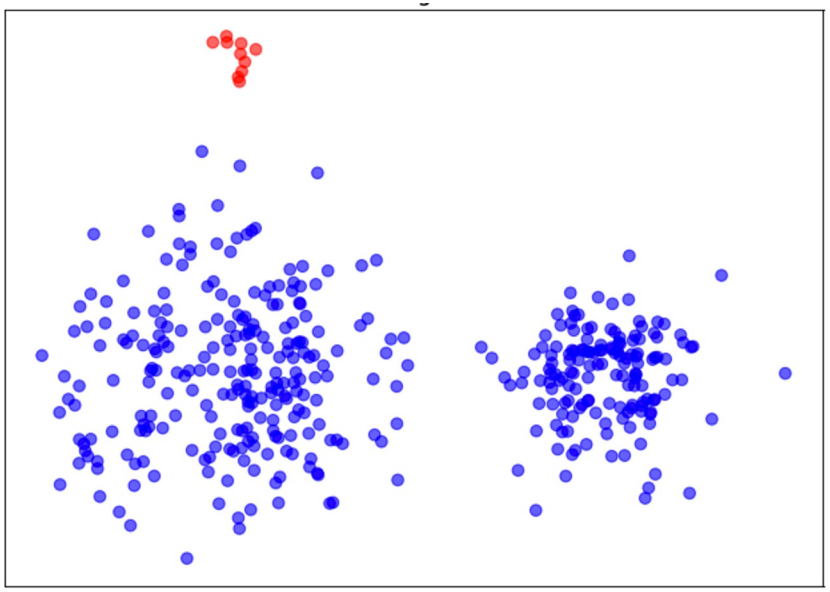
Github link

Relationships between OOD tasks



Anomaly Types

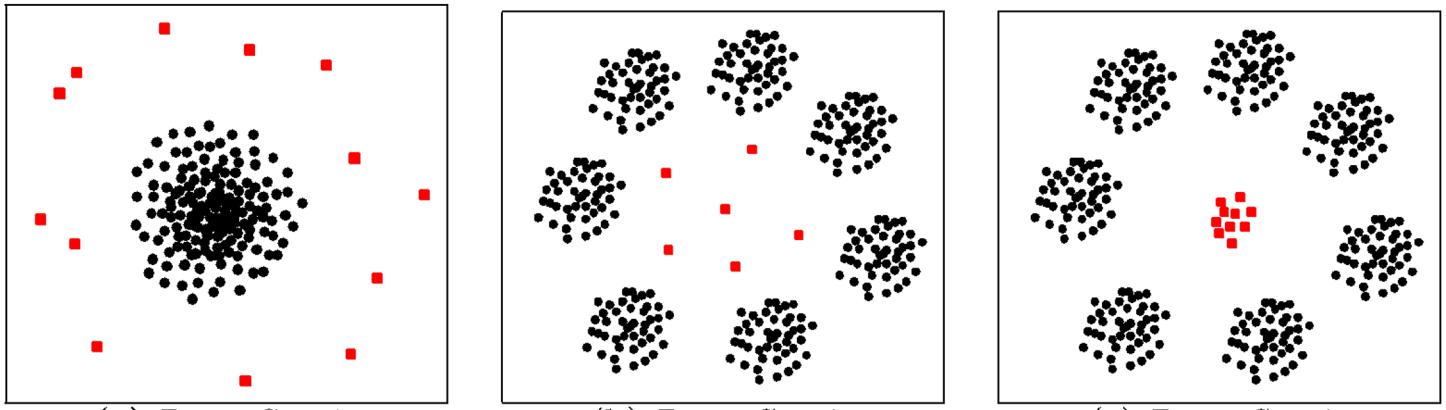
- Point anomalies
  - Any type of data



May-22	1:14 pm	FOOD	Monaco Café	$1,127.80
May-22	2:14 pm	WINE	Wine Bistro	$28.00
...				
Jun-14	2:14 pm	MISC	Mobil Mart	$75.00
Jun-14	2:05 pm	MISC	Mobil Mart	$75.00
Jun-15	2:06 pm	MISC	Mobil Mart	$75.00
Jun-15	11:49 pm	MISC	Mobil Mart	$75.00
May-28	6:14 pm	WINE	Acton shop	$31.00
May-29	8:39 pm	FOOD	Crossroads	$128.00
Jun-16	11:14 am	MISC	Mobil Mart	$75.00
Jun-16	11:49 am	MISC	Mobil Mart	$75.00

Point Anomaly

Collective Anomaly

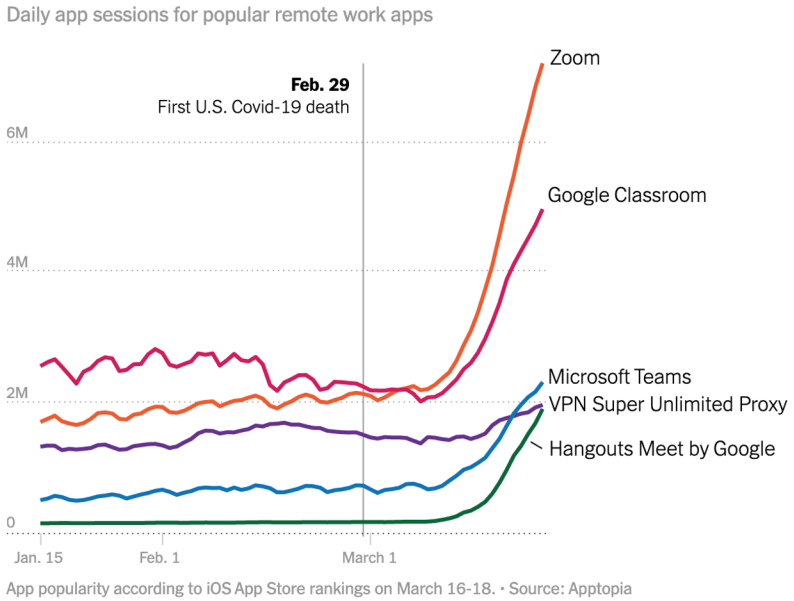


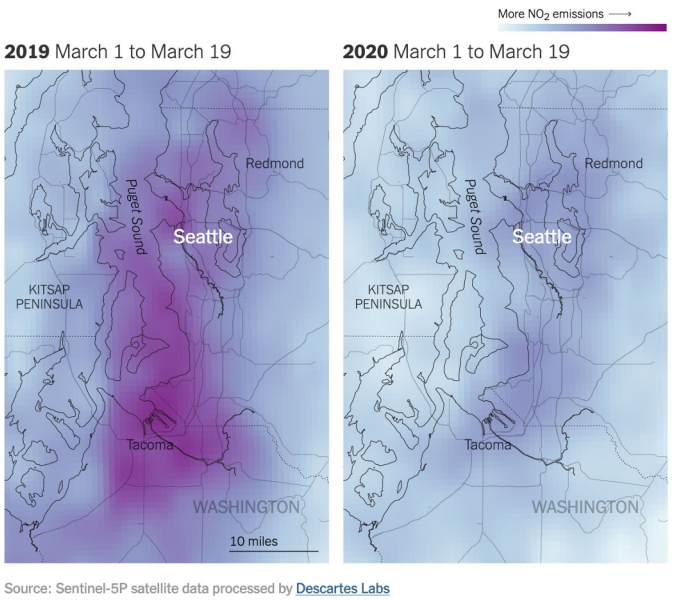
(a) Data Set 1 (b) Data Set 2 (c) Data Set 3

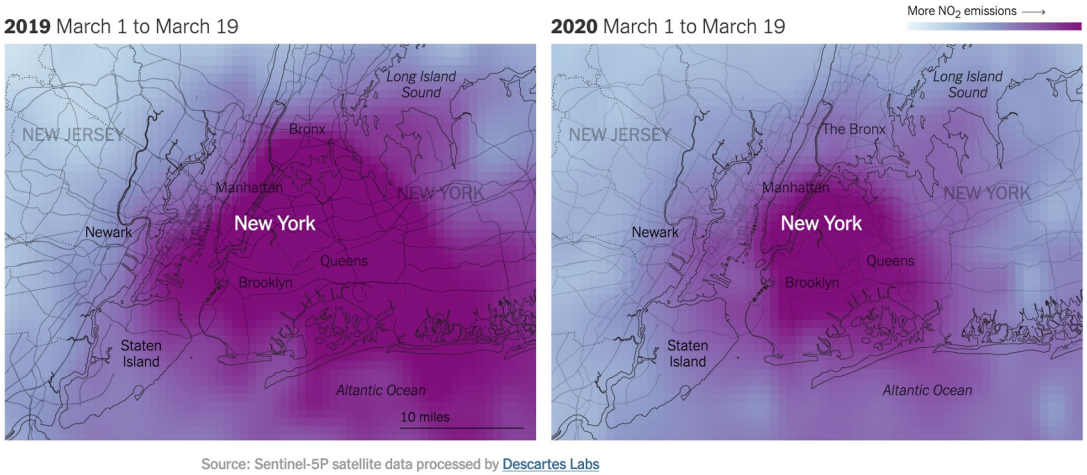
Chandola. V., et. al., Anomaly Detection: Survey, 2009

Anomaly Types

- Contextual anomalies
  - Structured data
  - Context as
    - Time
    - Space
    - Person



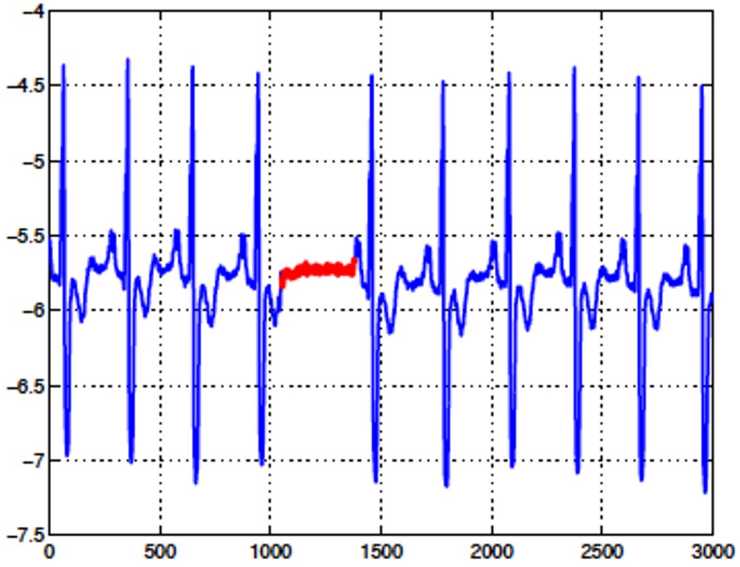




https://www.nytimes.com/interactive/2020/03/22/climate/coronavirus-usa-traffic.html,
https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html

Anomaly Types

- Collective anomalies
  - Structured data
  - problems that may not be identified based on a single data source or in a single location



Chandola. V., et. al., Anomaly Detection: Survey, 2009; *Zheng, Yu et al.*
"Detecting collective anomalies from multiple spatio-temporal datasets across different domains." GIS '15 (2015).

Anomaly detection settings

- Definition: “anomaly” is a data point generated by a process that is different than the process generating the “nominal” data
- Given:
  - Training data $\{x_1, x_2, x_3, \ldots, x_N\}$
- Goal:
  - Identify ‘anomalous’ data points.

labels (nominal, anomaly) OR anomaly score

Approaches for anomaly detection

- Based on labels availability
- Supervised
  - Labeled nominals and anomalies
  - Like class imbalance problem (i.e. spam <10%)
- Semi-supervised
  - Nominals are labeled
- Unsupervised
  - No labels provided

Can’t I formulate it as supervised learning?

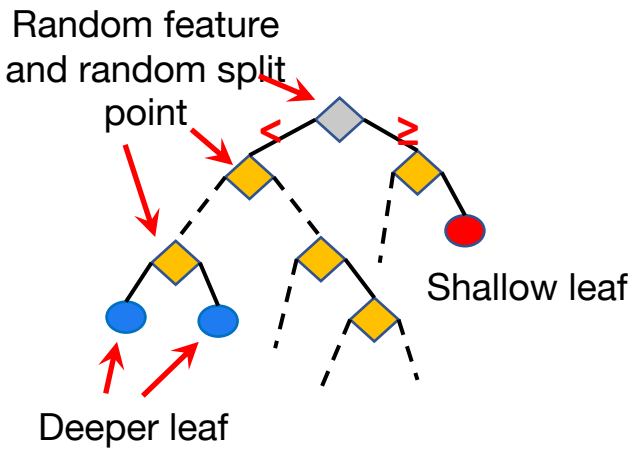
- Anomalies are rare.
- Labeled anomalies will not always be helpful.
- Anomalies present in test data is not guaranteed to follow the same pattern.

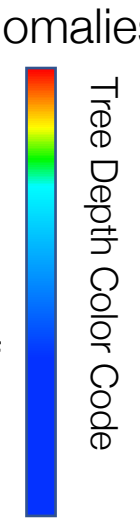
Anomaly detection algorithms

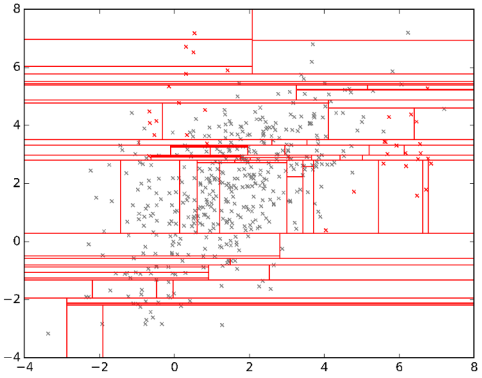
- Statistics based (Histograms, robust statistics)
- Neighbors / distance based (e.g., LOF and variants, ABOD)
- Density based (KDE, GMM)
- Reconstruction errors (e.g., using PCA)
- Projection based (Isolation Forest, HS Forest, RF Trees, LODA, EGMM)
- ML / Classification based (One-class SVM, SVDD, SVM, RF, pseudo anomalies)

Isolation Tree

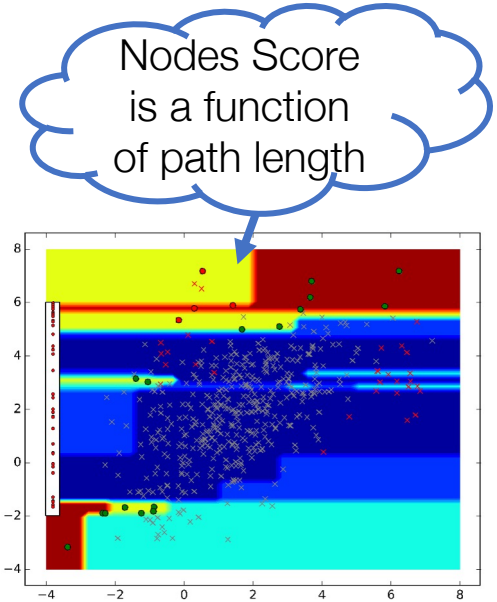
- Key Idea
  - Anomalies can be easily **Isolated** from nominals
  - Degree of anomaly is inversely proportional to depth







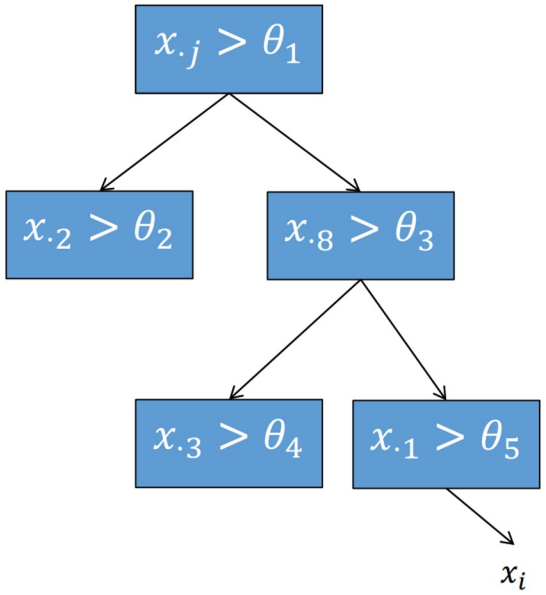
Partitioned subspaces



Depth Colored subspaces

Isolation Tree steps

- Construct a fully random binary tree
  - choose attribute j at random
  - choose splitting threshold θ_1 uniformly from [min x_j , max x_j]
  - until every data point is in its own leaf
  - let $d(x_i)$ be the depth of point x_i



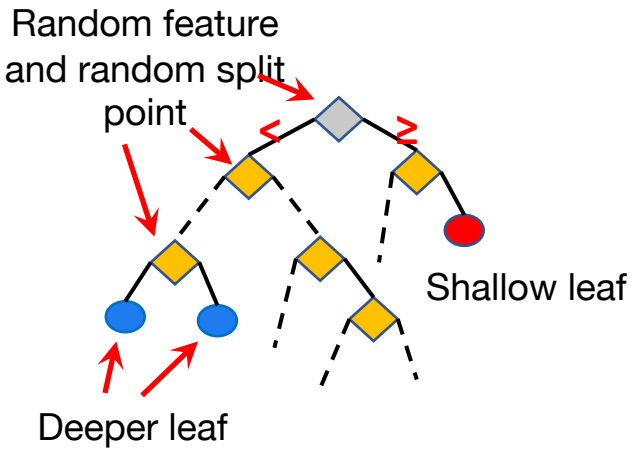
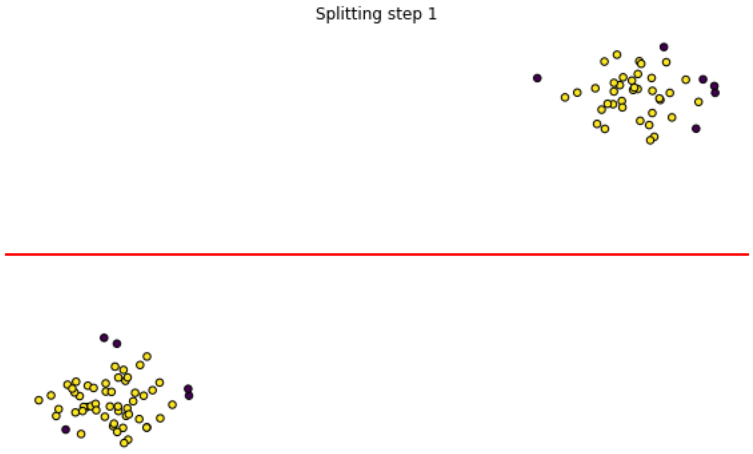
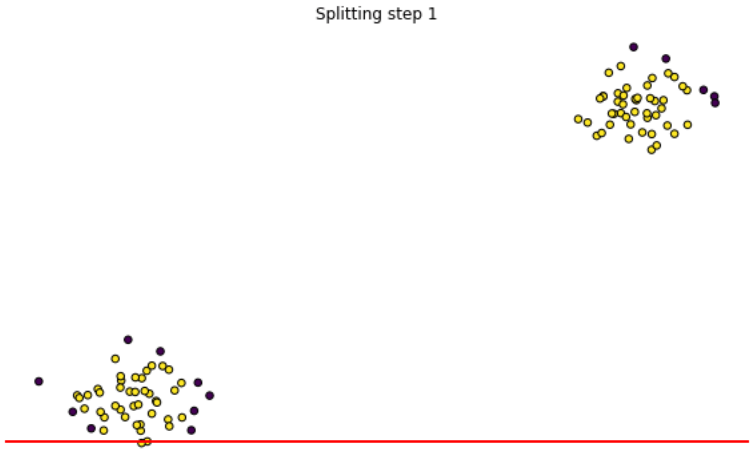


Figure and slide credit Thomas G. Dietterich

Demo

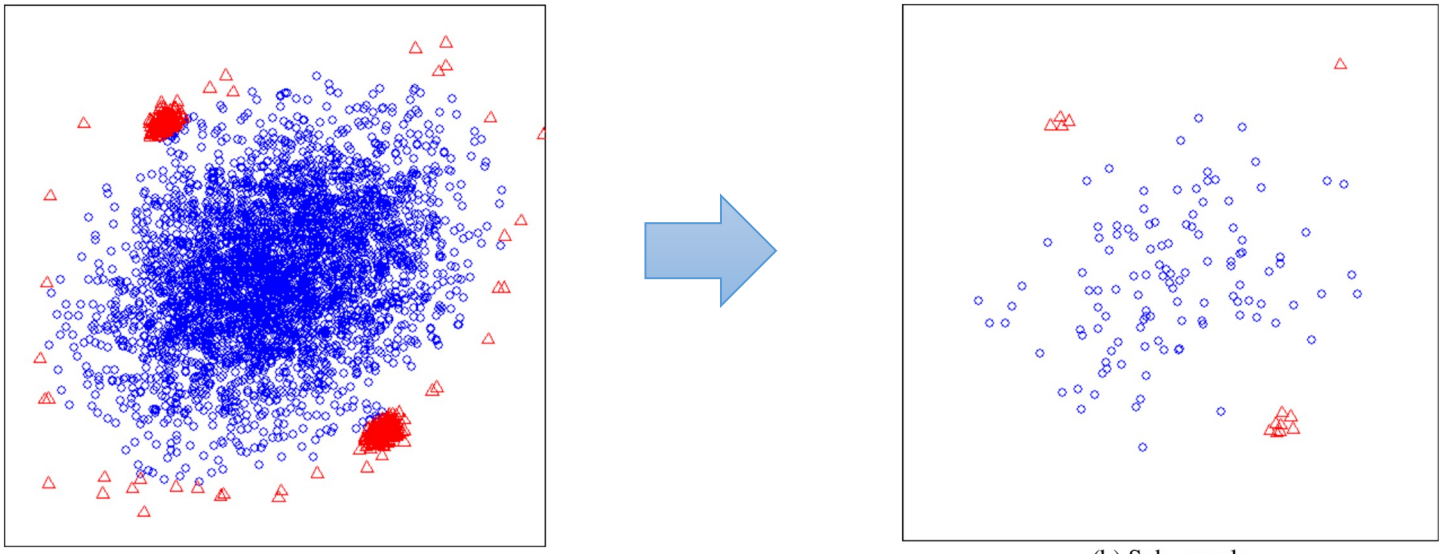
How splitting works for Isolation Forest?





What are the key ingredients?

- Subsampling



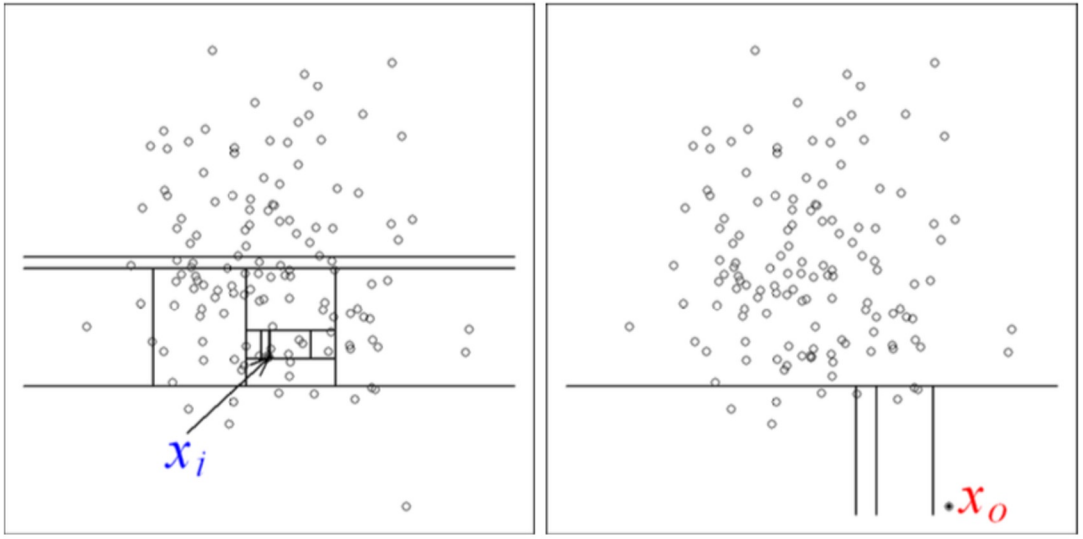
(a) Original sample
(4096 instances)

(b) Sub-sample
(128 instances)

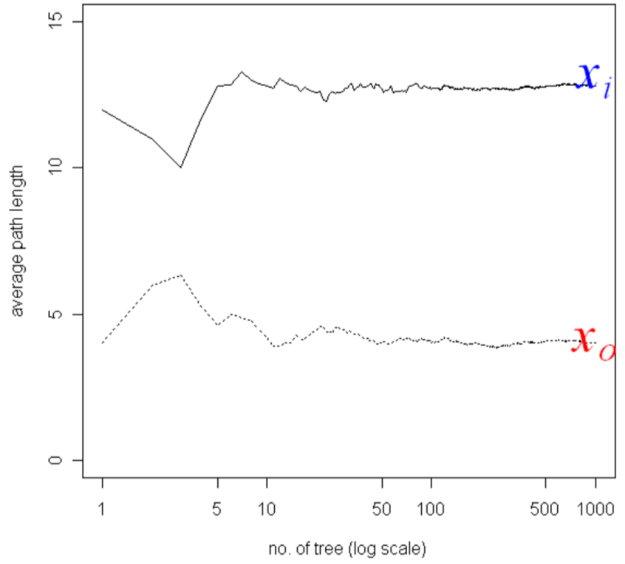
- Why is that necessary?
  - It seems $2^8 = 256$ points contain enough information to perform anomaly detection
  - Saves computational cost and time.
  - Adding more points didn't improve the results.

What are the key ingredients?

- Path length



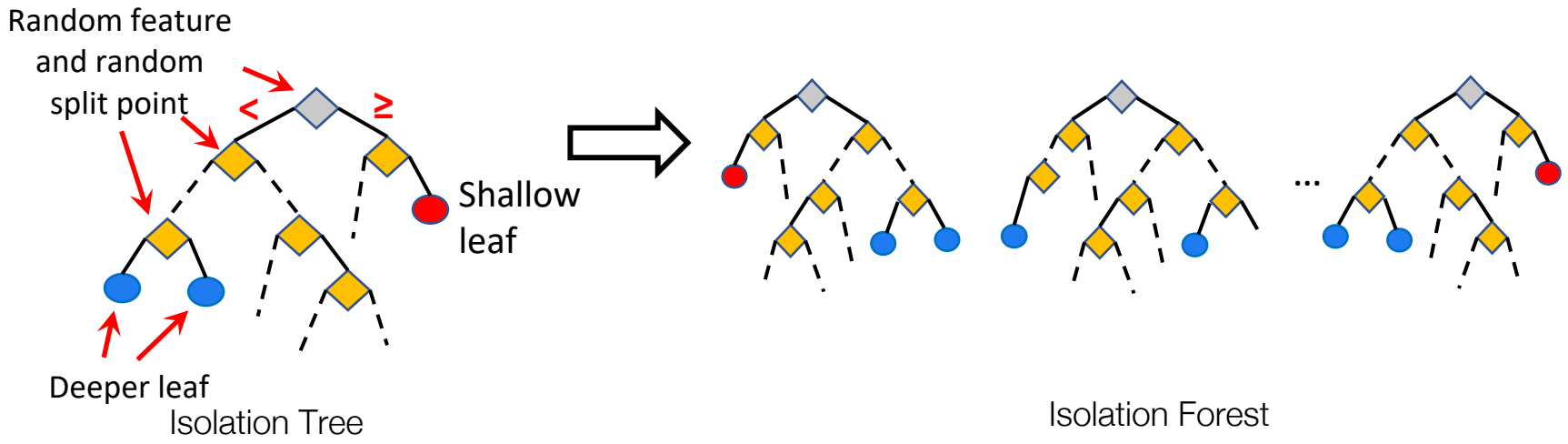
(a) Isolating x_i (b) Isolating x_o



(c) Average path lengths converge

- How it helps?
  - The smaller the path length the more anomalous it is.

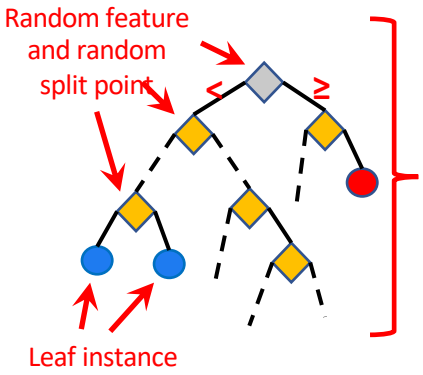
Isolation Forest



- State of the art unsupervised approach
- Assumptions
  - # of anomalies are small
  - Features are distinguishable
- Assigns uniform weights for subspace score

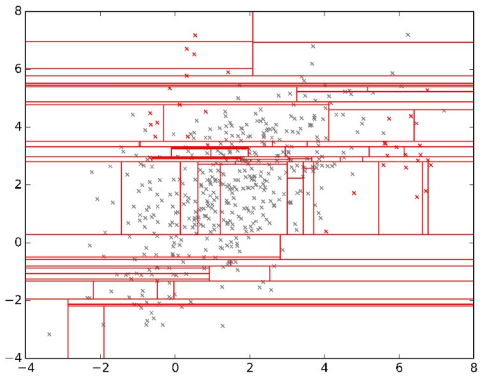
Techniques: Ensemble based (contd..)

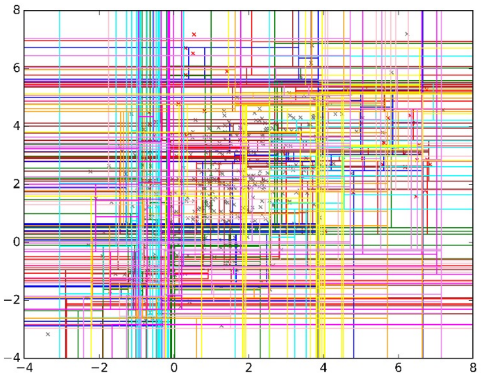
- Isolation Forest

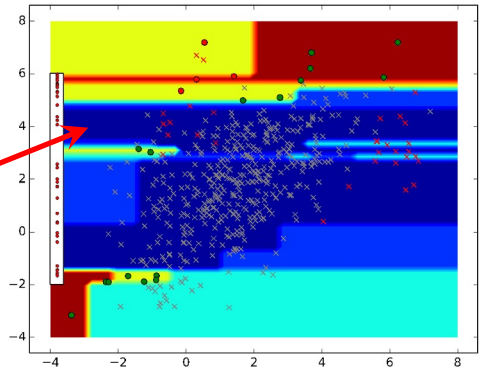


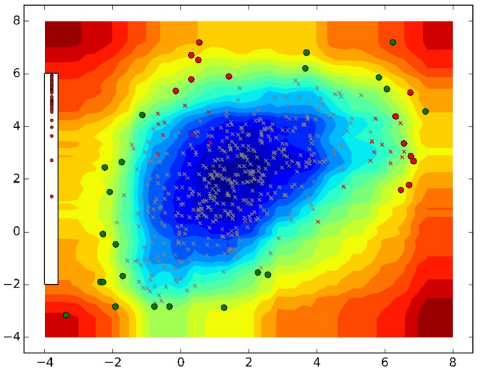
A node's score is a function of path length

Typically 100 trees in practice



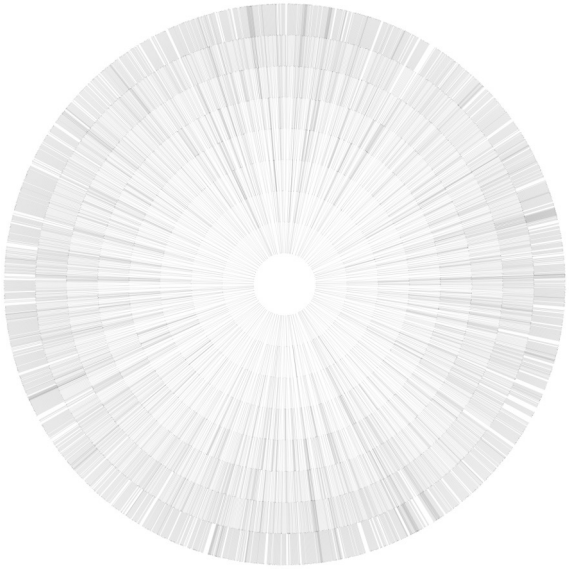




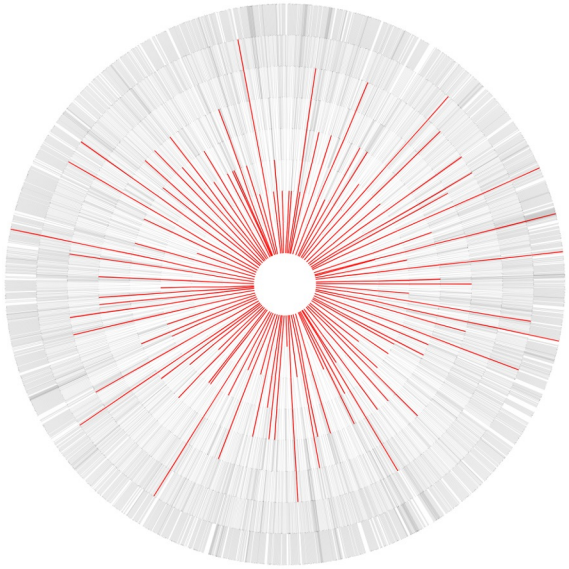


F T Liu, K M Ting, and Z-H Zhou, Isolation Forest, ICDM 2008

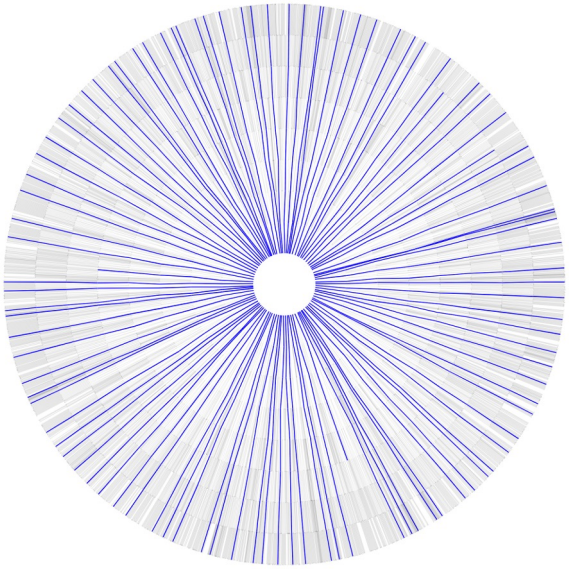
Techniques: Ensemble based (contd..)



Path length for
all data points



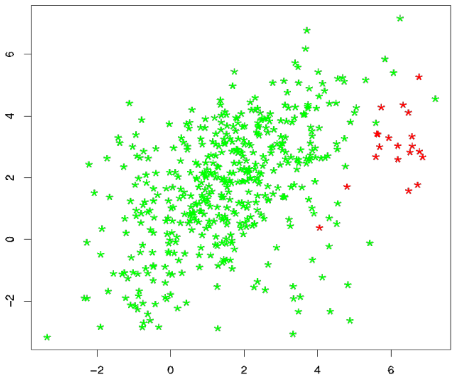
Path length for Anomalous
data point



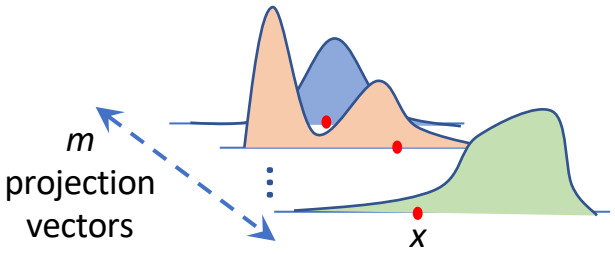
Path length for Nominal
data point

Credit: [Matias Carrasco Kind]

Loda: Lightweight on-line detector of anomalies



Synthetic data



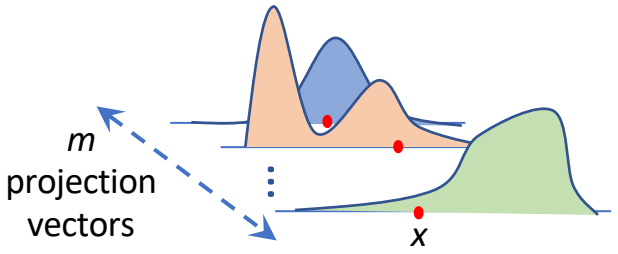
- $\Pi_1, \ldots, \Pi_M$ set of M sparse random projections
- Let $w_m = 0, \ldots, 0$
- Choose d elements of w_m and set them to normal random variate
- $\Pi_m(x) = w_m \cdot x$
- $f_1, \ldots, f_M$ corresponding 1-dimensional density estimators
  - Pevny uses optimal histograms
- $S(x) = -\frac{1}{M}\sum m \log f_m(x)$ average "surprise"

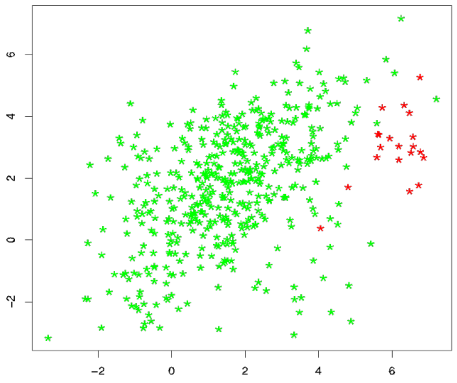
Techniques: Ensemble based (contd..)

- Loda: Lightweight online detector of anomalies *[Pevny 2015]*

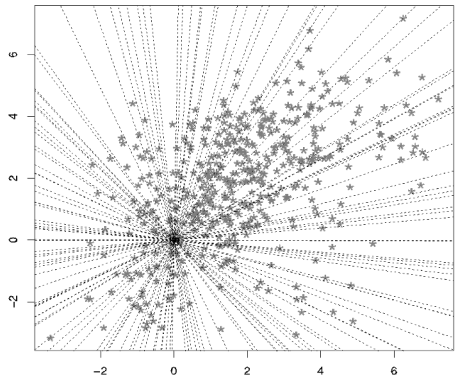
$$Score(x) = \frac{1}{m}[f_1(x) + f_2(x) + \cdots + f_m(x)]$$

Negative log pdf value for a 1D histogram formed by a sparse random projection of x

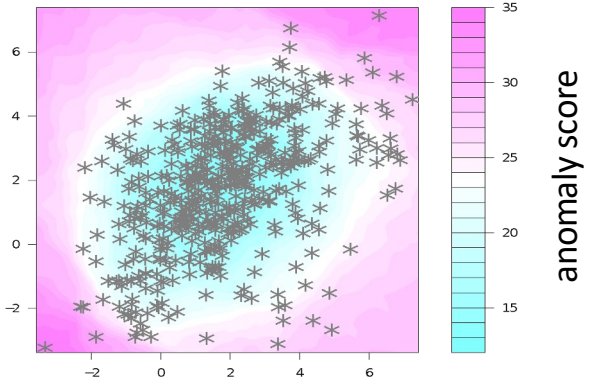




Synthetic data



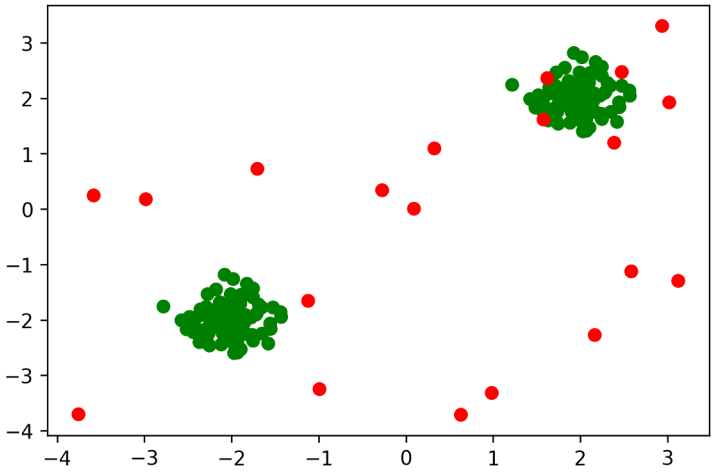
Loda projections



Loda anomaly scores

Pevny, T., Loda: Lightweight on-line detector of anomalies, Machine Learning, 2016

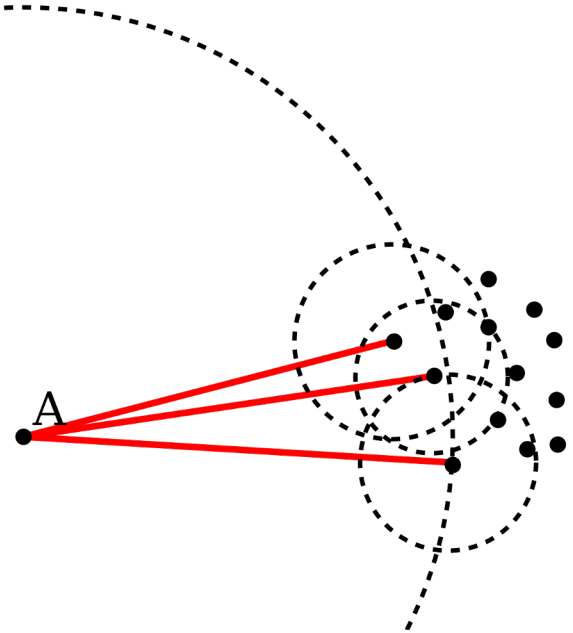
Local Outlier Factor

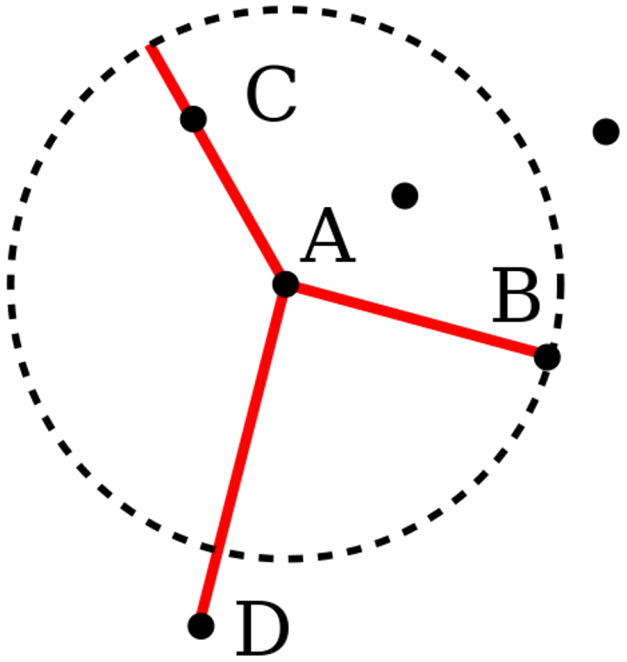


- Distance based approach
- Takes nearest neighbors' density (k neighbors')
- Different metrics for distance
- Key assumption:
  - density around a non outlier object ≅ density around its neighbors,
  - density around an outlier object ≠ density around its neighbors

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. (2000). *LOF: Identifying Density-based Local Outliers* (PDF). *Proceedings of the 2000 ACM SIGMOD*

Local Outlier Factor

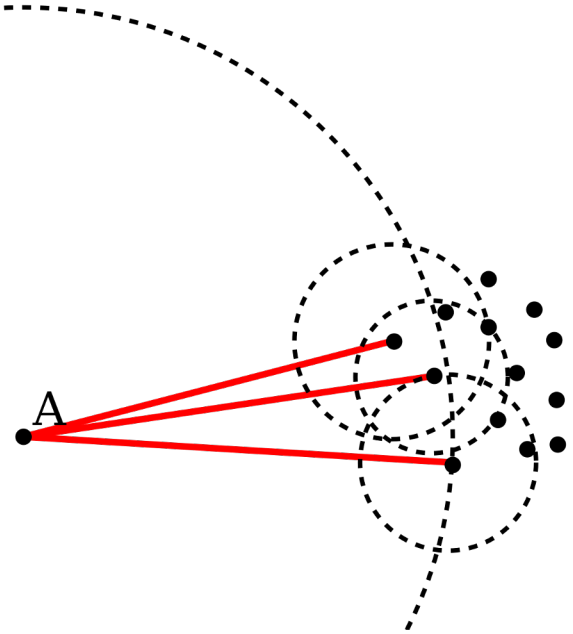


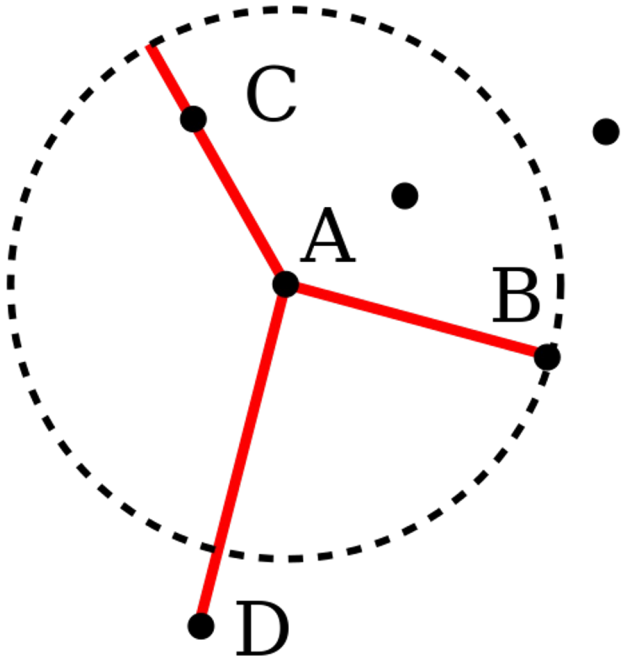


- $Local\ (reachability) density(x) = \frac{No\ of\ neighbors}{\Sigma reachable\ distance\ for\ all\ neighbor}$
- $LOF(x) = \frac{\text{average local density of k−nearest neighbors'}}{\text{local density of x}}$

Figures: https://en.wikipedia.org/wiki/Local_outlier_factor

Local Outlier Factor





- LOF Score
  - ~ 1 means Similar density as neighbors
  - < 1 means Higher density than neighbors (Inlier)
  - > 1 means Lower density than neighbors (Outlier)

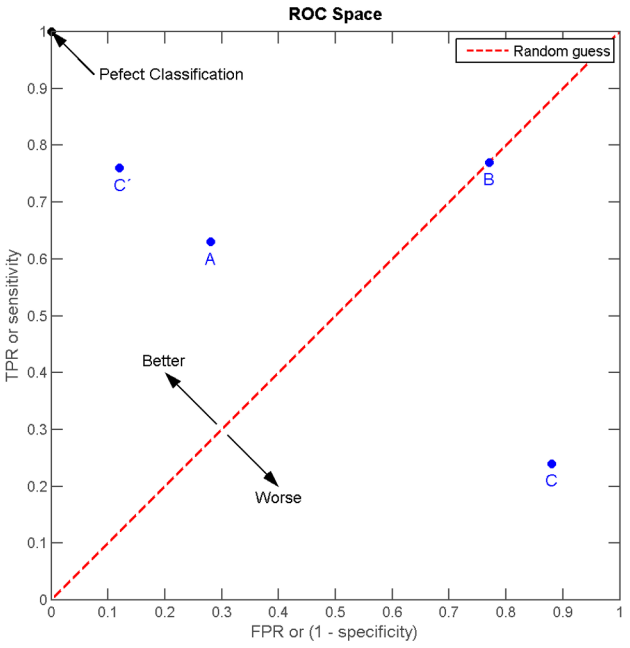
How to evaluate an anomaly detector?

- Metrics used mostly:
  - Precision@k, Recall@k
  - AUC (Area Under Curve)

	Actual Positives	Actual Negatives
Positive Predictions	True Positives (TP)	False Positives (FP)
Negative Predictions	False Negatives (FN)	True Negatives (TN)

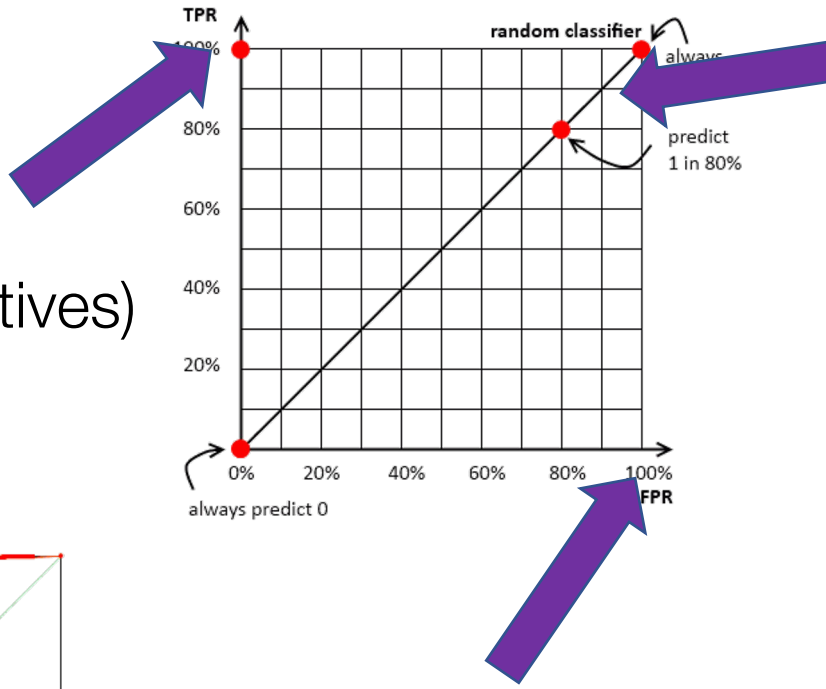
True positive rate
TPR= TP/(FN+TP)
False positive rate
FPR = FP/ (TN+FP)

- What is an AUC?
  - Measure for evaluating the performance of a classifier
  - probability that a randomly-chosen anomaly point is ranked above a randomly-chosen nominal point
  - The expectation that a uniformly drawn random positive is ranked before a uniformly drawn random negative.

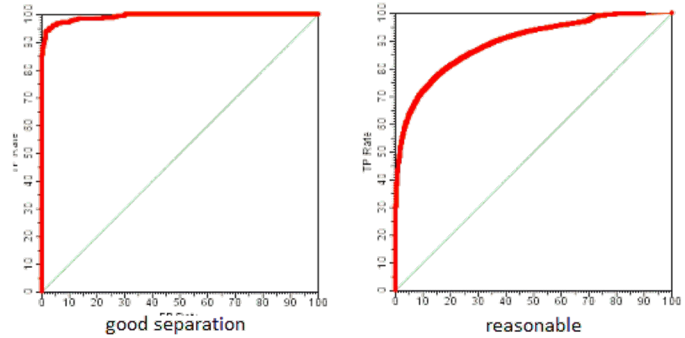


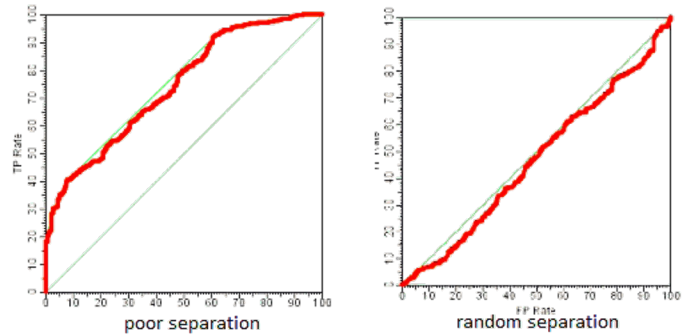
How to evaluate an anomaly detector?

- AUC Score
  - 1 (perfect classifier)
  - 0.5 (random classifier)
  - 0 (negatives comes before positives)

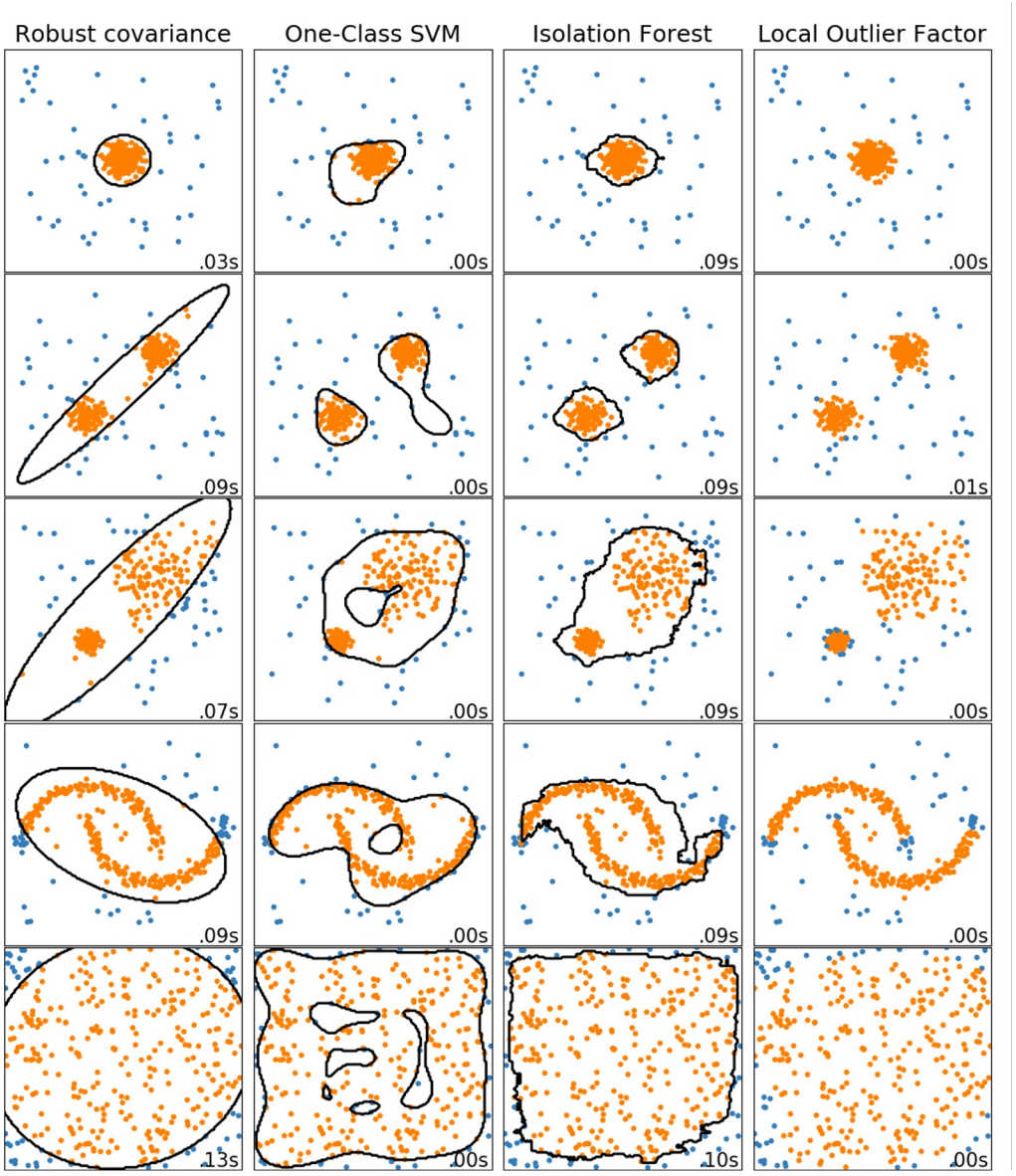


- Some more examples

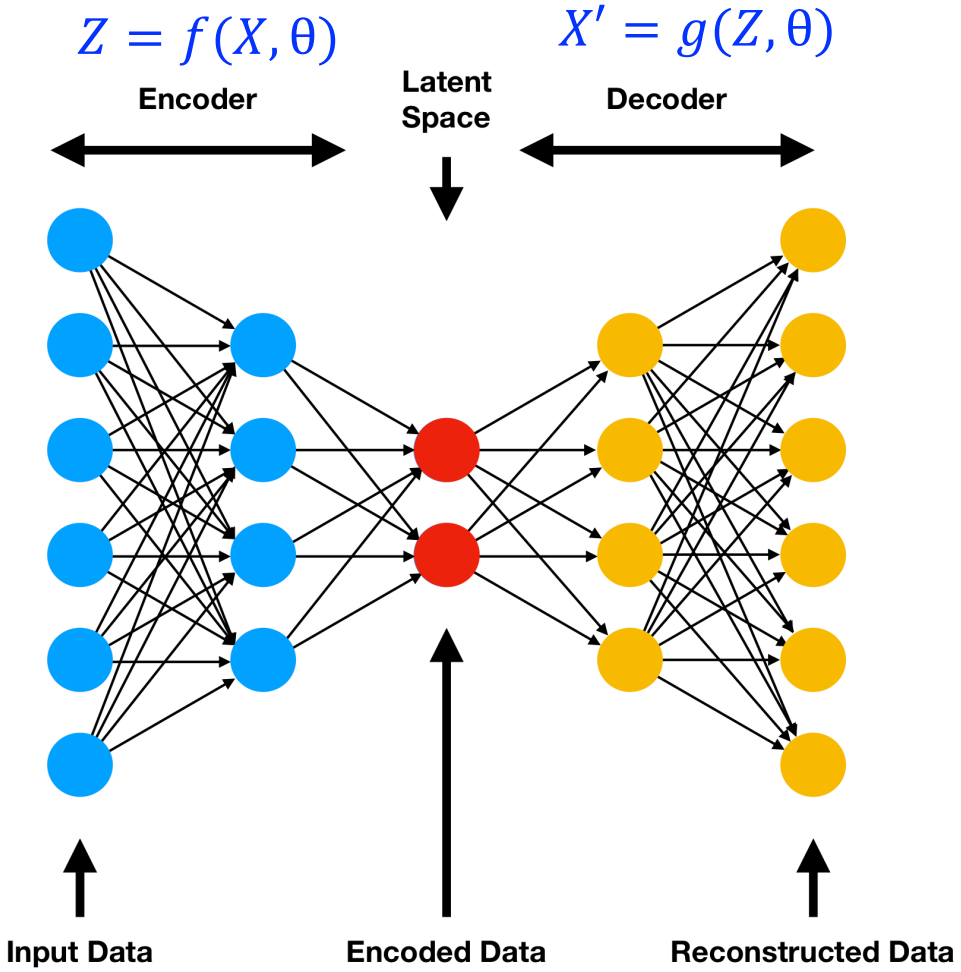




Some more examples at `scikit-learn`



Auto encoder

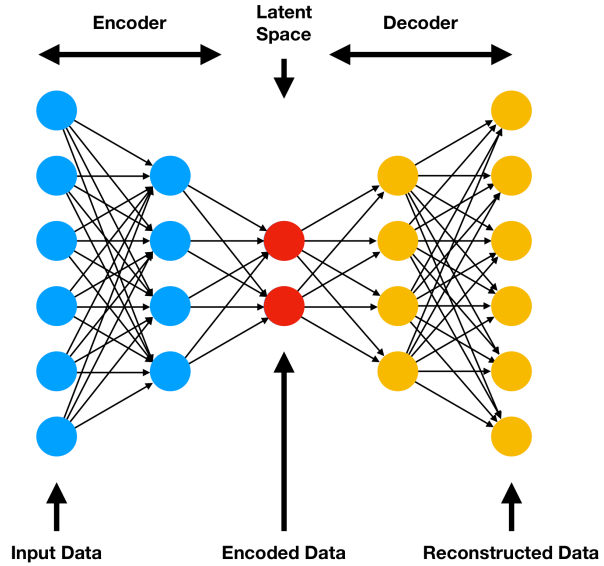


- Maps input data to a **latent space**
- Output data was **reconstructed** from the latent space.
- Dim(latent space) << Dim (input space)

How to detect anomaly using AE?

- Training
  - Using normal data/samples from normal class
- Testing
  - Any data that **can't be reconstructed back** are considered as anomaly.

$Z = f(X, \theta)$ $X' = g(Z, \theta)$



- $|X' - X| > \delta$ then anomaly!

Demo