Data Mining Anomaly Detection: Finding “weirdness”

João Brandão

DS PORTUGAL

BOSCH

Anomaly

"Something that deviates from what is standard, normal, or expected"
Oxford Dictionary

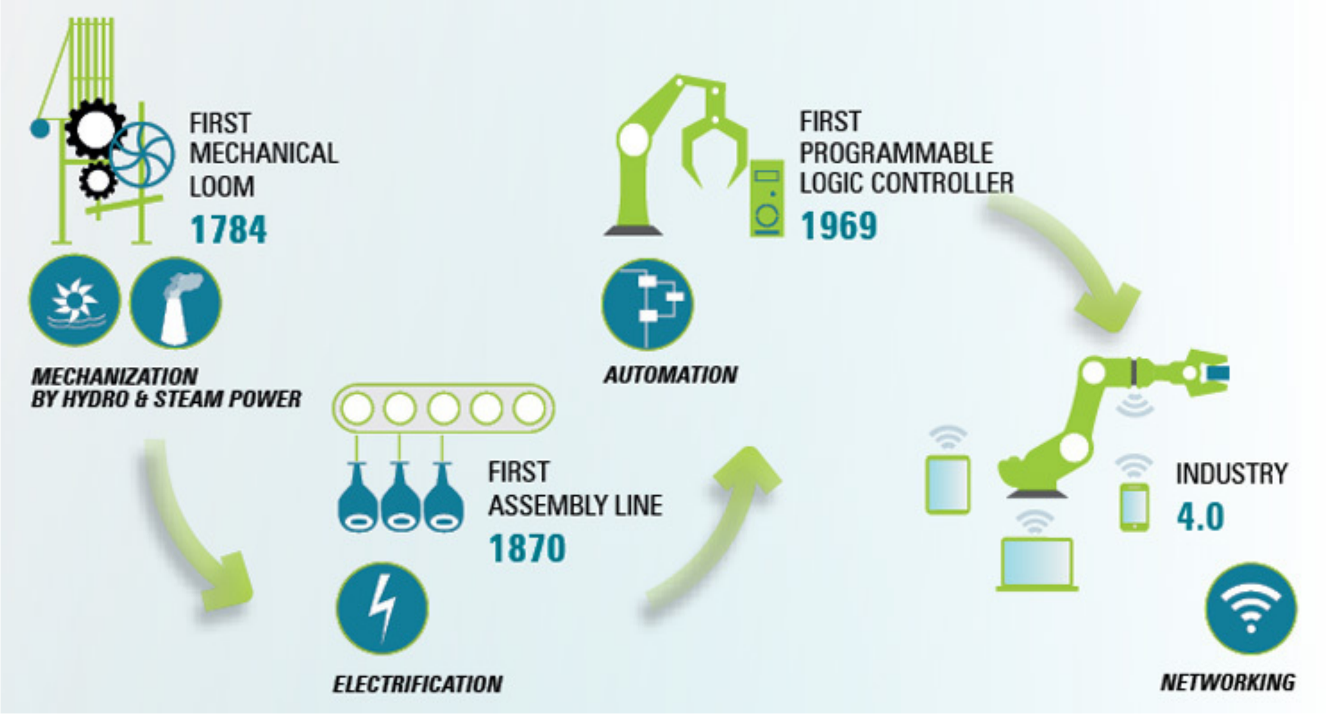
Anomaly

- Outliers
- Abnormalities
- Exceptions
- Discordant observations
- Surprises

Importance

Importance

Importance



DS PORTUGAL

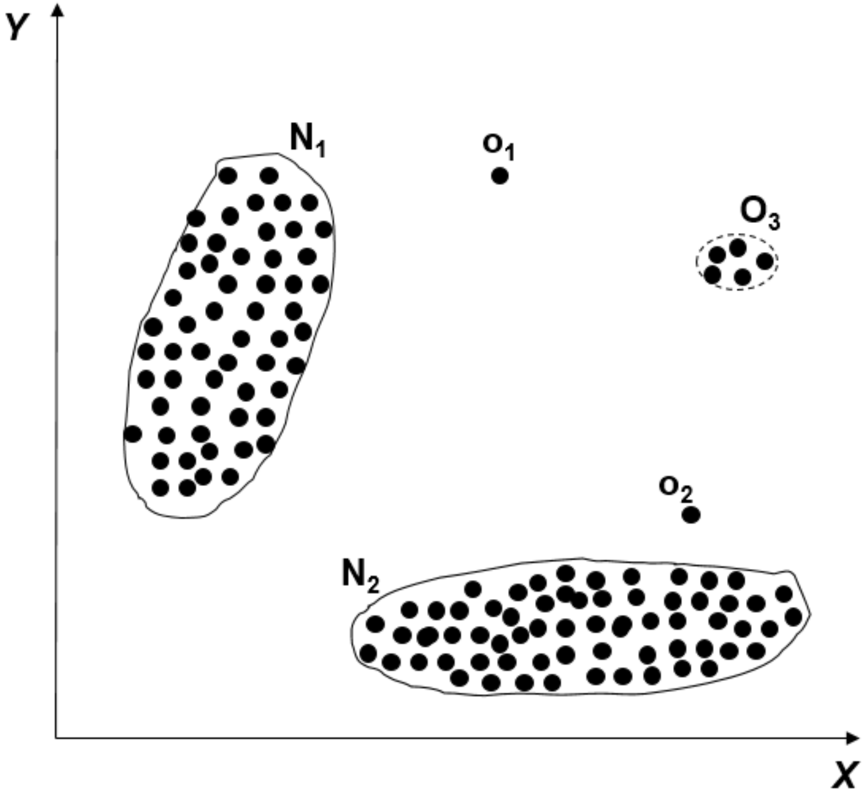
Applicable Domains

- Manufacturing Process
- Machine Monitoring
- Fraud
- Security
- Healthcare

Anomalies Types

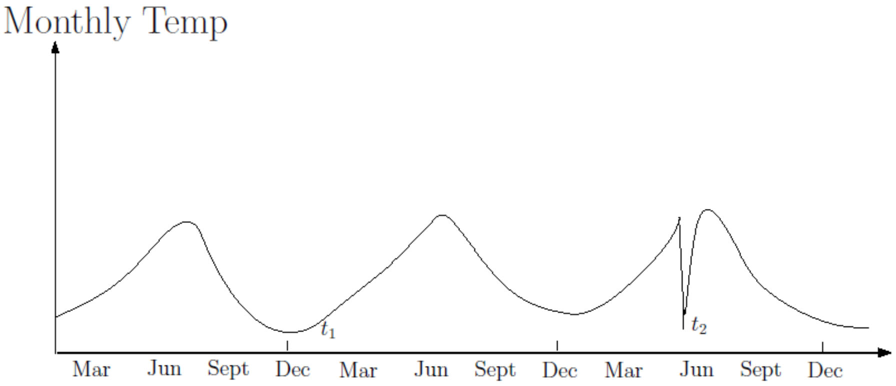
- Point Anomalies

  Data point(s) considered anomalous with respect to the rest of the data.



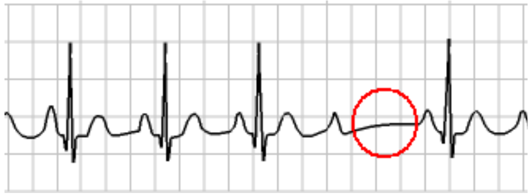
Anomalies Types

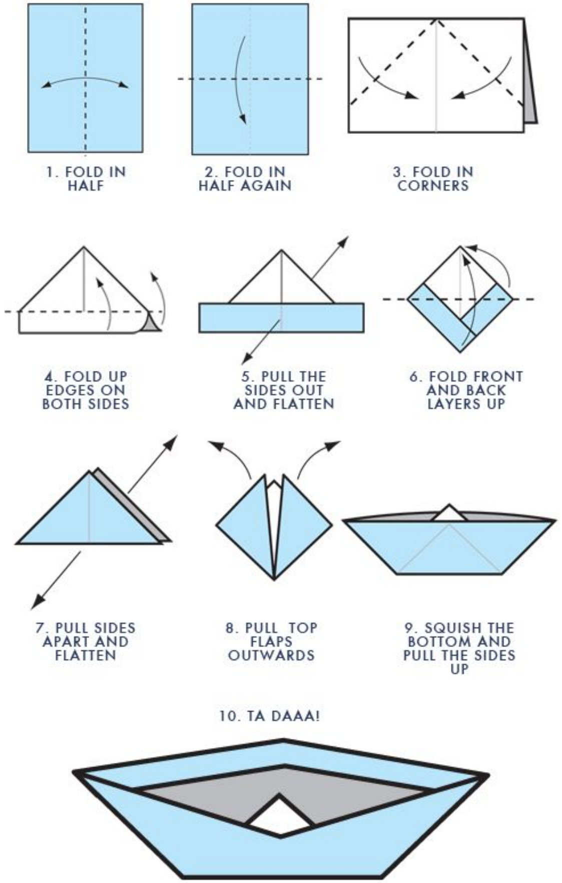
- Contextual Anomalies
  - Context variables
    - Longitude ,latitude, sequence position
  - Behavior variables
    - Amount produced, temperature, duration, ..



Anomalies Types

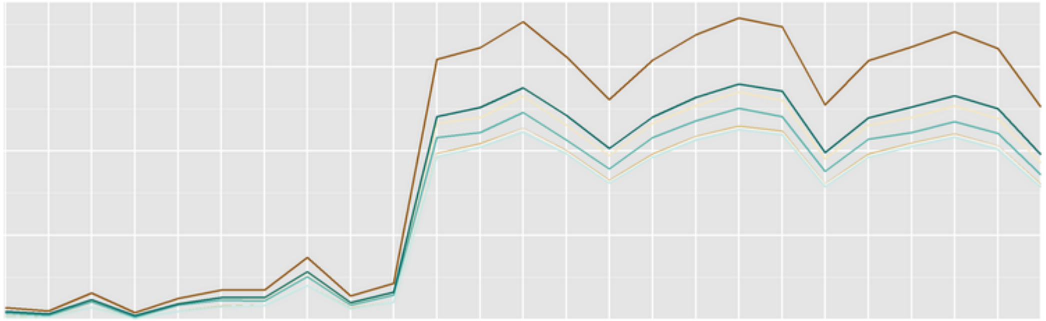
- Collective Anomalies
  - Relationship among data instances
  - Sequence, spatial, combinations, …




1. FOLD IN HALF
2. FOLD IN HALF AGAIN
3. FOLD IN CORNERS
4. FOLD UP EDGES ON BOTH SIDES
5. PULL THE SIDES OUT AND FLATTEN
6. FOLD FRONT AND BACK LAYERS UP
7. PULL SIDES APART AND FLATTEN
8. PULL TOP FLAPS OUTWARDS
9. SQUISH THE BOTTOM AND PULL THE SIDES UP
10. TA DAAA!

Challenges

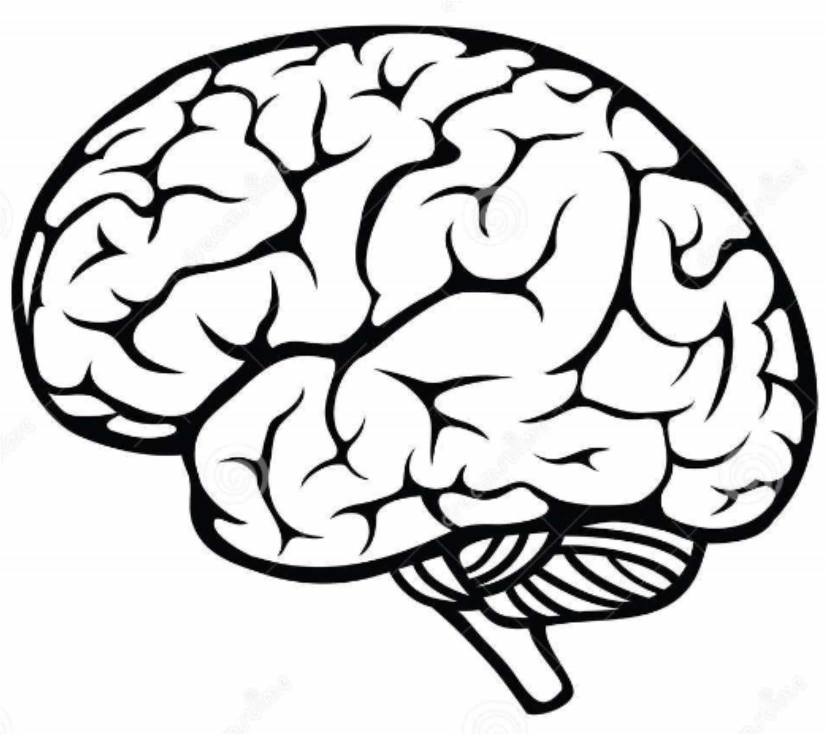
- Defining a normal region
- The evolution of normal
- Anomaly adaptation
- Application domain specificity
- No labeled data
- Rare **is not** anomalous



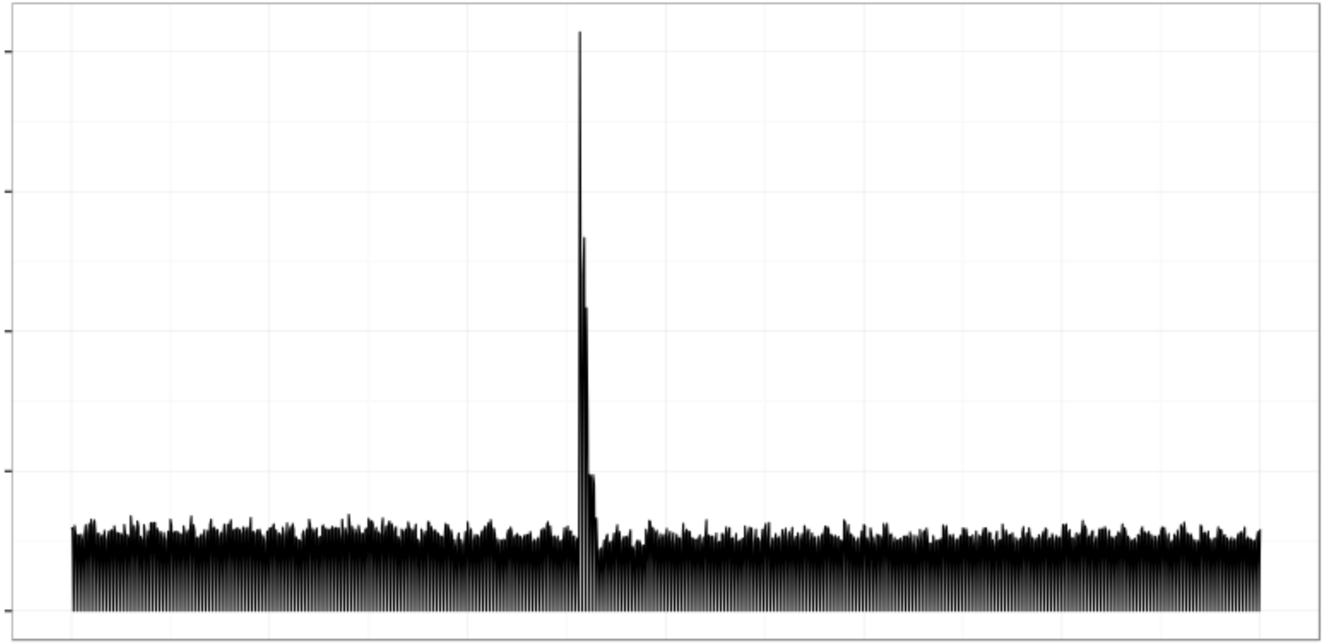
Data Nature

- Nature of input data
  - Univariate vs Multivariate
  - Categorical, nominal, continuous,..
- Related Instances
  - Temporal, spatial, spatiotemporal

Techniques



Techniques – Visual Detection



Techniques – Visual Detection

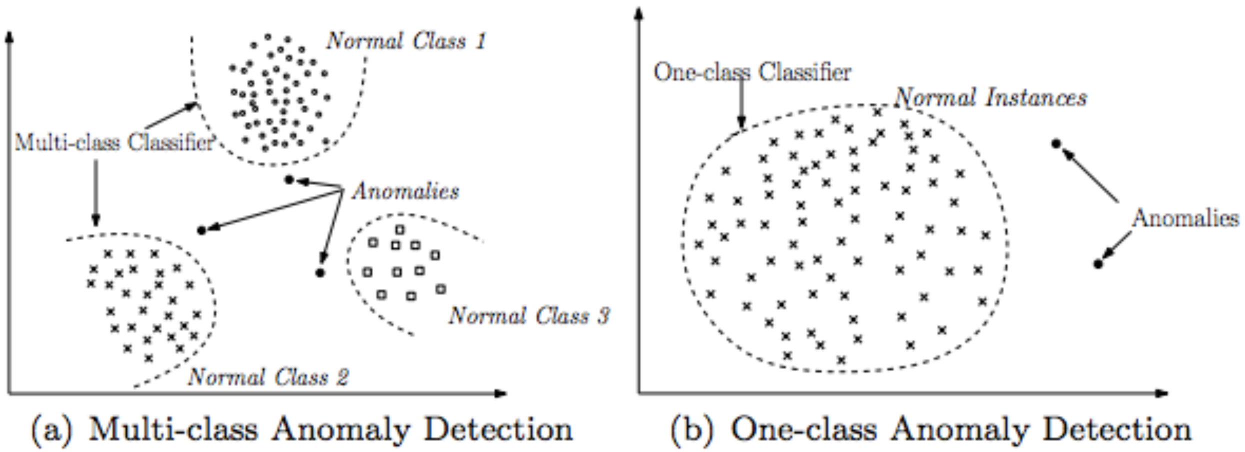
Techniques – Visual Detection

Techniques

- Machine Learning
- Statistics
- Information theory
- Spectral theory

Classification

- One-Class Anomaly
- Multi-Class Anomaly

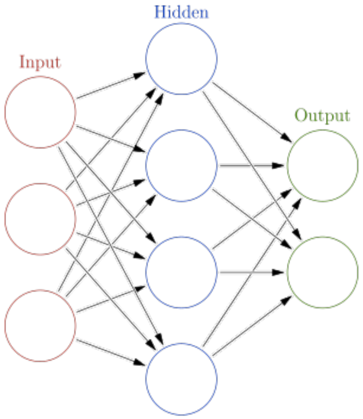


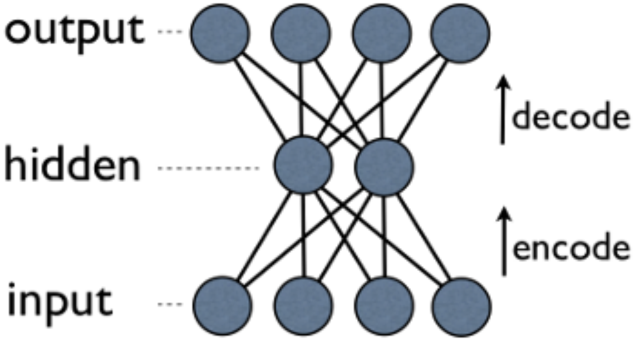
(a) Multi-class Anomaly Detection (b) One-class Anomaly Detection

Classification - Neuronal Network

- Multi -Class
  - Train classes with normal data
  - Test of accept/reject

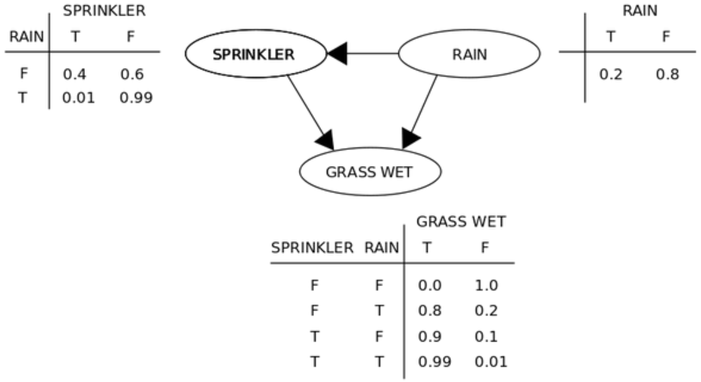
- One-Class
  - Replicator Neural Networks
  - Auto Encoders
  - Look at the error %





Classification - Naive Bayesian Network

- Multi-class
- Relationship between events
- Prior probabilities
- To occurrence of certain events influence the probability of other events occurring.
- Based on observed properties



	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99

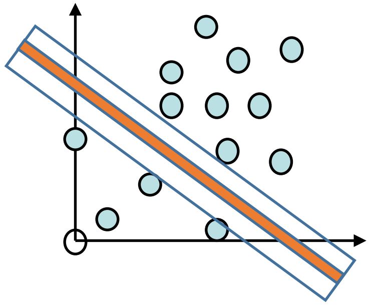
RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Classification - SVM

- Support Vector Machine
  - One-class
  - Maximize weigh if the margin

- Normal data records belong to high density data regions

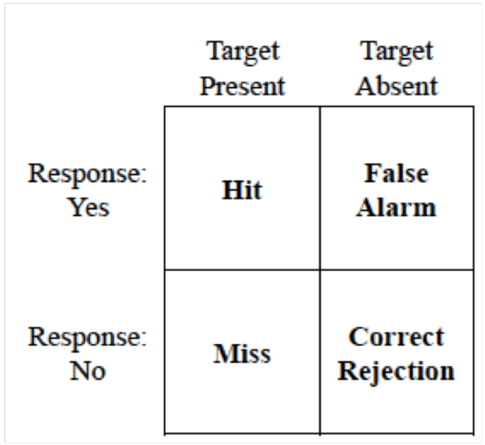


Evaluation

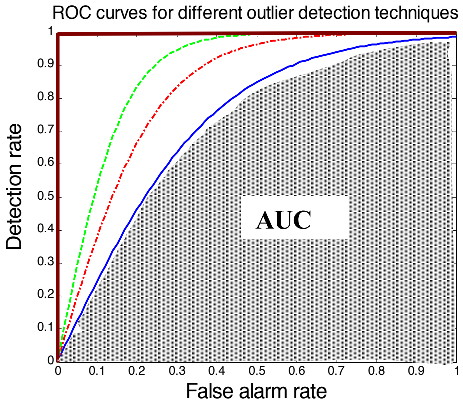
- Accuracy is not enough
    - 99% is normal data

- Signal Detection Theory

- Detection rate (recall)
    - Hits/(Hits + Missed)
- False Alarm rate
    - False Alarm/(False Alarm + Correct Rejections)



	Target Present	Target Absent
Response: Yes	**Hit**	**False Alarm**
Response: No	**Miss**	**Correct Rejection**



Nearest Neighbor

Distance

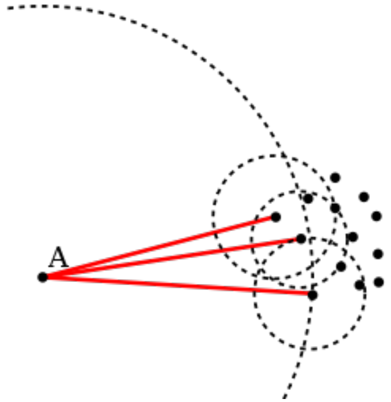
- Anomalies occur far from their closest neighbors
- Distance (or similarity)
  - Euclidian distance between data distances
  - Matching Coefficient for categorical attributes.
- Score
  - Total distance of data instance to its k-th nearest neighbor

Density

- Normal data instances occur in dense neighborhoods
- Global density
  - Count the number of nearest neighbors (n) that are not more than d distance
  - Nº Equal Attributes for same Category (Categorical Attributes)
- Score
  - Inverse Density

Nearest Neighbor

- Local Outlier Factor
  - Density based techniques perform poorly if the data has multiple regions
  - Compare the local density of a point with the densities of its neighbors
  - Local Density = k/volume of the hyper-sphere

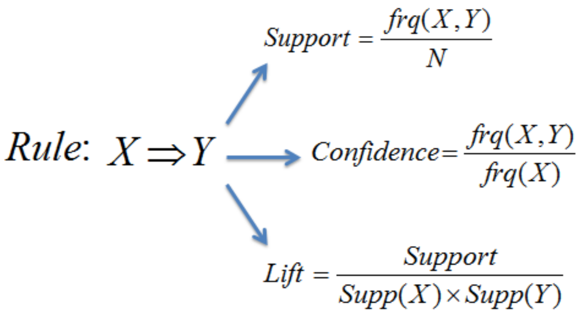


Association Rules

- **Low support**
  - Not usual happen
  - Is **not usual** to snow in Braga

- **1/Confidence**
  - When happens X
    usually Y doesn't happen
  - When it snows usually people
    do not wear a t-shirt outside

- **Create control limits**



$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Statistical Techniques

- Checking Normality
  - Histogram
  - Quantile plot
  - Z-Value

- Decide use Parametric or Non-Parametric
  - It is very important for small datasets

- Data Transformation
  - Log(x) or Log(x+c)

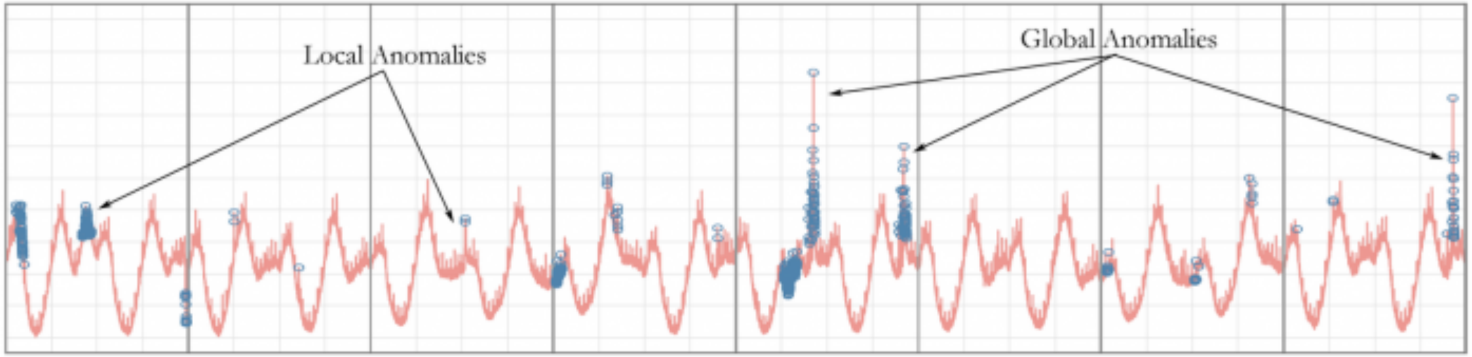
Statistical Techniques

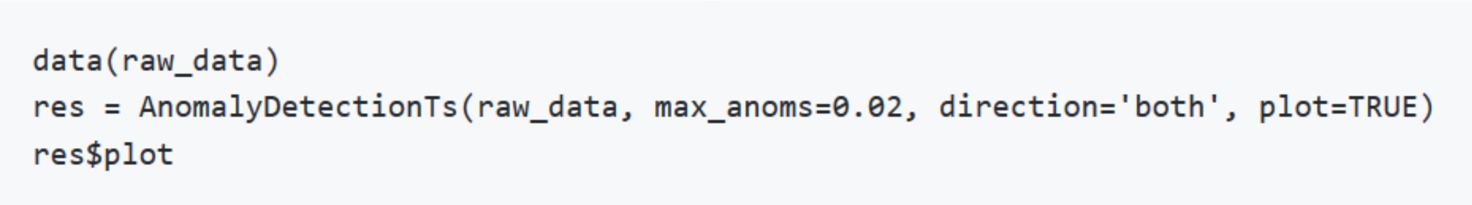
- Gaussian distribution
  - Uses medium and st. dev
  - Recalculate medium, st. dev with a sliding window

- Anomalies change a lot the St. Dev and the Mean
  - Use Median Absolute Deviation
  - Recalculate median and st. dev.
  - New (modified) z-score

Time Series - Seasonal Hybrid ESD

- Generalized ESD
- Twitter
- Global vs Local Anomalies





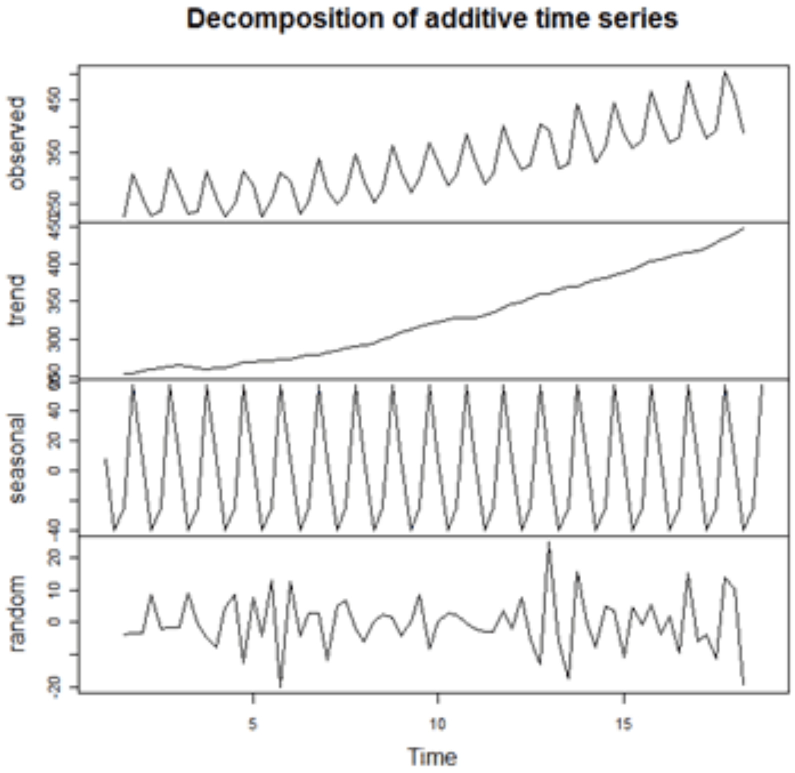
```
data(raw_data)
res = AnomalyDetectionTs(raw_data, max_anoms=0.02, direction='both', plot=TRUE)
res$plot
```

Time Series

- Estimate % anomalies that you are looking for

- G-Score (Absolute Dev. Z-Score)

- Seasonal Decomposition
  - Trend
  - Seasonal
  - Residual (or Random) Components



Others

- Information Theory
  - Use entropy concepts
  - Measure quantity of new information
  - Unsupervised

- Spectral
  - Dimensional reduction
  - Principal component analysis (PCA)
  - How changes vectors change with new data point

Output

- Score
  - Each instance is given an anomaly score
  - Threshold alert needed

- Label
  - Each instance is flagged as normal/anomaly
  - Usually used in classification

Quick overview

- Very Important subject for the present and future
- Many challenges, many options
- You don’t have a do all things algorithm
- Online anomaly detection challenge
- Fixed rules, still have a place
- Visual detection, still have a place

Thank

jpabrandao@gmail.com

linkedin.com/in/jpabrandao