# CS 522 – Selected Topics in CS

## Lecture 02 – Preprocessing

January 24, 2022

# Objectives

- Machine Learning: Review
- Missing Values Treatment
- Outlier Detection

# + Review: (What is Machine Learning?)

- *"A Computer program is said to **Learn from Experience** with respect to some **class of Task T** and Performance measure **P**, if its performance at task in **T**, as measured by **P,** improves with experience **E".***

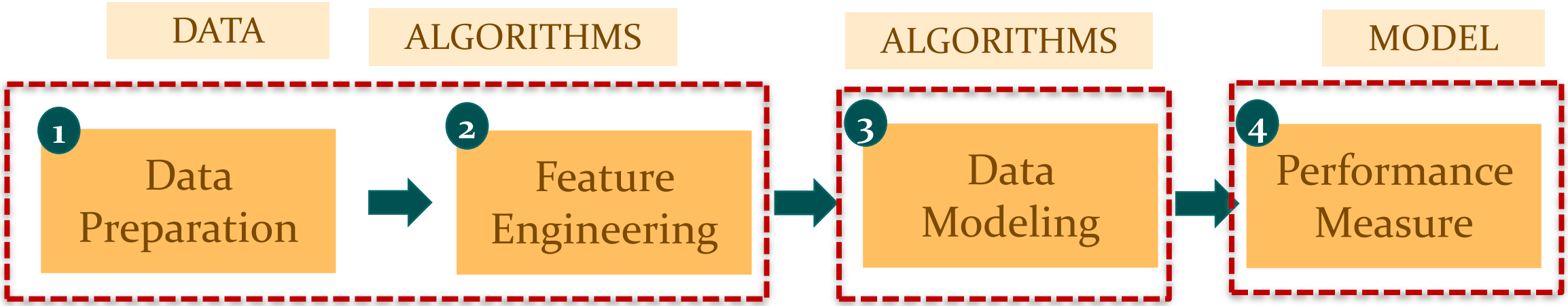
  Tom M. Mitchel, Computer Scientist, 1997

## What is Machine Learning Model?


- *"A **Machine Learning model** intends to determine the **optimal structure** in a dataset to **achieve** an **assigned task**.*
- *It results from **Learning algorithms** applied on a **training dataset**.*

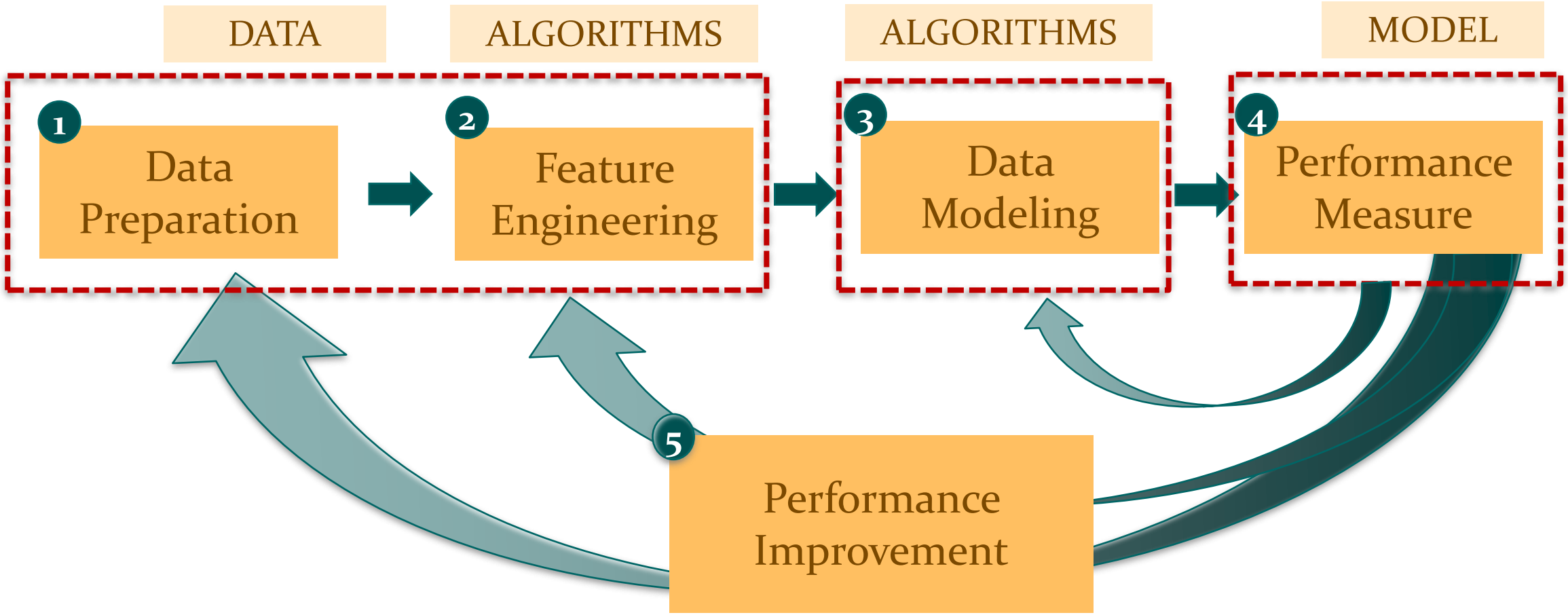
# Machine Learning Model

- Machine Learning Model contains 4 basic steps.



# Machine Learning Model

- Machine Learning Model is an iterative process.




- Until the model reaches a satisfying performance!

# Machine Learning Model

1 Data Preparation

- How can you ***import*** your ***raw*** data?
- What are the ***most common*** data ***cleaning methods***?

2 Feature Engineering

- How do you ***turn raw data*** into ***relevant data***?
- Turing data to ***meaningful*** for a ***learning algorithm***?
- How can you make the ***difference*** between ***useful*** and ***useless*** data in a huge dataset?

3 Data Modeling

- What are the different types of ***machine learning algorithms***?
- Which one should you ***choose*** to build your model?

# Machine Learning Model

4 Performance Measure

- What is the ***right method*** to ***access the performance*** of your ML algorithm?
- Which ***indicator*** should you ***use***?

5 Performance Improvement

- What are the ***reasons why your ML model is not performing well***?
- What are the ***most common techniques to improve the performance***?

# Three Things about ML

- ***Feature :*** Representation of raw data
- ***Model:*** Mathematical summary of features
- ***Making Something that work:*** Choosing the right model and features, given data and Task

# What is Features?

- The ***initial pick*** of feature is always an ***expression*** of ***prior knowledge.***
- ***Images*** → pixels, contours, textures, etc.
- ***Signal*** → samples, spectrograms, etc.
- ***Time series*** → ticks, trends, reversals, etc.
- ***Biological data*** → DNA, marker sequences, genes, etc.
- ***Text data*** → words, grammatical classes and relations, etc.

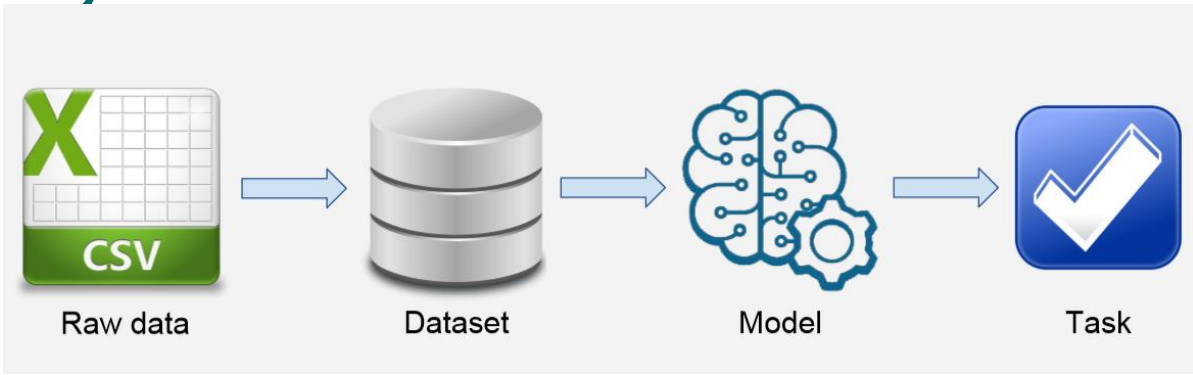
# Problem: Where to focus attention ?

- *Garbage In Garbage Out (GIGO)*
- *"Sometimes, less is better!".*

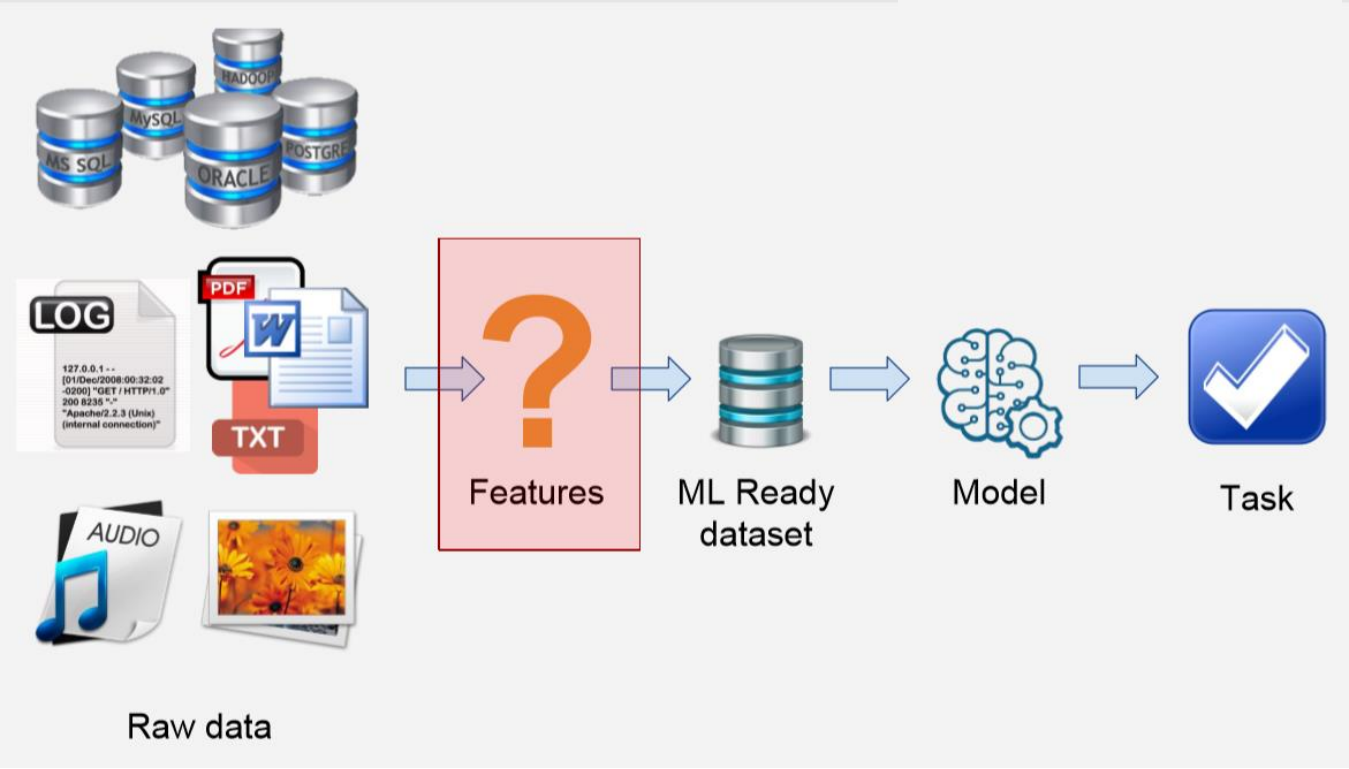
- A universal problem of intelligent (learning) agents is *where to focus their attention*.
- What *aspects* of the *problem* at hand are *important*/necessary to *solve* it?
- *Discriminate* between the *relevant* and *irrelevant* parts of *experience*.

# + Dream Vs. Reality

***DREAM***




***REALITY***



# + Missing Values Treatment

- Why missing value treatment is required ?
- Why data has missing values?
- Which are the methods to treat missing value ?

# + Missing Values Treatment

- Missing values are representative of the ***messiness*** of real-world data.
- There can be a ***multitude of reasons*** why they occur—ranging from
  - ***human errors*** during data entry,
  - ***incorrect sensor*** readings,
  - to ***software bugs*** in the data processing pipeline.
- Treating missing data is the ***fundamental*** and ***core element*** for the ***data analysis*** and / or ***machine learning***

# Why missing values treatment is required?

- *Missing data* in the training data set can *reduce the power / fit* of a *model* or can *lead to a biased model* because we have *not analysed* the *behaviour and relationship* with other variables *correctly*.
- It can lead to *wrong prediction or classification*.
- Example:

Results with not treated missing values. The inference from this data set is that the chances of playing cricket by males is higher than females.

| Name | Weight | Gender | Play Cricket/ Not |
|---|---|---|---|
| Mr. Amit | 58 | M | Y |
| Mr. Anil | 61 | M | Y |
| Miss Swati | 58 | F | N |
| Miss Richa | 55 | | Y |
| Mr. Steve | 55 | M | N |
| Miss Reena | 64 | F | Y |
| Miss Rashmi | 57 | | Y |
| Mr. Kunal | 57 | M | N |

| Gender | #Students | #Play Cricket | %Play Cricket |
|---|---|---|---|
| F | 2 | 1 | 50% |
| M | 4 | 2 | 50% |
| Missing | 2 | 2 | 100% |

| Name | Weight | Gender | Play Cricket/ Not |
|---|---|---|---|
| Mr. Amit | 58 | M | Y |
| Mr. Anil | 61 | M | Y |
| Miss Swati | 58 | F | N |
| Miss Richa | 55 | F | Y |
| Mr. Steve | 55 | M | N |
| Miss Reena | 64 | F | Y |
| Miss Rashmi | 57 | F | Y |
| Mr. Kunal | 57 | M | N |

| Gender | #Students | #Play Cricket | %Play Cricket |
|---|---|---|---|
| F | 4 | 3 | 75% |
| M | 4 | 2 | 50% |

Result with treated missing values, we can see that females have higher chances of playing cricket compared to males.

# + Dealing With Missing Values

- Some of your columns will ***certainly*** contains ***missing values***, often represented as '***NaN', empty column, zeros***.

$$\textbf{\textit{Compute Ratio(Missing Values)}} R_m = \frac{\textbf{\textit{Number of Missing Values}}}{\textbf{\textit{Total Number of Values}}}$$

- If $R_m$ ***is high***, You might need to ***remove*** the whole ***Column***.
- If $R_m$ is ***reasonable low***, to ***avoid losing data***, you can ***impute*** the ***mean***, the ***median*** or the ***most frequent*** value in place of the missing value

# Methods to Treat Missing Values ?

- The best is to get the ***actual value that was missing*** by going ***back*** to the ***Data Extraction & Collection*** stage and ***correcting possible errors during these stages***.
  - ***Which is not possible in most of the cases***

- There are ***two main*** techniques to treat missing data.
  - Deletion
  - Imputation

# 1. Deletion

- Unless the nature of missing data is '**Missing completely at random**', the best avoidable method in many cases is deletion.

- ***Listwise:*** In this case, rows containing missing variables are deleted. It suffers the ***maximum information loss***.

| User | Device | OS | Transactions |
|---|---|---|---|
| ~~A~~ | ~~Mobile~~ | ~~NA~~ | ~~5~~ |
| B | Mobile | Android | 3 |
| ~~C~~ | ~~NA~~ | ~~iOS~~ | ~~2~~ |
| D | Tablet | Android | 1 |
| E | Mobile | iOS | 4 |

- ***Pairwise:*** In this case, only the ***missing observations*** are ***ignored***, and analysis is done on ***variables present***. The problem is that even though it takes the available cases, ***one can't compare analyses because the sample is different every time***.

| User | Device | OS | Transactions |
|---|---|---|---|
| A | Mobile | ~~NA~~ | 5 |
| B | Mobile | Android | 3 |
| C | ~~NA~~ | iOS | 2 |
| D | Tablet | Android | 1 |
| E | Mobile | iOS | 4 |

- ***Deleting Columns:*** In most cases if the missing data constitutes more than 90% of the data then the column is dropped as it would not contribute to the mode.

# 2. Imputation

- Replacing With Mean/Median/Mode
- Assigning A Unique Category
- Assigning A Most frequent Value
- Using Algorithms Which Support Missing Values

# 1. Replacing With Mean/Median/Mode

- This strategy can be applied on a ***feature*** which has ***numeric data*** like the age of a person or the ticket fare.
- We can calculate the ***mean, median or mode of the feature*** and ***replace*** it with the missing values.
- This is an ***approximation*** which can add ***variance*** to the data set.
- It is a ***statistical approach of handling the missing values***

| OS | Revenue |
|---|---|
| Android | 1,804 |
| iOS | 3,027 |
| iOS | 8,788 |
| Android | NA |
| Android | 3,735 |
| Android | 1,056 |
| iOS | 9,319 |
| Android | 6,199 |
| Android | 2,235 |
| iOS | NA |
| Android | 1,146 |

| OS | Global Mean | Group Mean |
|---|---|---|
| Android | 1,804 | 1,804 |
| iOS | 3,027 | 3,027 |
| iOS | 8,788 | 8,788 |
| Android | 4,145 | 2,696 |
| Android | 3,735 | 3,735 |
| Android | 1,056 | 1,056 |
| iOS | 9,319 | 9,319 |
| Android | 6,199 | 6,199 |
| Android | 2,235 | 2,235 |
| iOS | 4,145 | 7,045 |
| Android | 1,146 | 1,146 |

# 2. Assigning a Unique Category

- A ***categorical feature*** will have a ***definite number of possibilities***, such as gender, for example.
- Since they have a definite number of classes, we can ***assign another class*** for the missing values.
- Missing values can be treated as a ***separate category*** by itself.
- The missing values which can be replaced with a new category, say, ***U for 'unknown'***.
- This strategy will ***add more information*** into the dataset which will ***result in the change of variance***.

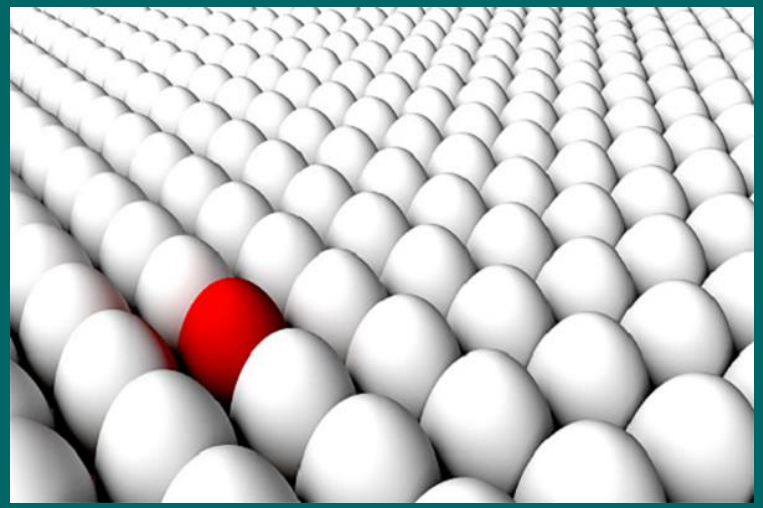
# 3. Assigning a Most frequent Value

- **Frequent Value:** The standard thing to do is to replace the missing entry with the most frequent one

# 4. Using Prediction Algorithm

- **Prediction models:** We can create a predictive model to estimate values that will substitute the missing data.
- **KNN** is a machine learning algorithm which works on the principle of ***distance measure***.
- This algorithm can be used when there are ***nulls*** present in the dataset.
- While the algorithm is applied, KNN considers the missing values by taking ***the majority of the K nearest values***.
- **RandomForest:** This model produces a ***robust result*** because it works well on ***non-linear and the categorical data***.
- It adapts to the data structure taking into consideration of the ***high variance or the bias, producing better results on large datasets.***

# Outlier Detection



- Outliers and Outlier Analysis
- What Are Outliers?
- Types of Outliers
- Challenges of Outlier Detection
- Outlier Detection Methods
- Application of Outlier Detection
- Evaluation

# Anomalies/Outliers

نغرق فى فيضان الداتا

- ***We are drowning in the deluge of data that are being collected world-wide, while starving for knowledge at the same time****

بشكل نسبى نادر f

- Anomalous events occur ***relatively infrequently***
- However, when they do occur, ***their consequences can be quite dramatic and quite often in a negative sense***

غير مبالى

لا يحدث بشكل متكرر

**"Mining needle in a haystack. So much hay and so little time"**

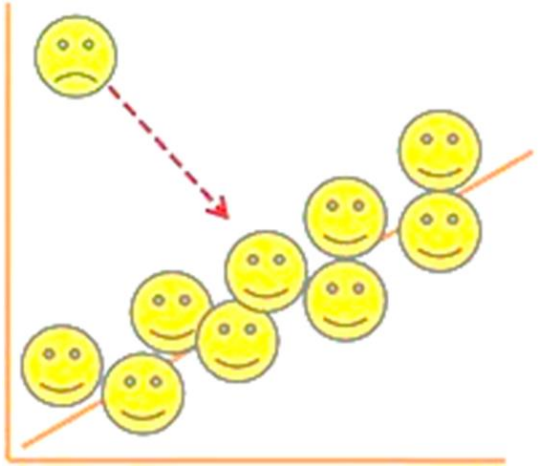
- ***Anomaly*** is a pattern in the ***data that does not conform*** to the ***expected behavior*** also referred to as outliers

# What is an Outlier?

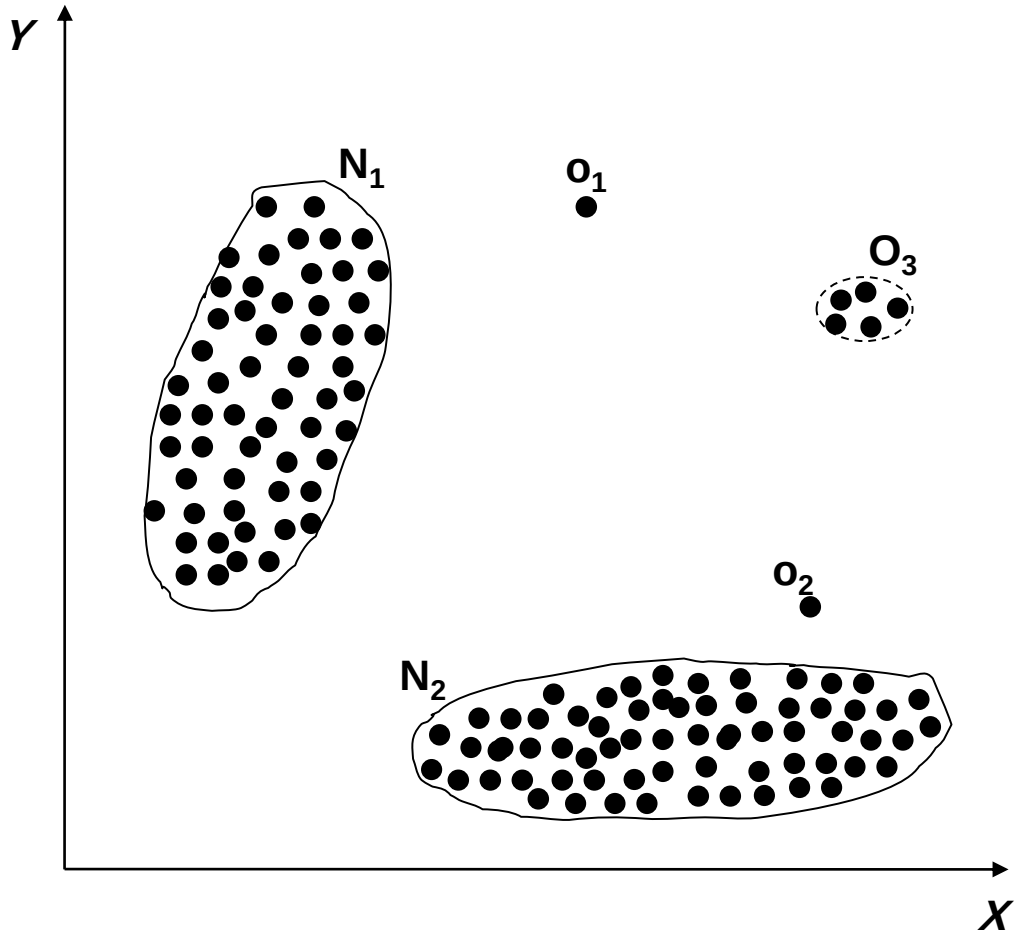
- **Outlier** is a commonly used terminology by analysts and data scientists as it needs close attention else it can result in wildly wrong estimations.

- **Simply speaking**, Outlier is an observation that appears far ***away*** and ***diverges*** from an ***overall pattern*** in a sample.

How do you even detect the presence of outliers and how extreme they are?



# + Example




- $N_1$ and $N_2$ are regions of normal behavior
- Points $o_1$ and $o_2$ are anomalies
- Points in region $O_3$ are anomalies

- Example: Age of a person

# Anomalies/Outliers

- **What are outliers?**
  - An outlier is a ***data object*** that ***deviates significantly*** from the ***rest of the objects,*** as if it were generated by a ***different mechanism***.
  - We may refer to data objects that are not outliers as "***normal***" or ***expected data***. Similarly, we may refer to outliers as "***abnormal***" data.
- Also referred to as outliers, ***exceptions***, ***peculiarities***, ***surprise***, etc.
- **Outliers are different from noisy data**
- ***"What is noise?"***
  - Noise is a ***random error*** or ***variance*** in a measured variable.
- ***Outliers are interesting:*** an outlier violates the mechanism that generates the normal data.
- Noise is ***not interesting*** in data ***analysis***.

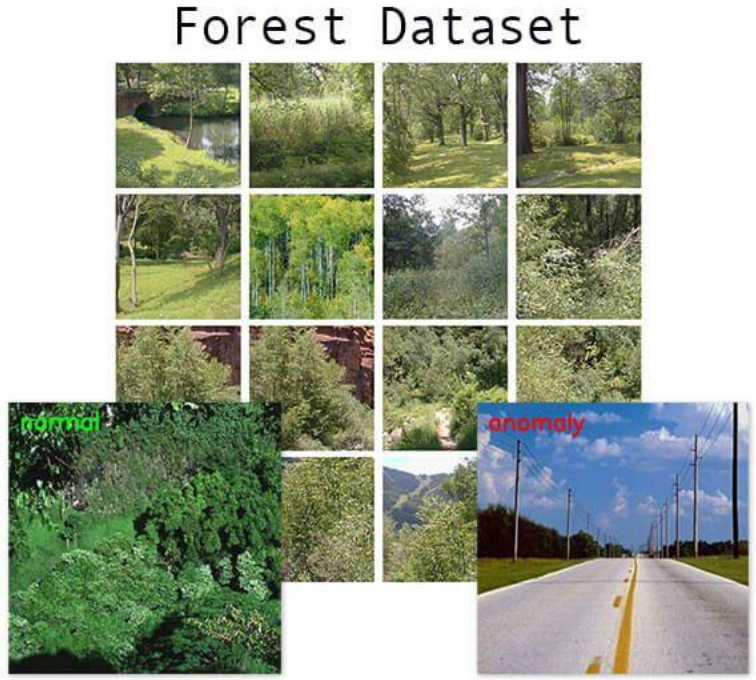
# + Anomalies/Outliers

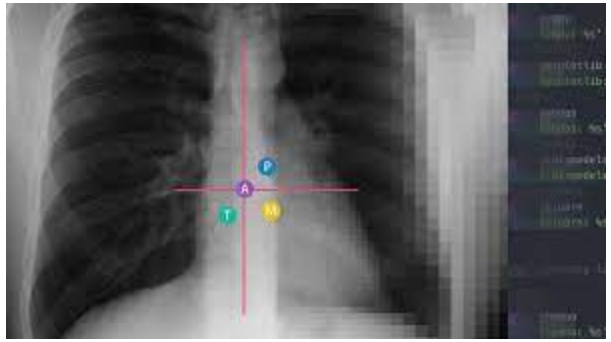
- Outliers are *interesting* because they are *suspected* of *not* being *generated* by the *same* mechanisms as the rest of the data.
- Therefore, in outlier detection, it is important to *justify why the outliers detected are generated by some other mechanisms.*
- This is often achieved by making various assumptions on the rest of the data and showing that the outliers detected violate those assumptions significantly.

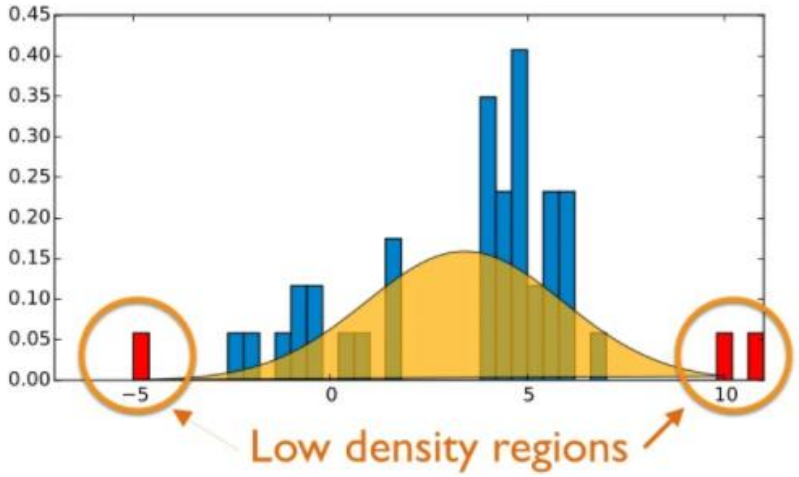
# + Novelty Detection

- Outlier detection is also related to ***novelty detection*** in ***evolving*** data sets.
- For example, by monitoring a ***social media web site*** where ***new*** content is ***incoming***, novelty detection may identify ***new topics*** and ***trends*** in a ***timely*** manner.
- Novel topics may initially appear as ***outliers***.
- To this extent, ***outlier detection and novelty detection*** share some similarity in ***modeling*** and ***detection*** methods.
- However, a ***critical difference between the two is that in novelty detection, once new topics are confirmed, they are usually incorporated into the model of normal behavior so that follow-up instances are not treated as outliers anymore***

# + Outliers/Anomalies





# Related problems

- Rare Class Mining
- Chance discovery اكتشاف الفرص
- Novelty Detection اكتشاف الجديد
- Exception Mining التنقيب عن الاستثناءات

# + Key Challenges

- Defining a ***representative normal region*** is ***challenging***.
- The ***boundary*** between ***normal*** and ***outlying*** behavior is often ***not precise***.
- The ***exact notion*** of an outlier is ***different*** for ***different*** application ***domains***.
- Availability of ***labeled*** data for training/validation
- ***Malicious adversaries***.

الاشياء الغريبة المختبئة

- Data might contain ***noise***.
- Normal behavior keeps ***evolving***.

# Aspects of Anomaly Detection Problem

- Nature of input data
- Availability of supervision
- Type of anomaly: point, contextual, structural
- Output of anomaly detection
- Evaluation of anomaly detection techniques

# 1. Nature of the Data: Input Data

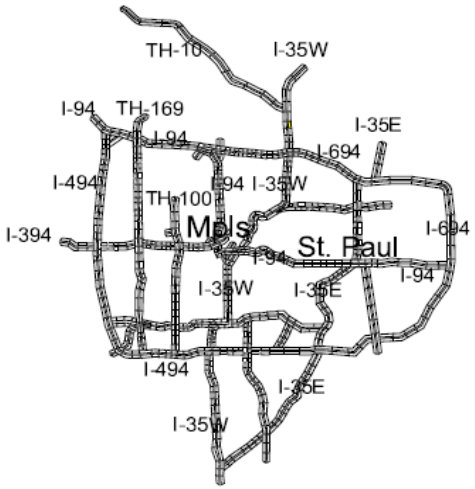
- Most common form of data handled by anomaly detection techniques is ***Record Data***
  - Univariate
  - Multivariate
- Nature of attributes
  - Binary
  - Categorical
  - Continuous
  - Hybrid

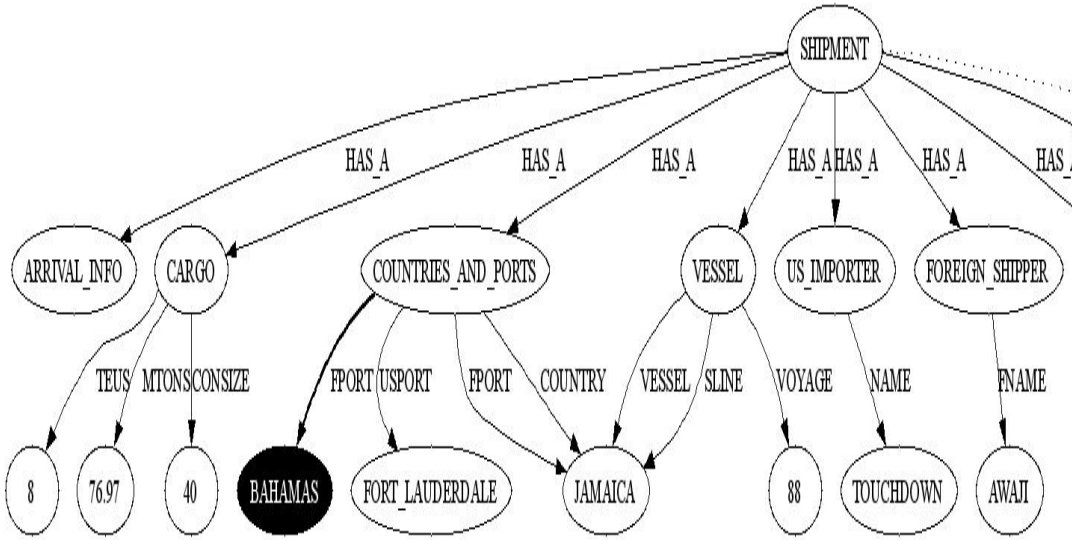
| | categorical | continuous | categorical | continuous | Binary |
|---|---|---|---|---|---|
| *Tid* | SrcIP | Start time | Dest IP | Dest Port | Number of bytes | Attack |
| 1 | 206.135.38.95 | 11:07:20 | 160.94.179.223 | 139 | 192 | No |
| 2 | 206.163.37.95 | 11:13:56 | 160.94.179.219 | 139 | 195 | No |
| 3 | 206.163.37.95 | 11:14:29 | 160.94.179.217 | 139 | 180 | No |
| 4 | 206.163.37.95 | 11:14:30 | 160.94.179.255 | 139 | 199 | No |
| 5 | 206.163.37.95 | 11:14:32 | 160.94.179.254 | 139 | 19 | Yes |
| 6 | 206.163.37.95 | 11:14:35 | 160.94.179.253 | 139 | 177 | No |
| 7 | 206.163.37.95 | 11:14:36 | 160.94.179.252 | 139 | 172 | No |
| 8 | 206.163.37.95 | 11:14:38 | 160.94.179.251 | 139 | 285 | Yes |
| 9 | 206.163.37.95 | 11:14:41 | 160.94.179.250 | 139 | 195 | No |
| 10 | 206.163.37.95 | 11:14:44 | 160.94.179.249 | 139 | 163 | Yes |

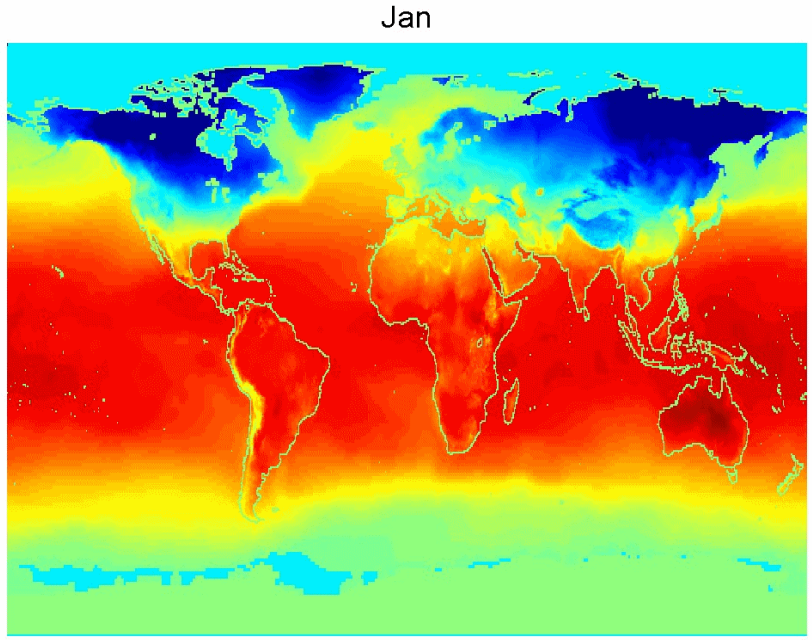
# 1. Nature of the Data: Input Data

- Relationship among data instances
  - Sequential
    - Temporal
  - Spatial
  - Spatio-temporal
  - Graph

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```



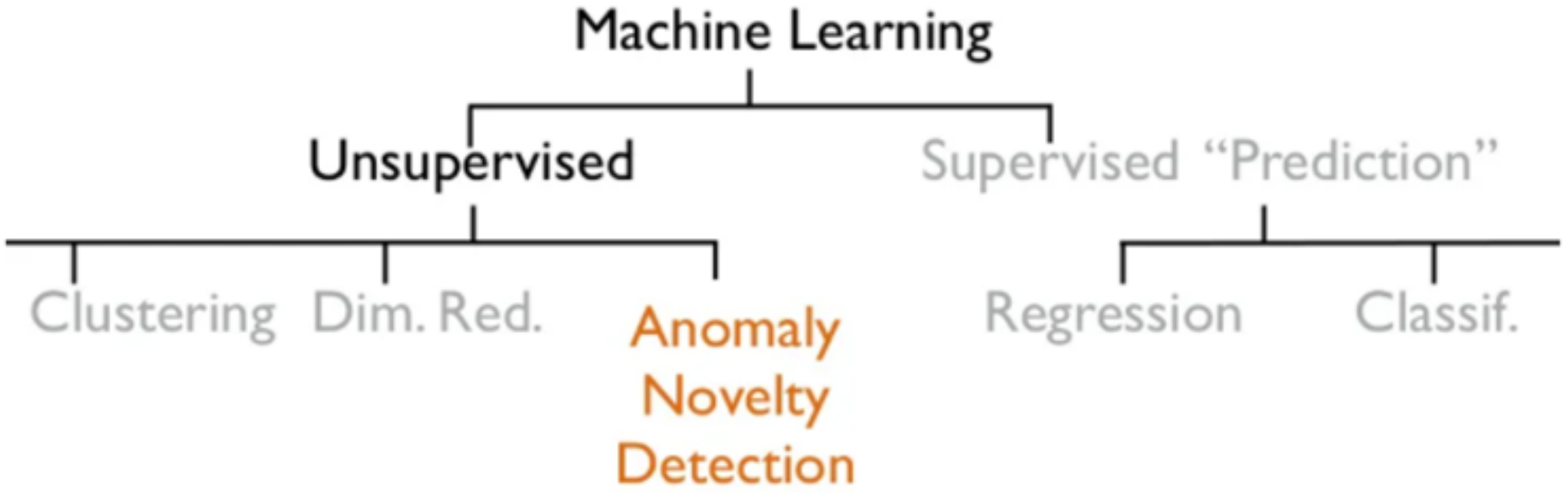



# 2. Availability of supervision: Data Labels

- ***Supervised Anomaly Detection***
  - Labels available for both normal data and anomalies
  - Similar to rare class mining/imbalanced classification
- ***Semi-supervised (Anomaly/novelty Detection)***
  - Labels available only for normal data.
  - The algorithms learns on normal data only
- ***Unsupervised Anomaly Detection (Outlier Detection)***
  - No labels assumed (training set=normal data + abnormal Data)
  - Based on the assumption that anomalies are very rare compared to normal data

# Machine Learning Taxonomy



# 3. Types of Outlier/Anomaly

- ***Three kinds:***
  - ***Global Outliers*** (Point Anomalies)
  - ***Contextual Outliers*** (Conditional Anomalies)
  - ***Collective Outliers***
- A data set may have ***multiple*** types of ***outlier***
- ***One*** object may ***belong*** to ***more*** than ***one*** type of ***outlier***

Global anomalies affect the entire system uniformly. Contextual anomalies occur within specific contexts or subsets of data. Collective anomalies involve collective behavior of multiple data points or entities

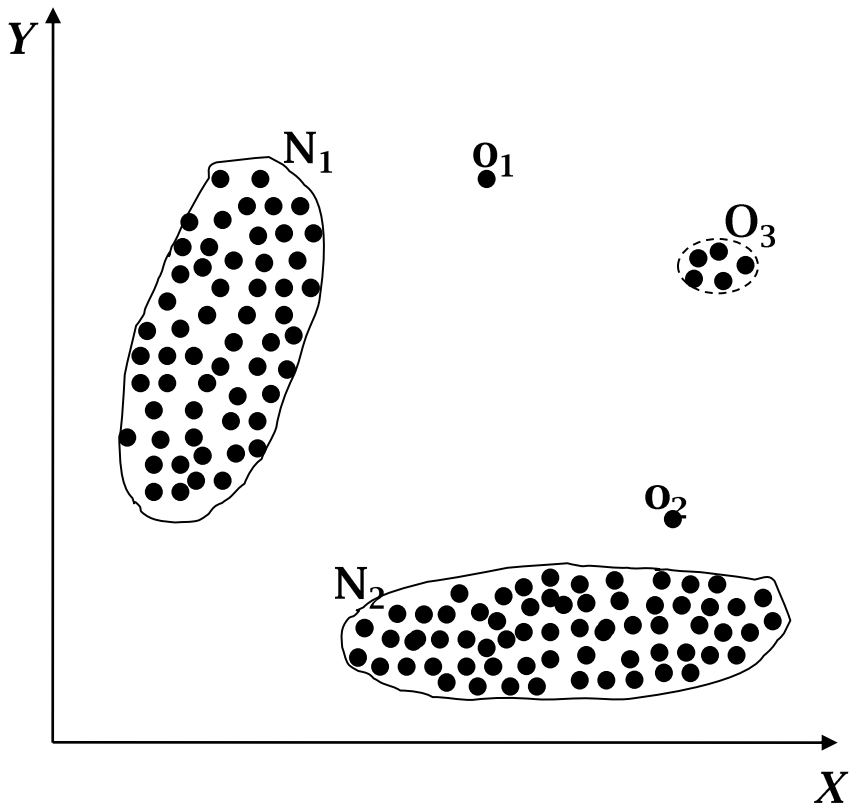
# + Point Anomalies

- Global outlier (or point anomaly)
  - An outlier object ***significantly deviates*** from the ***rest*** of the data set

تنحرف بشكل كبير

- ***Challenge***: find an appropriate ***measurement*** of ***deviation***
- An individual data instance is anomalous w.r.t. the data.

نقاط متطرفة

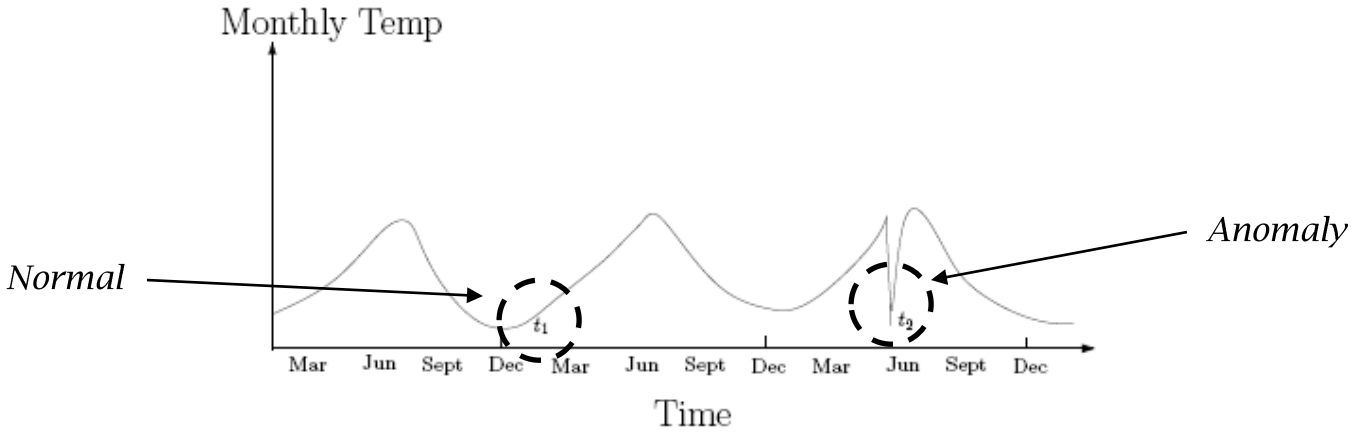


# + Contextual Anomalies

- An individual data instance is anomalous ***within a context***
- Requires a ***notion*** of ***context***

تتطلب فهم السياق الكلى للداتا

- Also referred to as ***conditional anomalies****




** Xiuyao Song, Mingxi Wu, Christopher Jermaine, Sanjay Ranka, Conditional Anomaly Detection, IEEE Transactions on Data and Knowledge Engineering, 2006.*

# Contextual Anomalies

- An outlier object ***deviates*** significantly based on a ***selected context***.
  - ***Is 10C in Vancouver an outlier? (depending on summer or winter?)***
- ***Attributes*** of data objects should be divided into ***two groups***.
  - ***Contextual attributes:*** defines the context, e.g., time & location.
  - ***Behavioral attributes:*** characteristics of the object, used in outlier evaluation, e.g., temperature.
- Contextual outliers are a generalization of local outlier—whose density significantly deviates from its local area.
- ***Challenge***: how to define or formulate meaningful context?

# + Global Outlier

- ***Global outlier detection*** can be regarded as a ***special case*** of ***contextual*** outlier detection where the set of ***contextual attributes is empty***.
- In other words, global outlier detection uses the ***whole data set*** as the ***context***.
- Contextual outlier analysis provides ***flexibility to users*** in that one can ***examine*** outliers in ***different contexts***, which can be ***highly desirable*** in ***many applications***.

# + Applications of Anomaly Detection

- Network intrusion detection
- Insurance / Credit card fraud detection
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining
- …

# + Application: Intrusion Detection


- ***Intrusion Detection***
  - Process of monitoring the ***events*** occurring in a ***computer system*** or network and ***analyzing*** them for ***intrusions***
  - Intrusions are defined as ***attempts to bypass the security mechanisms*** of a computer or network
- ***Challenges***
  - Traditional ***signature-based intrusion detection systems*** are based on ***signatures*** of known ***attacks*** and cannot ***detect emerging cyber threats***
  - Substantial latency in deployment of newly created signatures across the computer system
- ***Anomaly detection can alleviate these limitations***

# + Applications of Anomaly Detection

- ***Fraud detection:***
  - Fraud detection refers to detection of ***criminal activities*** occurring in ***commercial organizations***.
  - Malicious users might be the ***actual customers*** of the ***organization*** or might be ***posing*** as a ***customer*** (also ***known as identity theft***).


- ***Types of fraud***
  - Credit card fraud.
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading التداول الداخلى من موظفى الشركة

- ***Challenges***
  - ***Fast*** and ***accurate*** real-time detection.
  - ***Misclassification*** cost is very ***high***

# + Applications of Anomaly Detection

- ***Industrial damage detection***
  - Refers to detection of different ***faults*** and ***failures*** in ***complex industrial systems***, ***structural*** damages, ***intrusions*** in ***electronic security systems***, ***suspicious*** events in video surveillance, ***abnormal*** energy consumption, ***etc***.

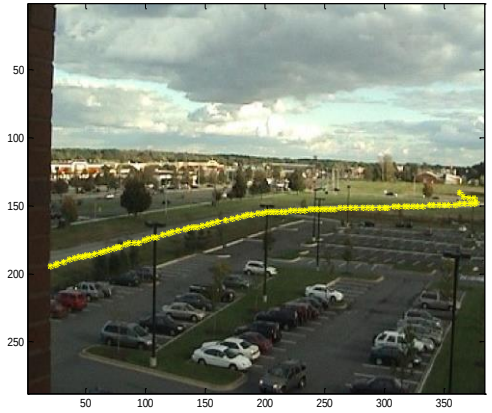
- ***Example:***
  - Aircraft Safety  Anomalous Aircraft (Engine) / Fleet Usage. سلامة الطائرات - والاساطيل
  - Anomalies in engine combustion data. شذود فى بيانات احتراق المحرك
  - Total aircraft health and usage management  Key

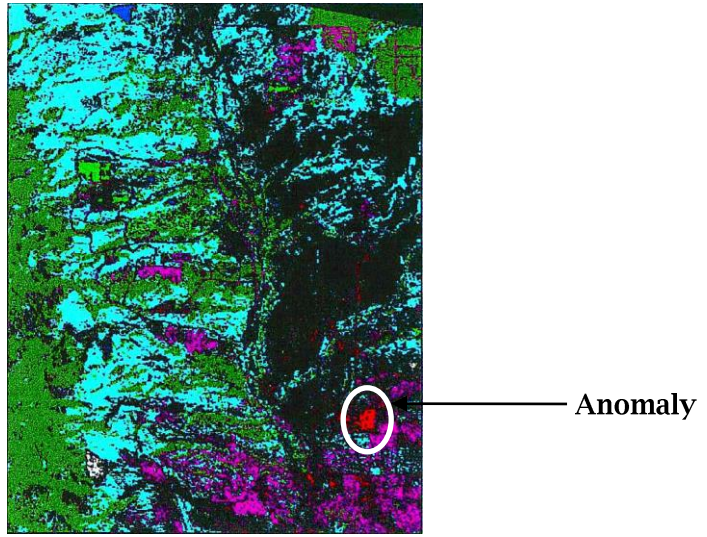
- ***Challenges:***
  - Data is ***extremely huge***, ***noisy*** and ***unlabeled***.
  - Most of applications ***exhibit temporal behavior***. سلوك غريب عن المعتاد
  - Detecting anomalous events typically require ***immediate intervention***

# Image Processing

- Detecting outliers in an image monitored over time
- Detecting anomalous regions within an ***image***
- Used in
  - ***mammography image analysis***
  - ***video surveillance***
  - ***satellite image analysis***
- ***Key Challenges***
  - Detecting ***collective anomalies***
  - Data sets are ***very large***

مشاكل الثدى



# Challenges of Anomaly detection

- Modeling ***normal objects*** and ***outliers*** properly.
  - Hard to ***enumerate all possible normal behaviors in an application***.
  - The ***border*** between ***normal*** and ***outlier*** objects is often a gray area
- Application-specific outlier detection.
  - Choice of ***distance measure*** among objects and the model of ***relationship*** among objects are often ***application–dependent***.
- ***Example: clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations***

# Challenges of Anomaly detection

- ***Handling noise in outlier detection.***
  - ***Noise*** may ***distort*** the normal objects and ***blur*** the ***distinction*** between ***normal*** objects and ***outliers***.
  - Noise may help ***hide outliers*** and ***reduce*** the ***effectiveness*** of outlier detection.
- ***Understandability***
  - Understand why these are outliers: ***Justification of the detection***.
  - Specify the ***degree*** of an ***outlier***: the ***unlikelihood*** of the object being generated by a ***normal*** mechanism.

احتمالية

# + Methods for anomaly detection

- Outlier Detection Methods.
- ***Whether user–labeled examples of outliers can be obtained.***
  - Supervised, Semi-Supervised, and Unsupervised Methods.
- ***Assumptions about normal data and outliers.***
  - Statistical Methods, Proximity-Based Methods, and Clustering-Based Methods.

# + Supervised Methods

- ***Modeling outlier detection as a classification problem.***
  - Samples examined by domain experts used for training & testing
- Methods for ***Learning*** a ***classifier*** for ***outlier detection*** effectively:
  - Model normal objects & report those not matching the model as outliers, or
  - Model outliers and treat those not matching the model as normal
- ***Challenges***
  - ***Imbalanced classes***, i.e., outliers are rare: Boost the outlier class and make up some artificial outliers.
  - Catch as many outliers as possible, i.e., ***recall*** is more important than ***accuracy*** (i.e., not mislabeling normal objects as outliers)

# + Unsupervised Methods

- Assume the normal objects are somewhat "***clustered***" into ***multiple groups***, each having some ***distinct features***
- An outlier is expected to be ***far away*** from any groups of normal objects
- ***Weakness***: ***Cannot detect collective outlier effectively***
  - Normal objects ***may not*** share any ***strong patterns***, but the collective outliers may share ***high similarity in a small area***
- Many clustering methods can be adapted for unsupervised methods.
- ***Find clusters, then outliers: not belonging to any cluster***

# + Unsupervised Methods: Challenges

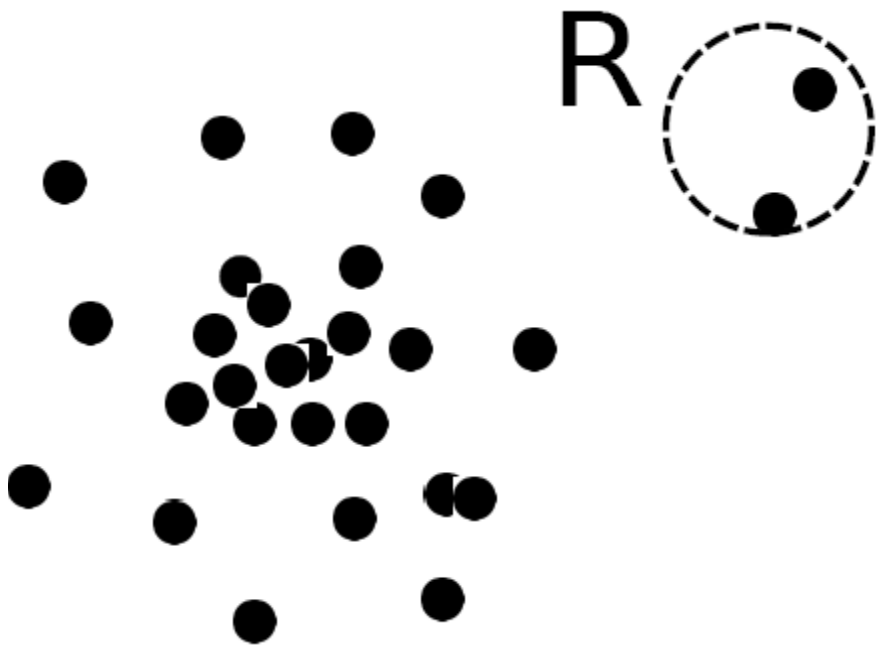
- In *some intrusion or virus detection*, *normal* activities are *diverse*. مختلفة
  - Unsupervised methods may have a *high false positive rate* but still *miss* many *real outliers*.
  - Supervised methods can be *more effective*, e.g., *identify attacking some key resources*
- *Challenges*.
  - Hard to *distinguish noise from outliers*.
  - Costly since *first clustering*: but far less outliers than normal objects مكلفة فى حالة الغريب يكون اقل
- *Newer methods: tackle outliers directly*

# + Semi-Supervised Methods

- In many applications, the number of ***labeled data is often small***
  - Labels could be on outliers only, normal objects only, or both·
- If some labeled ***normal objects are available***
  - Use the ***labeled examples*** and the ***proximate unlabeled*** objects to train a model for ***normal*** objects.
  - Those not ***fitting the model of normal objects*** are detected as ***outliers***
- If only ***some labeled outliers*** are available, a ***small number of labeled outliers many not cover the possible outliers well***.
  - To improve the ***quality of outlier detection***, one can get help from models for normal objects learned from unsupervised methods

# + Proximity-based Methods

- An object is an ***outlier*** if the ***nearest neighbors*** of the object are ***far away***, i.e., the proximity of the object is ***significantly deviates*** from the ***proximity*** of ***most*** of the ***other*** objects in the ***same data set.***



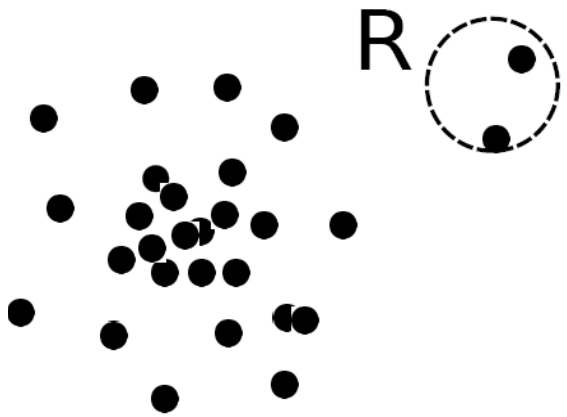
# Challenges: Proximity based

- The ***effectiveness*** of proximity-based methods highly relies on the ***proximity measure***.
- In some applications, ***proximity or distance measures cannot*** be ***obtained easily***.
- Often have a ***difficulty in identifying a group of outliers that stay close to each other***.
- Two major types of proximity-based outlier detection methods.
  - ***Distance–based vs. density–based***

# + Clustering-based Methods

- Normal data belong to ***large and dense clusters***, whereas ***outliers*** belong to ***small*** or ***sparse clusters***, or ***do not belong to any clusters***.

- Clustering based
  - Nearest-neighbor based
  - Density based

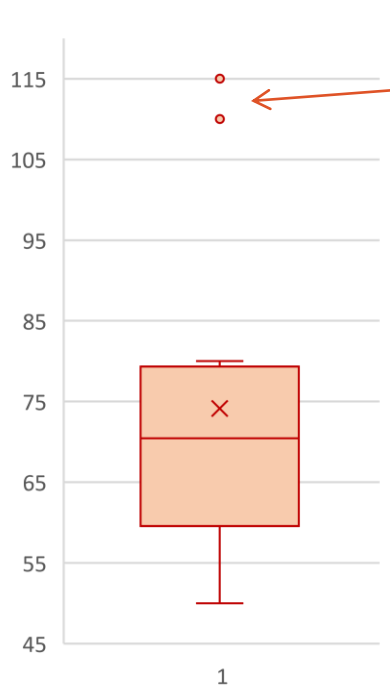


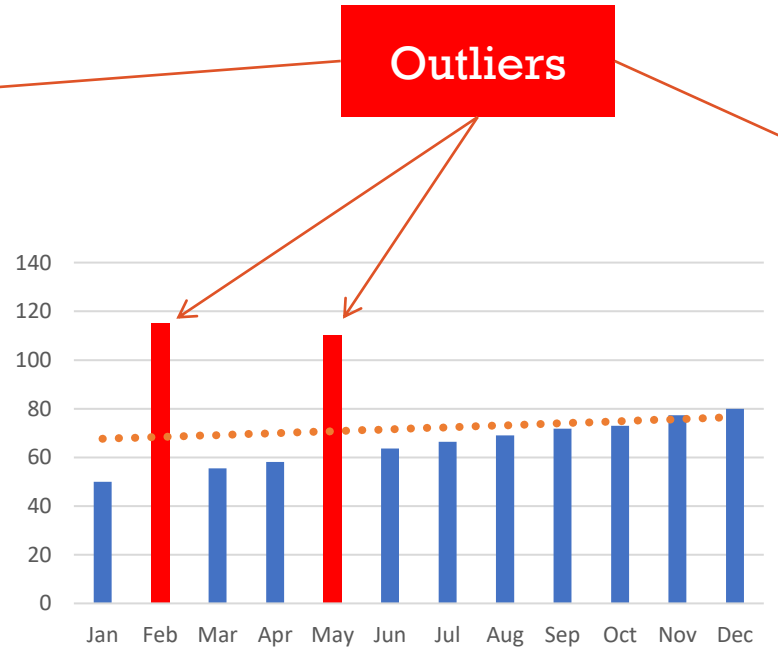

- ***Challenges***
  - Clustering is expensive: straightforward adaption of a clustering method for outlier detection can be costly and does not scale up well for large data sets.

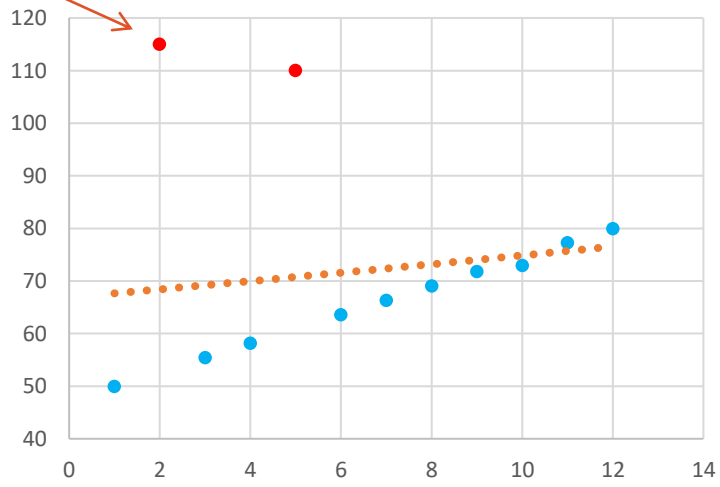
# Statistical Outlier Analysis

- Most commonly used method to ***detect*** outliers is visualization.
  - Various visualization methods, like ***Box–plot, Histogram, Scatter Plot***.

| Quarter - Income | 2017 |
|---|---|
| Jan | 50 |
| Feb | 115 |
| Mar | 55 |
| Apr | 58 |
| May | 110 |
| Jun | 64 |
| Jul | 66 |
| Aug | 69 |
| Sep | 72 |
| Oct | 73 |
| Nov | 77 |
| Dec | 80 |




***Box-plot*** ***Histogram*** ***Scatter Plot***

# Statistical Outlier Analysis

- ***Apply a statistical test that depends on***
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)
- ***Limitation***
  - Most of the tests are for a single attribute
  - In many cases, data distribution may not be known
  - For high dimensional data, it may be difficult to estimate the true distribution

# + Statistical Outlier Analysis

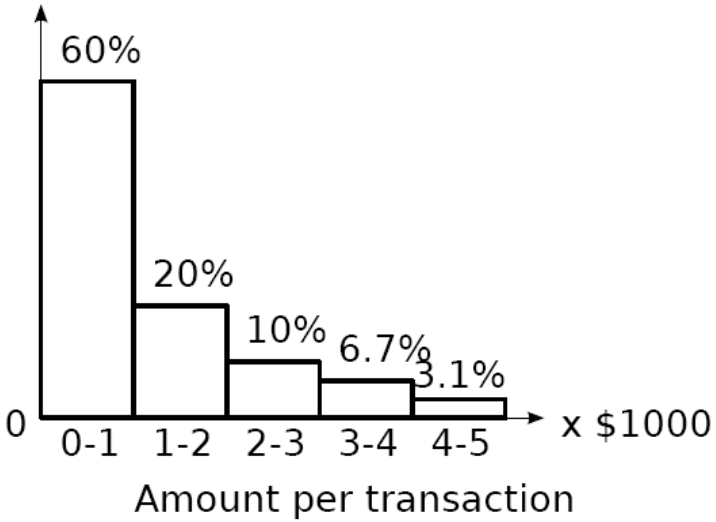
- ***Assumption***: the objects in a data set are generated by a ***(stochastic) process (a generative model).***
- Learn a ***generative model fitting*** the given data set, and then ***identify*** the ***objects*** in ***low probability regions of the model as outliers***.
- Two categories: ***parametric versus nonparametric***.
- Statistical methods (also known as model based methods) assume that the ***normal data*** follow ***some statistical model***.
  - The data not following the model are outliers

# + Parametric Methods

- Assumption: the normal data is generated by a ***parametric distribution*** with ***parameter θ.***
- The probability ***density function*** of the parametric distribution ***f(x | θ) gives the probability that object x is generated by the distribution***
- ***The smaller this value, the more likely x is an outlier***

# + Non-parametric Method

- Not assume an ***a-priori statistical model, instead, determine the model from the input data.***
  - Not completely parameter free but consider the number and nature of the parameters are flexible and not fixed in advance.
- Examples: ***histogram and kernel density estimation***.
- A transaction in the amount of $7,500 is an outlier, since only 0.2% transactions have an amount higher than $5,000



# + Challenges: Non Parametric method

- Hard to choose an ***appropriate bin size for histogram***.
  - ***Too small bin size*** → normal objects in empty/rare bins, false positive .
  - ***Too big bin size*** → outliers in some frequent bins, false negative

# Evaluation of Anomaly Detection – F-value

- ***Accuracy is not sufficient metric for evaluation***
- ***Example***: network traffic data set with ***99.9% of normal data*** and ***0.1% of intrusions***
- Trivial classifier that labels everything with the ***normal class can achieve 99.9% accuracy !!!!!***
- Focus on both ***recall and precision***
  - Recall (R) = TP/(TP + FN)
  - Precision (P) = TP/(TP + FP)
  - F – measure = 2*R*P/(R+P)

| |
|---|
| Anomaly class (AC) |
| Normal class (NC) |

| Confusion matrix | | Predicted class | |
|---|---|---|---|
| | | NC | AC |
| Actual class | NC | TN | FP |
| | AC | FN | TP |

# Evaluation of Anomaly Detection – ROC &AUC

- Standard measures for evaluating anomaly detection problems:
- ***Recall (Detection rate)*** – ratio between the number of correctly detected anomalies and the total number of anomalies
- ***False alarm (false positive) rate*** – ratio between the number of data records from normal class that are misclassified as anomalies and the total number of data records from normal class
- ***ROC Curve is a trade-off between detection rate and false alarm rate***
- ***Area under the ROC curve (AUC) is computed using a trapezoid rule***

**Will Discuss in detail in Classification Evaluation Measures**

# End of Lecture – 02